The Proof is in the Pudding

A Look at the Changing Nature of Mathematical Proof

Steven G. Krantz

January 9, 2008

To Jerry Lyons, mentor and friend.

Table of Contents

Preface

The title of this book is not entirely frivolous. There are many who will claim that the correct aphorism is "The proof of the pudding is in the eating." That it makes no sense to say, "The proof is in the pudding." Yet people say it all the time, and the intended meaning is always clear. So it is with mathematical proof. A *proof* in mathematics is a psychological device for convincing some person, or some audience, that a certain mathematical assertion is true. The structure, and the language used, in formulating that proof will be a product of the person creating it; but it also must be tailored to the audience that will be receiving it and evaluating it. Thus there is no "unique" or "right" or "best" proof of any given result. A proof is part of a situational ethic. Situations change, mathematical values and standards develop and evolve, and thus the very *way* that we do mathematics will alter and grow.

This is a book about the changing and growing nature of mathematical proof. In the earliest days of mathematics, "truths" were established heuristically and/or empirically. There was a heavy emphasis on calculation. There was almost no theory, and there was little in the way of mathematical notation as we know it today. Those who wanted to consider mathematical questions were thereby hindered: they had difficulty expressing their thoughts. They had particular trouble formulating general statements about mathematical ideas. Thus it was virtually impossible that they could state theorems and prove them.

Although there are some indications of proofs even on ancient Babylonian tablets from 1000 B.C.E., it seems that it is in ancient Greece that we find the identifiable provenance of the concept of proof. The earliest mathematical tablets contained numbers and elementary calculations. Because

of the paucity of texts that have survived, we do not know how it came about that someone decided that some of these mathematical procedures required *logical justification*. And we really do not know how the formal concept of proof evolved. The *Republic* of Plato contains a clear articulation of the proof concept. The *Physics* of Aristotle not only discusses proofs, but treats minute distinctions of proof methodology (see our Chapter 11). Many other of the ancient Greeks, including Eudoxus, Theaetetus, Thales, Euclid, and Pythagoras, either used proofs or referred to proofs. Protagoras was a sophist, whose work was recognized by Plato. His *Antilogies* were tightly knit logical arguments that could be thought of as the germs of proofs.

But it must be acknowledged that Euclid was the first to systematically use precise definitions, axioms, and strict rules of logic. And to systematically *prove* every statement (i.e., every theorem). Euclid's formalism, and his methodology, has become the model—even to the present day—for establishing mathematical facts.

What is interesting is that a mathematical statement of fact is a free-standing entity with intrinsic merit and value. But a proof is a device of communication. The creator or discoverer of this new mathematical result wants others to believe it and accept it. In the physical sciences—chemistry, biology, or physics for example—the method for achieving this end is the *reproducible experiment*.[1] For the mathematician, the reproducible experiment is a proof that others can read and understand and validate.

Thus a "proof" can, in principle, take many different forms. To be effective, it will have to depend on the language, training, and values of the "receiver" of the proof. A calculus student has little experience with rigor and formalism; thus a "proof" for a calculus student will take one form. A professional mathematician will have a different set of values and experiences, and certainly different training; so a proof for the mathematician will take a different form. In today's world there is considerable discussion—among *mathematicians*—about what constitutes a proof. And for physicists, who are our intellectual cousins, matters are even more confused. There are those workers in physics (such as Arthur Jaffe of Harvard, Charles Fefferman of Princeton, Ed Witten of the Institute for Advanced Study, Frank Wilczek of MIT, and Roger Penrose of Oxford) who believe that physical concepts

[1] More precisely, it is the reproducible experiment *with control*. For the careful scientist compares the results of his/her experiment with some standard or norm. That is the means of evaluating the result.

should be derived from first principles, just like theorems. There are other physicists—probably in the majority—who reject such a theoretical approach and instead insist that physics is an empirical mode of discourse. These two camps are in a protracted and never-ending battle over the turf of their subject. Roger Penrose's new book *The Road to Reality: A Complete Guide to the Laws of the Universe*, and the vehement reviews of it that have appeared, is but one symptom of the ongoing battle.

The idea of "proof" certainly appears in many aspects of life other than mathematics. In the courtroom, a lawyer (either for the prosecution or the defense) must establish his/her case by means of an accepted version of proof. For a criminal case this is "beyond a reasonable doubt" while for a civil case it is "the preponderance of evidence shows". Neither of these is mathematical proof, nor anything like it. For the real world has no formal definitions and no axioms; there is no sense of establishing facts by strict logical exegesis. The lawyer certainly uses logic—such as "the defendant is blind so he could not have driven to Topanga Canyon on the night of March 23" or "the defendant has no education and therefore could not have built the atomic bomb that was used to ..."—but his/her principal tools are *facts*. The lawyer proves the case beyond a reasonable doubt by amassing a preponderance of evidence in favor of that case.

At the same time, in ordinary, family-style parlance there is a notion of proof that is different from mathematical proof. A husband might say, "I believe that my wife is pregnant" while the wife may *know* that she is pregnant. Her pregnancy is not a permanent and immutable fact (like the Pythagorean theorem), but instead is a "temporary fact" that will be false after several months. Thus, in this circumstance, the concept of truth has a different meaning from the one that we use in mathematics, and the means of verification of a truth are also rather different. What we are really seeing here is the difference between knowledge and belief—something that never plays a formal role in mathematics.

It is also common for people to offer "proof of their love" for another individual. Clearly such a "proof" will not consist of a tightly linked chain of logical reasoning. Rather, it will involve emotions and events and promises and plans. There may be discussions of children, and care for aging parents, and relations with siblings. This is an entirely different kind of proof from the kind treated in the present book. It is in the spirit of this book in the sense that it is a "device for convincing someone that something is true." But it is *not* a mathematical proof.

The present book is concerned with mathematical proof. For more than 2000 years (since the time of Euclid), the concept of mathematical proof has not substantially changed. Traditional "proof" is what it has always been: a tightly knit sequence of statements knit together by strict rules of logic. It is noteworthy that the French school (embodied by Nicolas Bourbaki) and the German school (embodied by David Hilbert) gave us in the twentieth century a focused idea of what mathematics is, what the common body of terminology and basic concepts should be, and what the measure of rigor should be. But, until very recently, a proof was a proof; it followed a strict model and was formulated and recorded according to rigid rules.

The eminent French mathematician Jean Leray (1906–1998) perhaps sums up the value system of the modern mathematician:

> . . . all the different fields of mathematics are as inseparable as the different parts of a living organism; as a living organism mathematics has to be permanently recreated; each generation must reconstruct it wider, larger and more beautiful. The death of mathematical research would be the death of mathematical thinking which constitutes the structure of scientific language itself and by consequence the death of our scientific civilization. Therefore we must transmit to our children strength of character, moral values and drive towards an endeavouring life.

What Leray is telling us is that mathematical ideas travel well, and stand up under the test of time, just because we have such a rigorous and well-tested standard for formulating and recording the ideas. It is a grand tradition, and one well worth preserving.

The early twentieth century saw L. E. J. Brouwer's dramatic proof of his fixed-point theorem followed by his wholesale rejection of proof by contradiction (at least in the context of existence proofs—which is precisely what his fixed-point theorem was an instance of) and his creation of the intuitionist movement. This gauntlet was later taken up by Errett Bishop, and his *Foundations of Constructive Analysis* (written in 1967) has made quite a mark (see also the revised version, written jointly with Douglas Bridges, published in 1985). These ideas are of particular interest to the theoretical computer scientist, for proof by contradiction has no meaning in computer science (this despite the fact that Alan Turing cracked the Enigma Code by applying ideas of proof by contradiction in the context of computing machines).

In the past thirty years or so it has come about that we have re-thought, and re-invented, and certainly amplified our concept of proof. Certainly computers have played a strong and dynamic role in this re-orientation of the discipline. A computer can make hundreds of millions of calculations in a second. This opens up possibilities for trying things, and calculating things, and visualizing things, that were unthinkable fifty years ago. Of course it should be borne in mind that mathematical thinking involves concepts and reasoning, while a computer is a device for manipulating data. These two activities are quite different. It appears unlikely (see Roger Penrose's remarkable book *The Emperor's New Mind*) that a computer will ever be able to think, and to prove mathematical theorems, in the way that a human being performs these activities. Nonetheless, the computer can provide valuable information and insights. It can enable the user to see things that he/she would otherwise be unable to envision. It is a valuable tool. We shall certainly spend a good deal of time in this book pondering the role of the computer in modern human thought.

In endeavoring to understand the role of the computer in mathematical life, it is perhaps worth drawing an analogy with history. Tycho Brahe (1546–1601) was one of the great astronomers of the renaissance. Through painstaking scientific procedure, he recorded reams and reams of data about the motions of the planets. His gifted student Johannes Kepler was anxious to get his hands on Brahe's data, because he had ideas about formulating mathematical laws about the motions of the planets. But Brahe and Kepler were both strong-willed men. They could not see eye-to-eye on many things. And Brahe feared that Kepler would use his data to confirm the Copernican theory about the solar system (namely that the *sun*, not the earth, was the center of the system—a notion that ran counter to religious dogma). As a result, during Tycho Brahe's lifetime Kepler did not have access to Brahe's numbers.

But providence intervened in a strange way. Tycho Brahe had been given an island by his sponsor on which to build and run his observatory. As a result, Tycho was obliged to attend certain social functions—just to show his appreciation, and to report on his progress. At one such function, Tycho drank an excessive amount of beer, his bladder burst, and he died. Kepler was able to negotiate with Tycho Brahe's family to get the data that he so desperately needed. And thus the course of scientific history was forever altered.

Kepler did *not* use deductive reasoning, nor the axiomatic method, nor

the strategy of mathematical proof to derive his three laws of planetary motion. Instead he simply stared at the hundreds of pages of planetary data that Brahe had provided, and he performed myriad calculations. At around this same time John Napier (1550–1617) was developing his theory of logarithms. These are terrific calculational tools, and would have simplified Kepler's task immensely. But Kepler could not understand the derivation of logarithms, and so refused to use them. He did everything the hard way. Imagine what Kepler could have done with a computer!—but he probably would have refused to use one just because he didn't understand how the central processing unit worked.

In any event, we tell here of Kepler and Napier because the situation is perhaps a harbinger of modern agonizing over the use of computers in mathematics. There are those who argue that the computer can enable us to see things—both calculationally and visually—that we could not see before. And there are those who say that all those calculations are good and well, but they do not constitute a mathematical proof. Nonetheless it seems that the first can inform the second, and a productive symbiosis can be created. We shall discuss these matters in detail in the present book.

It is worthwhile at this juncture to enunciate Kepler's three very dramatic laws:

1. The orbit of each planet is in the shape of an ellipse. The sun is at one focus of that ellipse.

2. A line drawn from the center of the sun to the planet will sweep out area at a constant rate.

3. The square of the time for one full orbit is proportional to the cube of the length of the major axis of the elliptical orbit.

It was a few centuries later that Edmond Halley (1656–1742), one of Isaac Newton's (1642–1727) few friends, was conversing with him about various scientific issues. Halley asked the great scientist what must be the shape of the orbits of the planets, given Newton's seminal inverse-square law for gravitational attraction. Without hesitation, Newton replied, "Of course it is an ellipse." Halley was shocked. "But can you prove this?" queried Halley. Newton said that he had indeed derived a proof, but then he had thrown the notes away. Halley was beside himself. This was the problem that he and his collaborators had studied for a great many years with no progress. And

now the great Newton had solved the problem and then frivolously discarded the solution. Halley *insisted* that Newton reproduce the proof. Doing so required an enormous effort by Newton, and led in part to his writing of the celebrated *Principia*—perhaps the greatest scientific work ever written.

Now let us return to our consideration of changes that have come about in mathematics in the past thirty years, in part because of the advent of high-speed digital computers. Here is a litany of some of the components of this process:

a) In 1974 Appel and Haken [APH1] announced a proof of the 4-color conjecture. This is the question of how many colors are needed to color any map, so that adjacent countries are colored differently. Their proof used 1200 hours of computer time on a supercomputer at the University of Illinois. Mathematicians found this event puzzling, because this "proof" was not something that anyone could study or check. Or understand. To this day there does not exist a proof of the 4-color theorem that can be read and checked by humans.

b) Over time, people became more and more comfortable with the use of computers in proofs. In its early days, the theory of wavelets (for example) depended on the estimation of a certain constant—something that could only be done with a computer. De Branges's original proof of the Bieberbach conjecture [DEB2] seemed to depend on a result from special function theory that could only be verified with the aid of a computer (it was later discovered to be a result of Askey and Gasper that was proved in the traditional manner). Many results in turbulence theory, shallow-water waves, and other applied areas depend critically on computers. Airplane wings are designed with massive computer calculations—there is no other way to do it. There are many other examples.

c) There is a whole industry of people who use computers to search axiomatic systems for new true statements, and proofs thereof. Startling new results in projective geometry have been found, for instance, in this fashion. The important Robbins Conjecture in Boolean Algebra was established by this "computer search" technique.

d) The evolution of new teaching tools like the software `The Geometer's Sketchpad` has suggested to many—including Fields Medalist William Thurston—that traditional proofs may be set aside in favor of experimentation—the testing of thousands or millions of examples—on the computer.

Thus the use of the computer has truly re-oriented our view of what a proof might comprise. Again, the point is to convince someone else that something is true. There are evidently many different means of doing so.

Perhaps more interesting are some of the new social trends in mathematics and the resulting construction of nonstandard proofs (we shall discuss these in detail in the text that follows):

a) One of the great efforts of twentieth century mathematics has been the classification of the finite, simple groups. Daniel Gorenstein of Rutgers University was in some sense the lightning rod who orchestrated the effort. It is now considered to be complete. What is remarkable is that this is a single theorem that is the aggregate effort of many hundreds of mathematicians. The "proof" is in fact the union of hundreds of papers and tracts spanning more than 150 years. At the moment this proof comprises over 10,000 pages. It is still being organized and distilled down today. The final "proof for the record" will consist of several volumes. It is not clear that the living experts will survive long enough to see the fruition of this work.

b) Thomas Hales's resolution of the Kepler sphere-packing problem uses a great deal of computer calculation, much as with the 4-color theorem. It is particularly interesting that his proof supplants the earlier proof of Wu-Yi Hsiang that relied on spherical trigonometry and *no computer calculation whatever*. Hales allows that his "proof" cannot be checked in the usual fashion. He is organizing a worldwide group of volunteers called `FlySpeck` to engage in a checking procedure for his computer-based arguments. [The FlySpeck program of Thomas Hales derives its name from "FPK" which stands for "formal proof of Kepler".] In December, 2005, Dimkow and Bauer were able to certify a piece of Hales's computer code. That is the first step of `FlySpeck`.

Hales expects that the task will consume twenty years of work by scientists all over the world.

c) Grisha Perelman's "proof" of the Poincaré conjecture and the geometrization program of Thurston is currently in everyone's focus. In 2003, Perelman wrote three papers that describe how to use Richard Hamilton's theory of Ricci flows to carry out Thurston's idea (called the "geometrization program") of breaking up a 3-manifold into fundamental geometric pieces. One very important consequence of this result would be the fundamental Poincaré conjecture. Although Perelman's papers are vague and incomplete, they are full of imaginative and deep geometric ideas. This work set off a storm of activity and speculation about how the program might be assessed and validated. There are huge efforts now by John Lott and Bruce Kleiner (at the University of Michigan) and Gang Tian (Princeton) and John Morgan (Columbia) to complete the Hamilton/Perelman program and produce a *bona fide*, recorded proof that others can study and verify.

Kleiner and Lott's paper [KLL], which is 192 pages, has this avowed intent:

> The purpose of these notes is to provide the details that are missing in [40] and [41] [these are [PER1] and [PER2] in the present book], which contain Perelman's arguments for the Geometrization Conjecture.

It is not clear as of this writing that the world has accepted this contribution as a *bona fide* proof of the Geometrization Conjecture.

The Morgan/Tian book, comprising 473 pages, has been completed and submitted to the Clay Mathematics Institute (see [MOT]). It is now being considered for publication by the American Mathematical Society.

An independent effort by Cao and Zhu has resulted in a 334-page paper that is published in the *Asian Journal of Mathematics*.[2] The latter work purports to prove *both* the geometrization conjecture *and* the Poincaré conjecture.

[2]According to an article in *The New Yorker* [NAG], Fields Medalist S. T. Yau has attempted to minimize Perelman's contributions to the solution of the Poincaré conjecture and play up the significance of the Cao/Zhu work (Cao was Yau's student). This all seems to be part of a program to promote Chinese mathematics.

It will be some years before we can be sure that any of these works is correct.

d) In fact Thurston's geometrization program is a tale in itself. He announced in the early 1980s that he had this result on the structure of 3-manifolds, and he knew how to prove it. The classical Poincaré conjecture would be an easy corollary of Thurston's geometrization program. He wrote an extensive set of notes [THU3]—of book length—and these were made available to the world by the Princeton math department. For a nominal fee, the department would send a copy to anyone who requested it. These notes, entitled *The Geometry and Topology of Three-Manifolds* [THU3], were extremely exciting and enticing. But almost nobody believed that they actually gave a proof of the geometrization theorem. Thurston ultimately became disillusioned by the process, because he believed that what he had written constituted a valid proof. He expressed his aggravation in the article *On Proof and progress in mathematics* [THU1]. There remains a cadre of very strong mathematicians who work to flesh out Thurston's program. It is also the case that Thurston, along with Silvio Levy, has written a more formal book *Three-Dimensional Geometry and Topology* [THU2] that is the first of several volumes that will provide all the details of Thurston's ideas. This first volume is very important, and presents a number of seminal ideas in a profound and original way. In fact that book recently won the prestigious AMS Book Prize. But it is really only a prelude to the proof of the actual theorem. Subsequent volumes have yet to appear.

There are a number of other fascinating components of this development. In 1993, John Horgan published an article in *Scientific American* called *The Death of Proof?* [HOR1]. In it he declared that traditional mathematical proofs no longer had any role in our thinking. A part, but not all, of Horgan's message was that any question that mathematicians might ask today can be answered by computers. In addition, proofs today were too long and complicated for anyone to understand, and anyway we now have better ways of doing things (again, computers come to the fore). This author wrote a rebuttal to Horgan entitled *The immortality of proof* [KRA1]. Horgan's ideas are no longer taken seriously—at least in the mathematics community.

John von Neumann (1903–1957) did not live to see the great diversification of mathematics that has taken place in the past few decades (although he *did* invent the stored-program computer). But he had concerns about the balkanization of the subject:

> I think that it is a relatively good approximation to truth—which is much too complicated to allow anything but approximations—that mathematical ideas originate in empirics, although the genealogy is sometimes long and obscure. But once they are so conceived, the subject begins to live a peculiar life of its own and is better compared to a creative one, governed by almost entirely aesthetical motivations, than to anything else and, in particular, to an empirical science. ... But there is a grave danger that the subject will develop along the line of least resistance, that the stream, so far from its source, will separate into a multitude of insignificant branches, and that the discipline will become a disorganised mass of details and complexities. In other words, at a great distance from its empirical source, or after much "abstract" inbreeding, a mathematical subject is in danger of degeneration. At the inception the style is usually classical; when it shows signs of becoming baroque, then the danger signal is up.

Traditional mathematics is unique among the sciences in that it uses a rigorous mode of discourse for establishing beyond any doubt that certain statements are true and correct. The development and refinement of that mode of discourse is one of the triumphs of the subject. But this great achievement is now being re-examined from a number of different points of view. As a result, new methods of proof are emerging; also the traditional modes of proof are being re-assessed, transmogrified, and developed. The upshot is a rich and varied tapestry of scientific methods that is re-shaping the subject of mathematics.

The purpose of this book is to explore all the ideas and developments outlined above. Along the way, we are able to acquaint the reader with the culture of mathematics: who mathematicians are, what they care about, and what they do. We also give indications of why mathematics is important, and why it is having such a powerful influence in the world today. We hope that, by reading this book, the reader will become acquainted with, and perhaps charmed by, the glory of this ancient subject. And will realized that there is so much more to learn.

Steven G. Krantz
St. Louis, Missouri

Acknowledgements

One of the pleasures of the writing life is getting criticism and input from other scholars. I thank John Bland, Robert Burckel, E. Brian Davies, Keith Devlin, Ed Dunne, Michael Eastwood, Jerry Folland, Jeremy Gray, Barry Mazur, Robert Strichartz, James Walker, and Doron Zeilberger for careful readings of my drafts and for contributing much wisdom and useful information. I thank Ed Dunne of the American Mathematical Society for many insights and for suggesting the topic of this book.

Ann Kostant of Birkhäuser was, as always, a proactive and supportive editor. She originally invited me to write a book for the Copernicus series, and gave terrific advice and encouragement during its development.

Chapter 0

What is a Proof and Why?

The proof of the pudding is in the eating.

Miguel Cervantes

By the work one knows the workman.

Jean de La Fontaine

In mathematics there are no true controversies.

Carl Friedrich Gauss

Logic is the art of going wrong with confidence.

Anonymous

It seems clear that mathematicians will have difficulty escaping from the Kantian fold. Even a Platonist must concede that mathematics is only accessible through the human mind, and thus all mathematics might be considered a Kantian experiment. We can debate whether Euclidean geometry is but an idealization of the geometry of nature (where a point has no length or breadth and a line has length but no breadth), or nature an imperfect reflection of "pure" geometrical objects, but in either case the objects of interest lie within the mind's eye.

J. Borwein, P. Borwein, R. Girgensohn, and S. Parnes

To test man, the proofs shift.

Robert Browning.

Newton was a most fortunate man because there is just one universe and Newton had discovered its laws.

Pierre-Simon Laplace

Had, having, and in quest to have, extreme; A bliss in proof,—and prov'd, a very woe; Before, a joy propos'd; behind, a dream.

William Shakespeare

The posing of conjectures is the most obvious mathematical activity which does not involve a proof. Conjectures range from brilliant to boring, from impossible to obvious. They are filtered by the interest they inspire, rather than by editors and referees. The better ones have inspired the development

of whole fields.

Arthur Jaffe and Frank Quinn

Conjecture has long been accepted and honored in mathematics, but the customs are clear. If a mathematician has really studied the subject and made advances therein, then he is entitled to formulate an insight as a conjecture, which usually has the form of a specific proposed theorem. . . . But the next step must be proof and not more speculation.

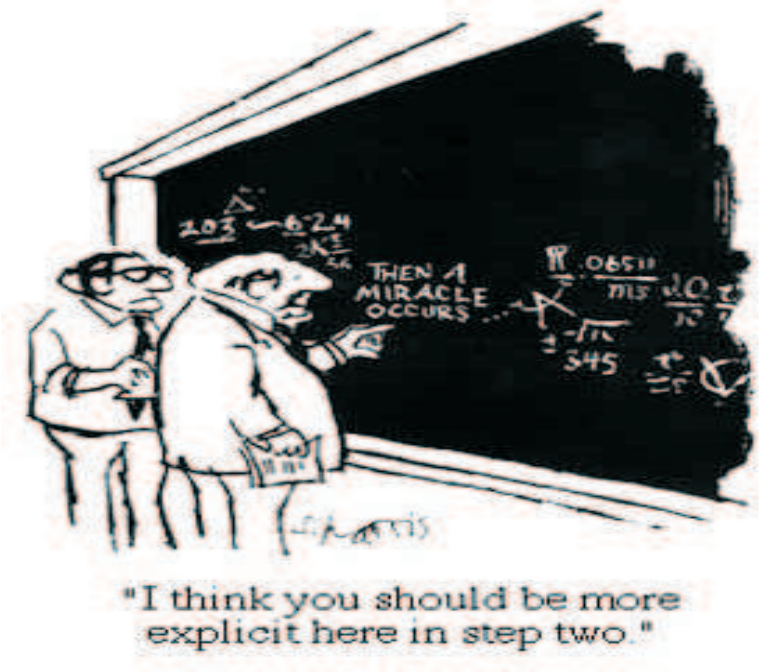
Saunders Mac Lane

0.1 What is a Mathematician?

A well-meaning mother was once heard to tell her child that a mathematician is someone who does "scientific arithmetic". Others think that a mathematician is someone who spends all day hacking away at a computer.

Neither of these contentions is incorrect, but they do not begin to penetrate all that a mathematician really is. Paraphrasing Keith Devlin, we note that a mathematician is someone who:

- Observes and interprets phenomena.
- Analyzes scientific events and information.
- Formulates concepts.
- Generalizes concepts.
- Performs inductive reasoning.
- Performs analogical reasoning.
- Engages in trial and error (and evaluation).
- Models ideas and phenomena.
- Formulates problems.
- Abstracts from problems.
- Solves problems.



Sidney Harris Cartoon I

- Uses computation to draw analytical conclusions.
- Makes deductions.
- Makes guesses.
- Proves theorems.

And even this list is incomplete. A mathematician is a master of critical thinking, of analysis, and of deductive reasoning. These skills travel well, and can be applied in a large variety of situations—and in many different disciplines. Today, mathematical skills are being put to good use in medicine, physics, law, commerce, Internet design, engineering, chemistry, biological science, social science, anthropology, genetics, warfare, cryptography, plastic surgery, security analysis, data manipulation, computer science, and in many other disciplines and endeavors as well.

Certainly one of the astonishing and dramatic new uses of mathematics that has come about in the past twenty years is in finance. The Nobel-Prize-winning work of Fisher Black of Harvard and Myron Scholes of Stanford

has given rise to the first ever method for option pricing. This methodology is based on the theory of stochastic integrals—a part of abstract probability theory. Investment firms all over the world now routinely employ Ph.D. mathematicians. When we teach a measure theory course in the Math Department—something that was formerly the exclusive province of graduate students in mathematics studying for the Qualifying Exams—we find that the class is unusually large, and most of the students are from Economics and Finance.

Another part of the modern world that has been strongly influenced by mathematics, and which employs a goodly number of mathematicians with advanced training, is genetics and the Genome Project. Most people do not realize that a strand of DNA can have billions of gene sites on it. Matching up genetic markers is *not* like matching up your socks; in fact things must be done probabilistically. A good deal of statistical theory is used. Thus many Ph.D. mathematicians work on the genome project.

The focus of the present book is on the concept of *mathematical proof*. Although it is safe to say that most mathematical scientists do not[1] spend the bulk of their time proving theorems, it is nevertheless the case that *proof* is the *lingua franca* of mathematics. It is the web that holds the enterprise together. It is what makes the subject travel well, and guarantees that mathematical ideas will have some immortality.

There is no other scientific or analytical discipline that uses proof as readily and routinely as does mathematics. This is the device that makes theoretical mathematics special: the tightly knit chain of reasoning, following strict logical rules, that leads inexorably to a particular conclusion. It is *proof* that is our device for establishing the absolute and irrevocable truth of statements in our subject. This is the reason that we can depend on mathematics that was done by Euclid 2300 years ago as readily as we believe in the mathematics that is done today. No other discipline can make such an assertion (but see Section 0.9).

This book will acquaint the reader with who mathematicians are and what they do, using the concept of "proof" as a touchstone. Along the way, we will become acquainted with foibles and traits of particular mathematicians, and of the profession as a whole. It is an exciting journey, full of rewards

[1]This is because a great many mathematical scientists do not work at universities. They instead work for the National Security Agency (NSA), or the National Aeronautics and Space Administration (NASA), or Hughes Aircraft, or Microsoft.

and surprises.

0.2 The Concept of Proof

It is well to begin this discussion with an inspiring quotation from master mathematician Michael Atiyah (1929–) [ATI2]:

> We all know what we like in music, painting or poetry, but it is much harder to explain why we like it. The same is true in mathematics, which is, in part, an art form. We can identify a long list of desirable qualities: beauty, elegance, importance, originality, usefulness, depth, breadth, brevity, simplicity, clarity. However, a single work can hardly embody them all; in fact, some are mutually incompatible. Just as different qualities are appropriate in sonatas, quartets or symphonies, so mathematical compositions of varying types require different treatment. Architecture also provides a useful analogy. A cathedral, palace or castle calls for a very different treatment from an office block or private home. A building appeals to us because it has the right mix of attractive qualities for its purpose, but in the end, our aesthetic response is instinctive and subjective. The best critics frequently disagree.

The tradition of mathematics is a long and glorious one. Along with philosophy, it is the oldest venue of human intellectual inquiry. It is in the nature of the human condition to want to understand the world around us, and mathematics is a natural vehicle for doing so. But, for the ancients, mathematics was also a subject that was beautiful and worthwhile in its own right. A scholarly pursuit that had intrinsic merit and aesthetic appeal, mathematics was certainly worth studying for its own sake.

In its earliest days, mathematics was often bound up with practical questions. The Egyptians, as well as the Greeks, were concerned with surveying land. Refer to Figure 0.1 Thus it was natural to consider questions of geometry and trigonometry. Certainly triangles and rectangles came up in a natural way in this context, so early geometry concentrated on these constructs. Circles, too, were natural to consider—for the design of arenas and water tanks and other practical projects. So ancient geometry (and Euclid's axioms for geometry) discussed circles.

Sidney Harris Cartoon II

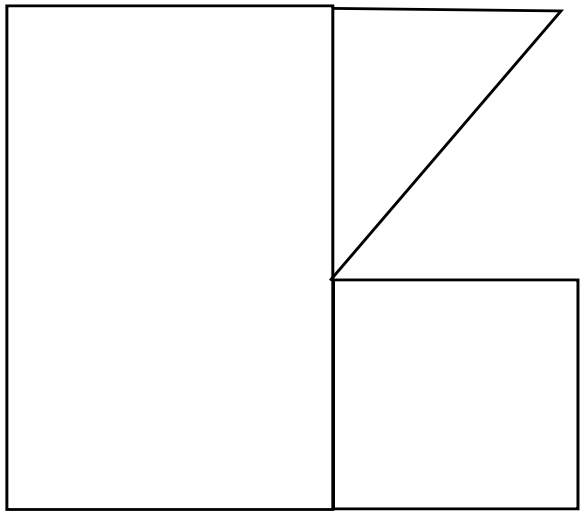


Figure 0.1

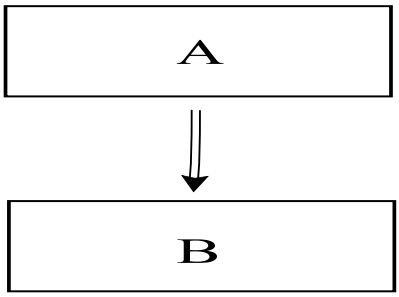


Figure 0.2

The earliest mathematics was phenomenological. If one could draw a plausible picture, then that was all the justification that was needed for a mathematical "fact". Sometimes one argued by analogy. Or by invoking the gods. The notion that mathematical statements could be *proved* was not yet an idea that had been developed. There was no standard for the concept of proof. The logical structure, the "rules of the game", had not yet been created. If one ancient Egyptian were to say to another, "I don't understand why this mathematical statement is true. Please prove it.", his request would have fallen on deaf ears. The concept of proof was not part of the working vocabulary of the ancient mathematician.

Well, what is a proof? Heuristically, a proof is a rhetorical device for convincing someone else that a mathematical statement is true or valid. And how might one do this? A moment's thought suggests that a natural way to prove that something new (call it **B**) is true is to relate it to something old (call it **A**) that has already been accepted as true. Thus arises the concept of *deriving* a new result from an old result. See Figure 0.2. The next question then is, "How was the old result verified?" Applying this regimen repeatedly, we find ourselves considering a chain of reasoning as in Figure 0.3. But then one cannot help but ask: "Where does the chain begin?" And this is a fundamental issue.

It will not do to say that the chain has no beginning: it extends infinitely far back into the fogs of time. Because if that were the case it would undercut our thinking of what a proof should be. We are endeavoring to justify new mathematical facts in terms of old mathematical facts. But if the reasoning regresses infinitely far back into the past, then we cannot in fact ever grasp a basis or initial justification for our reasoning.

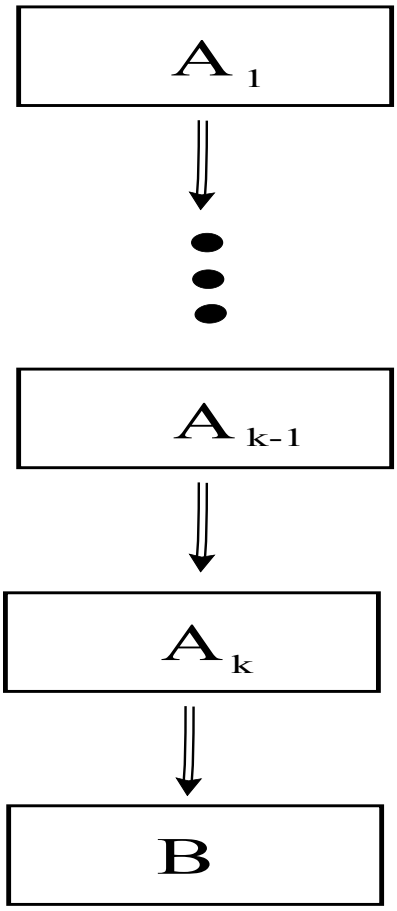


Figure 0.3

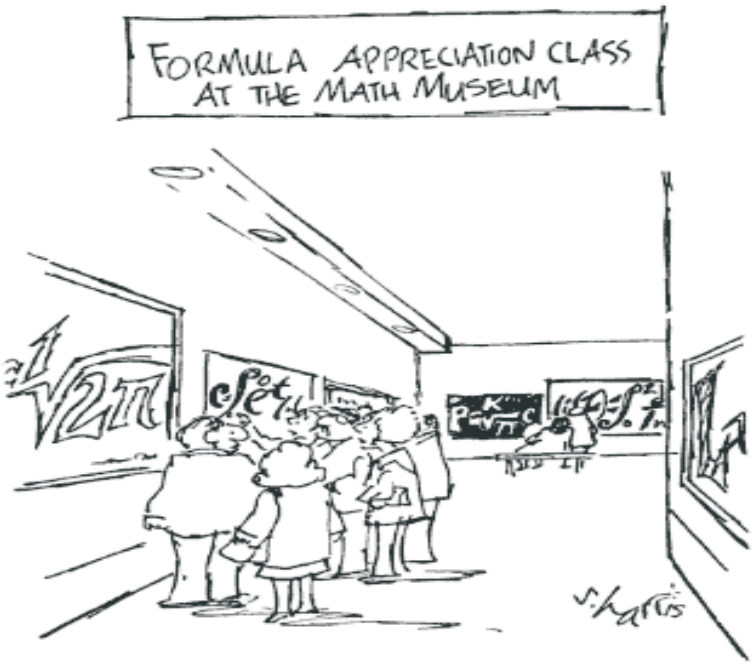
As a result of these questions, ancient mathematicians had to think hard about the nature of mathematical proof. Thales (640 B.C.E.–546 B.C.E.), Eudoxus (408 B.C.E.–355 B.C.E.), and Theaetetus of Athens (417 B.C.E.–369 B.C.E.) actually formulated theorems. Thales definitely proved some theorems in geometry (and these were later put into a broader context by Euclid). A theorem is the mathematician's formal enunciation of a fact or truth. But Eudoxus fell short in finding means to prove his theorems. His work had a distinctly practical bent, and he was particularly fond of calculations.

It was Euclid of Alexandria who first formalized the way that we now think about mathematics. Euclid had definitions and axioms and then theorems—in that order. There is no gainsaying the assertion that Euclid set the paradigm by which we have been practicing mathematics for 2300 years. This was mathematics done right. Now, following Euclid, in order to address the issue of the infinitely regressing chain of reasoning, we begin our studies by putting into place a set of *Definitions* and a set of *Axioms*.

What is a definition? A definition explains the meaning of a piece of terminology. There are logical problems with even this simple idea, for consider the first definition that we are going to formulate. Suppose that we wish to define a *rectangle*. This will be the first piece of terminology in our mathematical system. What words can we use to define it? Suppose that we define rectangle in terms of points and lines and planes. That begs the questions: What is a point? What is a line? What is a plane?

Thus we see that our *first* definition(s) must be formulated in terms of commonly accepted words that require no further explanation. It was Aristotle (384 B.C.E.–322 B.C.E.) who insisted that a definition must describe the concept being defined in terms of other concepts already known. This is often quite difficult. As an example, Euclid defined a *point* to be that which has no part. Thus he is using words *outside of mathematics*, that are a commonly accepted part of everyday argot, to explain the precise mathematical notion of "point".[2] Once "point" is defined, then one can use that term in later definitions. And one will also use everyday language that does not require further explication. That is how we build up our system of definitions.

[2]It is quite common, among those who study the foundations of mathematics, to refer to terms that are defined in non-mathematical language—that is, which cannot be defined in terms of other mathematical terms—as *undefined terms*. The concept of "set", which is discussed elsewhere in this book, is an undefined term. So is "point".



Sidney Harris Cartoon III

The definitions give us then a language for doing mathematics. We formulate our results, or *theorems*, by using the words that have been established in the definitions. But wait, we are not yet ready for theorems. Because we have to lay cornerstones upon which our reasoning can develop. That is the purpose of axioms.

What is an axiom? An axiom[3] (or postulate[4]) is a mathematical statement of fact, formulated using the terminology that has been defined in the definitions, that is taken to be self-evident. An axiom embodies a crisp, clean mathematical assertion. One does not *prove* an axiom. One takes the axiom to be given, and to be so obvious and plausible that no proof is required.

One of the most famous axioms in all of mathematics is the *Parallel Postulate* of Euclid. The Parallel Postulate (in Playfair's formulation) asserts that if P is a point, and if ℓ is a line not passing through that point, then there is a second line ℓ' passing through P that is parallel to ℓ. See Figure 0.4. The Parallel Postulate is part of Euclid's geometry, so it is 2300 years old. And people wondered for over 2000 years whether this assertion should actually be an axiom. Perhaps it could be proved from the other four axioms of geometry (see Section 1.2 for a detailed treatment of Euclid's axioms). There were mighty struggles to provide such a proof, and many famous mistakes made (see [GRE] for some of the history). But, in 1826, Janos Bolyai and Nikolai Lobachevsky showed that the Parallel Postulate can never be proved. There are models for geometry in which all the other axioms of Euclid are true yet the Parallel Postulate is false. So the Parallel Postulate now stands as one of the axioms of our most commonly used geometry.

Generally speaking, in any subject area of mathematics, one begins with a brief list of definitions and a brief list of axioms. Once these are in place, and are accepted and understood, then one can begin proving theorems. And what is a proof? A proof is a rhetorical device for convincing another mathematician that a given statement (the theorem) is true. Thus a proof can take many different forms. The most traditional form of mathematical proof is that it is a tightly knit sequence of statements linked together by strict rules of logic. But the purpose of the present book is to discuss and consider what other forms a proof might take. Today, a proof could (and often does) take the traditional form that goes back 2300 years to the time of Euclid. But it

[3]The word "axiom" derives from a Greek word meaning "something worthy".

[4]The word "postulate" derives from a medieval Latin word meaning "to nominate" or "to demand".

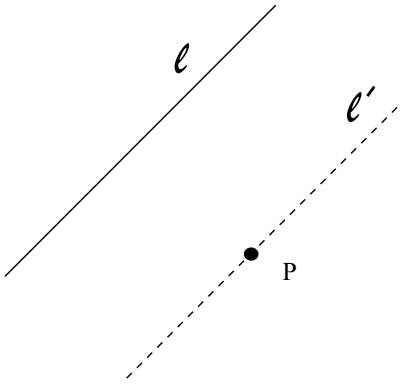


Figure 0.4

could also consist of a computer calculation. Or it could consist of constructing a physical model. Or it could consist of a computer *simulation* or *model*. Or it could consist of a computer algebra computation using `Mathematica` or `Maple` or `MatLab`. It could also consist of an agglomeration of these various techniques.

One of the main purposes of the present book is to present and examine all these many forms of mathematical proof, and the role that they play in modern mathematics. In spite of numerous changes and developments in the way that we view the technique of proof, this fundamental methodology remains a cornerstone of the infrastructure of mathematical reasoning. As we have indicated, a key part of any proof—no matter what form it may take—is logic. And what is logic? That is the subject of the next section.

The philosopher Karl Popper believed [POP] that *nothing* can ever be known with absolute certainty. He instead had a concept of "truth up to falsifiability". Mathematics, in its traditional modality, rejects this point of view. Mathematical assertions which are proved according to the accepted canons of mathematical reasoning are believed to be irrefutably true. And they will continue to be true. It is this immutable nature of mathematics that makes it unique among the human intellectual pursuits.

0.3 The Foundations of Logic

Today mathematical logic is a subject unto itself. It is a full-blown branch of mathematics, just like geometry or differential equations or algebra. But,

Sidney Harris Cartoon IV

for the purposes of practicing mathematicians, logic is a brief and accessible set of rules by which we live our lives.

The father of logic as we know it today was Aristotle (384 B.C.E.–322 B.C.E.). His *Organon* laid the foundations of what logic should be. Let us consider here what some of Aristotle's precepts were.

0.3.1 The Law of the Excluded Middle

One of Aristotle's rules of logic was that every sensible statement, that is clear and succinct and does not contain logical contradictions, is either true or false. There is no "middle ground" or "undecided status" for such a statement. Thus the assertion

> If there is life as we know it on Mars, then fish can fly.

is either true or false. The statement may seem frivolous. It may appear to be silly. There is no way to verify it, because we do not know (and we will not know any time soon) whether there is life as we know it on Mars.

But the statement makes perfect sense. So it must be true or false. We *do know* that fish cannot fly. But we cannot determine the truth or falsity of this statement because we do not know whether there is life as we know it on Mars.

You might be thinking, "Professor Krantz, that analysis is not correct. The correct truth value to assign to this sentence is 'Undecided'. We do not know about life on Mars so we cannot decide whether this statement is true. Perhaps in a couple of centuries we will have a better idea and we can assign a valid truth value to the sentence. But we cannot do so now. Thus the vote is 'undecided'."

Interesting reasoning, but this is not the point of view that we take in mathematics. Instead, our reasoning is that *God* knows everything—he certainly knows whether there is life as we know it on Mars—therefore *he* certainly knows whether the statement is true or false. The fact that we do *not* know is an unfortunate artifact of our mortality. But it does not change the basic fact that *This sentence is either true or false.* Period.

It may be worth noting that there *are* versions of logic which allow for a multi-valued truth function. Thus a statement is not merely assigned one of two truth values (i.e., "true" or "false"). Other truth values are allowed. For example, the statement "George W. Bush is President" is true right now, but will not be true in perpetuity. So we could have a truth value to indicate a transient truth. The book [KRA4] discusses multi-valued logics. Traditional mathematics uses a logic with only two truth values: *true* and *false*. Thus traditional mathematics rejects the notion that a sensible statement can have any undecided, or any transient, truth status.

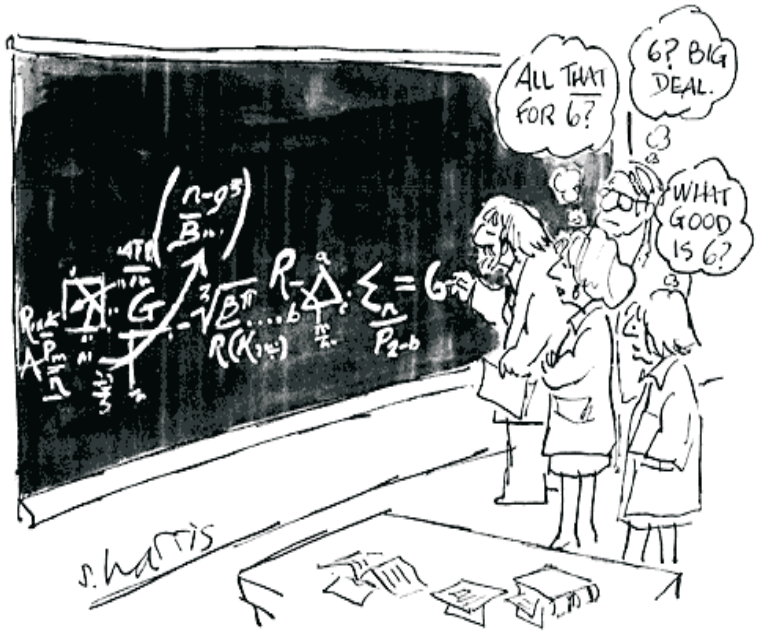
0.3.2 *Modus Ponendo Ponens* and Friends

The name *modus ponendo ponens*[5] is commonly applied to the most fundamental rule of logical reasoning. It says that if we know that "**A** implies **B**" and if we know "**A**" then we may conclude "**B**". This is most commonly summarized (using the standard logical notation of $\Rightarrow$ for "implies" and $\wedge$ for "and") as

$$\Big[(\mathbf{A} \Rightarrow \mathbf{B}) \ \wedge \ \mathbf{A}\Big] \ \Rightarrow \ \mathbf{B}\,.$$

We commonly use this mode of reasoning in everyday discourse. Unfortu-

[5]The translation of this Latin phrase is "mode that affirms".



Sidney Harris Cartoon V

nately, we also very frequently *mis*-use it. How often have you heard someone reason as follows?

- All communists eat breakfast.
- My distinguished opponent eats breakfast.
- Therefore my distinguished opponent is a communist.

You may laugh, but one encounters this type of thinking on news broadcasts, in the newspaper, and in everyday discourse on a regular basis. As an object lesson, let us do a logical analysis of this specious reasoning.

Let[6]

$$\mathbf{A}(x) \equiv x \text{ is a communist}\,,$$

$$\mathbf{B}(x) \equiv x \text{ eats breakfast}\,,$$

$$o \equiv \text{ my distinguished opponent}\,.$$

[6]Here we use the convenient mathematical notation $\equiv$ to mean "is defined to be."

Then we may diagram the reasoning above as

$$\begin{aligned} &\mathbf{A}(x) \;\Rightarrow\; \mathbf{B}(x) \\ &\mathbf{B}(o) \\ &\text{therefore } \mathbf{A}(o)\,. \end{aligned}$$

You can see now that we are misusing *modus ponendo ponens*. We have $\mathbf{A} \Rightarrow \mathbf{B}$ and $\mathbf{B}$ and we are concluding $\mathbf{A}$.

This is the very common error of confusing the converse with the contrapositive. Let us discuss this fundamental issue. If $\mathbf{A} \;\Rightarrow\; \mathbf{B}$ is a given implication then its *converse* is the implication $\mathbf{B} \;\Rightarrow\; \mathbf{A}$. Its *contrapositive* is the statement $\sim \mathbf{B} \;\Rightarrow \sim \mathbf{A}$, where $\sim$ stands for "not". One sometimes encounters the word "converse" in everyday conversation but rarely the word "contrapositive". So these concepts bear some discussion.

Consider the statement

Every healthy horse has four legs.

It is convenient to first rephrase this more simply as

A healthy horse has four legs.

If we let

$$\mathbf{A}(x) \equiv x \text{ is a healthy horse}\,,$$

$$\mathbf{B}(x) \equiv x \text{ has four legs}\,,$$

then the displayed sentence says

$$\mathbf{A}(x) \;\Rightarrow\; \mathbf{B}(x)\,.$$

The *converse* of this statement is

$$\mathbf{B}(x) \;\Rightarrow\; \mathbf{A}(x)\,,$$

or

An object with four legs is a healthy horse.

It is not difficult to see that the converse is a logically distinct statement from the original one that each healthy horse has four legs. And, whereas the original statement is true, the converse statement is *false*. It is *not* generally the case that a thing with four legs is a healthy horse. For example, most tables have four legs. But a table is not a healthy horse. A sheep has four legs. But a sheep is not a healthy horse.

The contrapositive is a different matter. The *contrapositive* of our statement is

$$\sim \mathbf{B}(x) \ \Rightarrow \sim \mathbf{A}(x)\,,$$

or

A thing that does not have four legs is not a healthy horse.

This is a different statement from the original sentence. But it is in fact *true*. If I encounter an object that does not have four legs, then I can be sure that it is *not* a healthy horse, because in fact any healthy horse has four legs. A moment's thought reveals that in fact this contrapositive statement is saying precisely the same thing as the original statement—just using slightly different language.

In fact *it is always the case that the contrapositive of an implication is logically equivalent with the original implication.* Such is not the case for the converse.

Let us return now to the discussion of whether or not our distinguished opponent is a communist. We began with $\mathbf{A} \ \Rightarrow \ \mathbf{B}$ and with $\mathbf{B}$ and we concluded $\mathbf{A}$. Thus we were misreading the implication as $\mathbf{B} \ \Rightarrow \ \mathbf{A}$. In other words, we were *misinterpreting the original implication as its converse.* It would be correct to interpret the original implication as $\sim \mathbf{B} \ \Rightarrow \sim \mathbf{A}$, because that is the contrapositive and is logically equivalent with the original implication. But of course $\sim \mathbf{B} \ \Rightarrow \sim \mathbf{A}$ and $\mathbf{B}$ taken together do not imply anything.

The logical rule *modus tollens*[7] is in fact nothing other than a restatement of *modus ponendo ponens*. It says that

$$\text{If } [(\mathbf{A} \ \Rightarrow \ \mathbf{B}) \text{ and } \sim \mathbf{B}], \text{ then } \sim \mathbf{A}.$$

Given the discussion we have had thus far, *modus tollens* is not difficult to understand. For $\mathbf{A} \ \Rightarrow \ \mathbf{B}$ is logically equivalent with its contrapositive

[7]The translation of this Latin phrase is "mode that denies."

Sidney Harris Cartoon VI

$\sim \mathbf{B} \Rightarrow \sim \mathbf{A}$. And if we also have $\sim \mathbf{B}$ then of course we may conclude (by *modus ponendo ponens*!) the statement $\sim \mathbf{A}$. That is what *modus tollens* says.

0.4 What Does a Proof Consist Of?

Most of the steps of a mathematical proof are applications of *modus ponendo ponens* or *modus tollens*. This is a slight oversimplification, as there are a great many proof techniques that have been developed over the past two centuries (see Chapter 11 for detailed discussion of some of these). These include proof by mathematical induction, proof by contradiction, proof by exhaustion, proof by enumeration, and many others. But they are all built on *modus ponendo ponens*.

It is really an elegant and powerful system. *Occam's Razor* is a logical principle posited in the fourteenth century (by William of Occam (1288 C.E.–1348 C.E.)) which advocates that your proof system should have the smallest possible set of axioms and logical rules. That way you minimize the possibility that there are internal contradictions built into the system, and also you make it easier to find the source of your ideas. Inspired both by Euclid's *Elements* and by Occam's Razor, mathematics has striven for all of modern time to keep the fundamentals of its subject as streamlined and elegant as possible. We want our list of definitions to be as short as possible, and we want our collection of axioms or postulates to be as concise and elegant as possible. If you open up a classic text on group theory—such as Marshall Hall's masterpiece [HAL], you will find that there are just three axioms on the first page. The entire 434-page book is built on just those three axioms.[8] Or instead have a look at Walter Rudin's classic *Principles of Mathematical Analysis* [RUD]. There the subject of real variables is built on just twelve axioms. Or look at a foundational book on set theory like Suppes [SUP] or Hrbacek and Jech [HRJ]. There we see the entire subject built on eight axioms.

[8] In fact there has recently been found a way to enunciate the premises of group theory using just *one* axiom, and *not* using the word "and". References for this work are [KUN], [HIN], and [MCC].

0.5 The Purpose of Proof

The experimental sciences (physics, biology, chemistry, for example) tend to use laboratory experiments or tests to check and verify assertions. The benchmark in these subjects is the *reproducible experiment with control*. In their published papers, these scientists will briefly describe what they have discovered, and how they carried out the steps of the corresponding experiment. They will describe the *control*, which is the standard against which the experimental results are compared. Those scientists who are interested can, on reading the article, then turn around and replicate the experiment in their own labs. The really classic, and fundamental and important, experiments become classroom material and are reproduced by students all over the world. Most experimental science is *not* derived from fundamental principles (like axioms). The intellectual process is more empirical, and the verification procedure is correspondingly practical and direct.

Theoretical physics is a bit different. Scientists like Stephen Hawking and Edward Witten and Roger Penrose never set foot in a laboratory. They just *think* about physics. They rely on the experimentalists to give them fodder for their ideas. The experimentalists will also help them to test their ideas. But they themselves do not engage in the benchmarking procedure.[9]

This process works reasonably well for theoretical physics, but not always. It took a great many years to find experimental evidence to support Einstein's general theory of relativity. The theory of *strings*, which is a fairly new set of ideas that may replace the classical atomic theory,[10] has been an exciting and fundamental part of physics now for over twenty years. But there is *no* experimental evidence to support any of the ideas of string theory. This is quite similar to the state in which general relativity sat for a good many years. String theory is a set of ideas waiting to be verified.

Mathematics is quite a different sort of intellectual enterprise. In mathematics we set our definitions and axioms in place *before* we do anything else. In particular, *before we endeavor to derive any results* we must engage in a certain amount of preparatory work. Then we give precise, elegant formula-

[9] The fascinating article [JAQ] considers whether, like physicists, mathematicians should be divided into "theorists" and "experimentalists".

[10] String theory is quite "far out." This new set of ideas is supposed to describe the fundamental composition of nature in the three-dimensional space in which we live. It provides an explanation for the provenance of gravity. But in fact a string lives in either ten- or twenty-six-dimensional space!

tions of statements and we prove them. Any statement in mathematics which lacks a proof has no currency.[11] Nobody will take it as valid. And nobody will use it in his/her own work. The proof is the final test of any new idea. And, once a proof is in place, that is the end of the discussion. Nobody will ever find a counterexample, nor ever gainsay that particular mathematical fact.

It should not be thought that the generation of mathematical proofs is mechanical. Far from it. A mathematician discovers ideas *intuitively*—just like anyone else. He or she will just "see" that something is true. This will be based on experience and insight developed over many years. Then the mathematician will have to think about why this new "fact" is true. At first, just the sketch of the idea of a proof may be jotted down. Over a long period of time, additional ideas will be generated and pieces of the proof will thereby be assembled. In the end, all the details will be filled in and a *bona fide*, rigorous proof, following the strict linear dictates of logic, will be the result.

Throughout history, there have been areas of mathematics where we could not decide what a proof was. We had neither the language nor the notation nor the concepts to write *anything* down rigorously. Probability suffered many missteps, and was fraught with conundrums and paradoxes, for hundreds of years until Andrei Kolmogorov (1903–1987) realized in the 1930s that measure theory was the right tool for describing probabilistic ideas. In the 1930s the Italian algebraic geometers used to decide what was a "theorem" by getting together in a group, discussing the matter, and then taking a vote. In fact it was worse than that. There was considerable, intense rivalry among Federigo Enriques (1871–1946), Guido Castelnuovo (1865–1952), and others. They would prove new results and announce them, but refuse to show anyone the proofs. Thus the Italian mathematicians would discuss the matter, offer speculations about how the new results could be proved, and then take a vote. One upshot of this feuding is that there was no verification process; there was no development of technique. The subject stagnated. The (currently-believed-to-be) correct tools for doing algebraic geometry were developed many years later by André Weil (1906–1998), Alexandre Grothendieck (1928–), Oskar Zariski (1899–1986), Jean-Pierre Serre (1926–), Claude Chevalley (1909–1984), and many others.

[11] But mathematicians certainly engage in heuristics, in algorithmic reasoning, and in conjectures. These are discussed elsewhere in the present book.

Michael Rabin has introduced a new twist into our subject by producing (in 1976) a *probabilistic proof* of a mathematical theorem. That is to say, he studied whether a certain large number p was prime. He devised an iterative procedure with the property that each application of the algorithm increased the probability that the number was prime. So, with enough iterations, you could make the probability as close to 1 as you wished. But you could never use his method to achieve mathematical certainty (unless the answer was "no"). Following in the same vein, Robert Solovay and Volker Strassen, in 1977, found a probabilistic means of examining the primality of large integers.

Mathematics (traditionally) is equipped with a sort of certainty that other sciences do not possess.[12] Mathematics lives completely inside a logical system that *we have created*. We have endowed the system with a reliability and reproducibility and portability that no other science can hope for.[13]

One of the intoxicating features of mathematical proof is its compelling, indeed its intoxicating, nature. The biographer John Aubrey tells of the philosopher Thomas Hobbes's (1588–1679) first encounter with the phenomenon:

> He was 40 yeares old before he looked on geometry; which happened accidentally. Being in a gentleman's library in . . . , Euclid's *Elements* lay open, and 'twas the 47 *El. libri I*. He read the proposition. "By G—," sayd he (he would now and then sweare, by way of emphasis), "this is impossible!" So he reads the demonstration of it, which referred him back to such a proposition; which proposition he read. That referred him back to another, which he also read. *Et sic deinceps*, that at last he was demonstratively convinced of that trueth. This made him in love with geometry.

In fact Hobbes was so taken with mathematics that he ended up adopting the mathematical methodology in his philosophy. He essayed to formulate a mathematical theory of ethics, so that one could make more decisions simply by solving an equation. This quest was less than successful.

[12]There are those who will argue that a nice, tight scientific computer program has the same sort of certainty as mathematics. This is certainly true in the sense that it is dependable, verifiable, and yields a predictable result.

[13]Vijay Sahni was in 1919 charged with blasphemy under an obscure law in New Jersey. His offense, evidently, was to contend that mathematical certainty can be applied to other aspects of human knowledge, including religion. This event was uncovered 61 years later by his grandson, who was taking a course on on infinity at Stanford University.

Another special feature of mathematics is its timelessness. The theorems that Euclid and Pythagoras proved 2500 years ago are still valid today; and we use them with confidence because we know that they are just as true today as they were when those great masters first discovered them. Other sciences are quite different. The medical or computer science literature of even three years ago is considered to be virtually useless. Because what people thought was correct a few years ago has already changed and migrated and transmogrified. Mathematics, by contrast, is here forever.

What is marvelous is that, in spite of the appearance of some artificiality in this process, mathematics provides beautiful models for nature (see the lovely essay [WIG], which discusses this point). Over and over again, and more with each passing year, mathematics has helped to explain how the world around us works. Just a few examples illustrate the point:

- Isaac Newton derived Kepler's three laws of planetary motion from just his universal law of gravitation and calculus.
- There is a complete mathematical theory of the refraction of light (due to Isaac Newton, Willebrord Snell, and Pierre de Fermat).
- There is a mathematical theory of the propagation of heat.
- There is a mathematical theory of electromagnetic waves.
- All of classical field theory from physics is formulated in terms of mathematics.
- Einstein's field equations are analyzed using mathematics.
- The motion of falling bodies and projectiles is completely analyzable with mathematics.
- The technology for locating distant submarines using radar and sonar waves is all founded in mathematics.
- The theory of image processing and image compression is all founded in mathematics.
- The design of music CDs is all based on Fourier analysis and coding theory, both branches of mathematics.

The list could go on and on.

The key point to be understood here is that *proof* is central to what modern mathematics is about, and what makes it reliable and reproducible. No other science depends on proof, and therefore no other science has the bulletproof solidity of mathematics (see also Section 0.9). But mathematics is *applied* in a variety of ways, in a vast panorama of disciplines. And the applications are many and varied. Other disciplines often like to reduce their theories to mathematics because it gives the subject a certain elegance and solidity—and it looks really jazzy.

There are two aspects of proof to be borne in mind. One is that it is our *lingua franca*. It is the mathematical mode of discourse. It is our tried-and-true methodology for recording discoveries in a bullet-proof fashion that will stand the test of time. The second, and for the working mathematician the most important, aspect of proof is that the proof of a new theorem explains *why* the result is true. In the end what we seek is new understanding, and "proof" provides us with that golden nugget. See [BRE] for a delightful discussion of these ideas.

An engineer may use mathematics in a heuristic fashion to get the results he/she needs. A physicist may use approximations to achieve his/her goals. The people who apply mathematics do not, generally speaking, prove theorems.[14] But they make use of mathematical ideas. And they know that they can depend on mathematics because of the subject's internal coherence.

We close this section with a timeline of seminal events in the history of mathematical proof. All of these events are treated in this book.

[14]Certainly there are exceptions to this statement. The engineering journal *IEEE Transactions in Signal Processing* often contains papers with mathematical proofs. A friend of mine recently had a paper rejected there for lack of a proof.

A TIMELINE OF MATHEMATICAL PROOF

Event	Date
Babylonian tablet with first proof	$\sim$ 1000 B.C.E.
Thales uses proofs	$\sim$ 600 B.C.E.
Pythagoras proves that $\sqrt{2}$ is irrational	$\sim$ 529 B.C.E.
Protagoras presents some of the first formal proofs	$\sim$ 430 B.C.E.
Hippocrates invents proof by contradiction	$\sim$ 420 B.C.E.
Plato's *Republic* discusses proof concept	$\sim$ 380 B.C.E.
Eudoxus develops the idea of theorem	$\sim$ 368 B.C.E.
Aristotle delineates proof techniques (in *Physics*)	$\sim$ 330 B.C.E.
Euclid writes *The Elements*	$\sim$ 285 B.C.E.
Eratosthenes creates his sieve	$\sim$ 236 B.C.E.
Al-Khwarizmi develops algebra	$\sim$ 820 C.E.
Kepler formulates his sphere-packing conjecture	1611 C.E.
Fermat lays foundations for differential calculus	1637 C.E.
Isaac Newton and Gottfried Leibniz invent calculus	$\sim$ 1666 C.E.
Gauss proves Fundamental Theorem of Algebra	1801 C.E.
Abel proves unsolvability of quintic equation	1824 C.E.
Galois records ideas of Galois theory (including groups)	1832 C.E.
Babbage produces analytical engine	1833 C.E.
Riemann formulates the Riemann hypothesis	1859 C.E.
Cantor publishes his ideas about cardinality and infinity	1873 C.E.
Hadamard and de la Vallee Poussin prove prime number theorem	1896 C.E.
Hilbert publishes *Grundlagen der Geometrie*	1899 C.E.
Hilbert delivers address with 23 seminal problems	1900 C.E.
Dehn solves Hilbert's third problem	1900 C.E.
Russell's paradox created	1902 C.E.
Frege's *The Foundations of Arithmetic* published	1903 C.E.
Poincare formulates his conjecture about spheres	1904 C.E.
L. E. J. Brouwer founds the intuitionist movement	1908 C.E.
Whitehead and Russell write *Principia Mathematica*	1910 C.E.

International Business Machines (IBM) incorporated	1911 C.E.
Banach-Tarski paradox is proved	1924 C.E.
Gödel publishes incompleteness theorem	1931 C.E.
First Bourbaki book appears	1939 C.E.
Erdős and Selberg give elementary proof of prime number theorem	1949 C.E.
von Neumann and Goldstine produce stored-program computer	1952 C.E.
A. Robinson introduces nonstandard analysis	1966 C.E.
E. Bishop publishes *Foundations of Constructive Analysis*	1967 C.E.
Cook invents **NP**-Completeness	1971 C.E.
Appel and Haken give computer proof of 4-color theorem	1976 C.E.
Jobs and Wozniak invent personal computer	1977 C.E.
Thurston formulates the geometrization program	1980 C.E.
Knuth invents TEX	1984 C.E.
de Branges proves Bieberbach conjecture	1984 C.E.
Hoffman, Hoffman, & Meeks use computer to generate embedded minimal surfaces	1990 C.E.
Horgan publishes *The Death of Proof?*	1993 C.E.
Hsiang publishes "solution" of Kepler problem	1993 C.E.
Andrew Wiles proves Fermat's last theorem	1994 C.E.
Finite, simple groups classified	1994 C.E.
McCune creates `Otter` Theorem-Proving Software	1994 C.E.
McCune proves Robbins conjecture using `EQP`	1997 C.E.
Almgren writes 1728-page regularity paper	1997 C.E.
Hsiang publishes book with "definitive" proof of Kepler	2001 C.E.
Grisha Perelman announces proof of Poincare Conjecture	2004 C.E.
Thomas Hales gives computer solution of Kepler problem	2006 C.E.

0.6 The Logical Basis for Mathematics

The late nineteenth century saw a great burgeoning of communication among mathematicians from different countries. Part and parcel of this new open culture was the realization that mathematics had become too scattered and fractured. Ideally, mathematics *should* be one, unified, pristine, logical edifice. It should all follow from a single set of definitions and a single set of

axioms. At least this was David Hilbert's dream.

Out of the intellectual milieu just described grew the seminal work of Frege on logical foundations (discussed in detail below). Another milestone in mathematical thought, created around this time, was *Principia Mathematica* [WRU] by Russell and Whitehead. Bertrand Russell, later to become a distinguished philosopher, was a student of the more senior Alfred North Whitehead at Cambridge University. They set out to derive all of basic mathematics, using only logic, from a minimal set of axioms. The end result was a massive, two-volume work that contains virtually no words. Only symbols! This was an exercise in pure mathematical logic taken to an exquisite extreme. One of the payoffs in this book, after about 1200 pages of hard work, was the theorem

$$2 + 2 = 4\,.$$

There are not many people today—even mathematicians—who study the book of Whitehead and Russell. But it is an important step in our development of mathematical rigor, and of what a proof should be. It represented, in its time, a pinnacle of the power of abstract logic.

It might be stressed that what Whitehead and Russell were doing was to produce a *strictly formal* development of mathematics. Their purpose was not to produce something readable, or comprehensible, or educational.[15] Their purpose instead was to record mathematics for the record, using the strictest rules of formal reasoning. A mathematical paper written today in the style of Whitehead and Russell would not be published. No journal would consider it—just because this mode of expression is not effective mathematical communication.

The mathematical proofs that we *do* publish today are distinctly less formal than the Whitehead/Russell model. Even though we adhere to strict rules of discourse, we also leave out steps and make small leaps and leave details to the reader—just because we want to get the ideas across in the most concise and elegant and effective manner possible. Often what we are publishing is a toolkit so that the reader can assemble his/her own proof. Such is quite analogous to what the chemist does: He/she, in a published paper, sketches how a certain experiment was performed (and of course describes what conclusions were drawn from it) so that the interested reader

[15] In fact—and this seems astonishing considering how important this work is—they had a difficult time getting *Principia Mathematica* published. They actually had to foot part of the publication cost themselves!

may reproduce the experiment if so desired. Often an important chemistry paper, describing years of hard work by dozens of people, will be just a few pages in length. This is an extreme application of Occam's Razor: the key ideas are recorded, so that other scientists may reproduce the experiment if needed.

0.7 The Experimental Nature of Mathematics

The discussion in the last section was accurate and complete, but it was not entirely truthful. Mathematicians *do* in fact engage in experimentation. How does this fit in with the strictly rigorous, axiomatic methodology that we have been describing?

What we have been discussing is the way that mathematics is *recorded*. It is because we use the axiomatic method and *proof* to archive our ideas that our subject is reliable, reproducible, and infallible. But this is not the way that mathematical facts are *discovered*. The working mathematician discovers new mathematical truths by *trying things*. He/she will work examples, talk to people, make conjectures, give lectures, attempt to formulate results, take a stab at proofs, derive partial results, make mistakes,[16] and try to learn from those mistakes. Probably the first ten attempts at the formulation of a new theorem will not be quite right. Hypotheses will have to be modified, perhaps strengthened. Conclusions may have to be altered or weakened. A theorem is arrived at and captured and finally formulated by means of trial and error. It frequently happens to the experienced mathematician that he/she will *know* something is true—will be able to picture it and describe it—but will not be able to formulate it precisely. Certainly it will be impossible at first to write down a rigorously formulated theorem.

In fact this is one of the most remarkable things about the professional mathematician. He/she spends a lifetime making mistakes and trying to learn from them. There is hardly any other profession that can make this claim. The mathematician in pursuit of a particular holy grail—a new theorem, or a new theory, or a new idea—could easily spend two or three years

[16] An old joke has it that a mathematician is cheap to fund because all he needs to do his work is paper, pencil, and a trash can. A philosopher is even cheaper because he doesn't need the trash can.

experimenting and trying things that do not work and falling on his/her face and picking himself/herself up again.

But here is the point. Once the mathematician has figured out what is going on, once he/she has finally arrived at a rigorous formulation and proof of his/her new idea, then it is time to engage the axiomatic method. The point is that the methodology of

$$\text{Definitions} \Rightarrow \text{Axioms} \Rightarrow \text{Theorem and Proof}$$

is the way that we *record* mathematics. It is the way that we can ensure permanence for our ideas, so that they will travel to and be comprehensible to future generations. It is *not* the way that we discover mathematics.

Today there is a remarkable mathematics journal called *Experimental Mathematics*. This journal—in a constructive way—flies in the face of mathematical tradition. For the tradition—going back to Euclid—has been to write up mathematics for the permanent record in a rigorous, formalized, axiomatic manner. One does so in such a way as to *not* reveal anything about how he/she arrived at the ideas, or about how many failed attempts he/she had, or what his/her partial results might have been. In short, a published mathematical paper is like a gleaming crystal ball, and the rest of the world is on the outside looking in.[17]

The periodical *Experimental Mathematics* turns the archetype just described on its head. For this forum encourages reports of partial results, descriptions of data generated by computer experiment, ideas learned from graphical images, assessments of numerical data, and analyses of physical experiments. The journal encourages speculation, and the presentation of partial or incomplete results. It publishes, for the most part, papers that other traditional mathematics journals will not consider. It has taken a bold step in acknowledging a part of the mathematical process that has never been formally accepted. And in doing so it has made a substantial and lasting contribution to our literature.

Experimental Mathematics is published by A K Peters, a daring and innovative mathematics publisher. Klaus Peters is himself a Ph.D. in mathemat-

[17]There is a grand tradition in mathematics of *not* leaving a trail of corn so that the reader may determine how the mathematical material was discovered or developed. Instead, the reader is supposed to figure it all out for himself. The result is a Darwinian world of survival of the fittest: only those with real mathematical talent can make their way through the rigors of the training procedure.

ics, and has shown extraordinary insight into the process of disseminating mathematical ideas. This journal is but one of his many innovations.

0.8 The Role of Conjectures

Of course the highest and finest form of guidance or direction that one mathematician can provide to the community of mathematicians is to prove a great theorem. This gives everyone something to think about, it points to new directions of research, and it raises ever more questions that are grist for the collective mill. But there are many other ways in which a mathematician can contribute to the common weal, exert some influence over the nature of his/her subject, make a difference in the directions of research. An instance of this type of activity is the formulation of conjectures.

If a mathematician has worked in a particular subject area for some years, then he/she will have a very strong sense of how the ideas fit together, which concepts are important, and which questions are the guiding principles for the subject. *If* that mathematician is a recognized leader in the field, *if* his/her opinions carry some weight, *if* perhaps this person is recognized as one of the creators of the subject area, then this person has the prerogative to make one or more conjectures (which other mathematicians will take seriously).

A *conjecture* is the postulation that something ought to be true (or perhaps false). A common way to pose a conjecture is to write a good paper and, in your summatory remarks at the end, to say that, "Here is the direction that I think the subject ought to go now. Here is what I think is true." And then you make a formal enunciation of a result. This is a result that you *cannot prove*—although you may be able to offer a plausibility argument, or the proof of a partial result, or at least some supportive evidence. Such a conjecture can have considerable influence over a subject, and can cause a good many people to shift the direction of their research.

Even though the academic subject of mathematics has few set rules, and even though there is room to let a thousand flowers bloom, it is well understood in the subject that you really ought not to be making conjectures unless you are a person of some substance. If everyone were running around making conjectures then the subject at hand would turn into a chaotic maelstrom, nobody would know what was true and what was false, everyone would get confused, and little progress would be made. So there is a *sotto voce* understanding that only certain sorts of people ought to be making conjectures.

Others should hold their peace.

Sometimes, if an eminent mathematician thinks he has proved a great result but turns out to be mistaken, then the mathematical community will exhibit its deference to the individual and call the result a conjecture named after that scholar. This is what happened with the Poincaré conjecture—discussed below. Poincaré thought he had proved it, but an error was discovered. So we now all tend to believe the result—just because Henri Poincaré did. And we call it the Poincaré conjecture. But it has not yet been proved.

Another instance of this phenomenon is the Riemann hypothesis. Riemann introduced in his paper [RIE] the basic ideas connected with the Riemann zeta function. He made speculations about the location of the zeros of this function (that is what the Riemann hypothesis is about). He went on to say that it would be desirable to have proofs of these assertions; he concludes by saying that his attempts at proof were unfruitful. He states that he will lay these matters aside, as they are not germane to his primary goal (which was to prove the prime number theorem). Sadly, Riemann died before he could achieve his goal.

0.8.1 Applied Mathematics

Up until about 1960, the vast majority of mathematical research in the United States was in pure mathematics. The tradition in Europe was rather different. Isaac Newton and Pierre de Fermat studied optics, Newton and Sonya Kowalevski studied celestial mechanics, George Green studied mathematical physics, Laplace studied celestial mechanics, Poincaré treated fluid mechanics and special relativity, Gauss contributed to the theory of geodesy and astrophysics, Turing worked on cryptography, and Cauchy even helped to develop the port facility for Napoleon's English invasion fleet. There have been several other instances throughout history of good mathematicians who were also interested in physics and engineering.

But in the early 1960s and earlier, few mathematics departments in the U.S. had any faculty who interacted with people from the physics department or the engineering department. Mathematicians in those days were content to sit in their offices and prove theorems about pure mathematics. They obtained the occasional diversion from chatting with their colleagues *in the mathematics department*. But in those days collaboration was the exception rather than the rule, so for the most part the mathematician was the ascetic lone wolf.

Starting in the early 1970s there was a distinct change in the viewpoint of what modern mathematics should be. Government funding agencies began to put pressure on universities, and on mathematics departments, to develop in "applied mathematics". Here *applied mathematics* is mathematics that is used on real-world problems. We mathematicians have long contented ourselves with saying that *all* mathematics can be applied; but this process sometimes takes a while and it is not our job to worry about what the mathematics we are doing is good for or how long it will take for the applications to develop.[18] We took comfort in citing all the many applications of Isaac Newton's mathematics, all the good uses that are made of George Green's and William Rowan Hamilton's and Arthur Cayley's mathematics. It was well known that the Courant Institute of Mathematical Sciences in New York City is a wellspring of excellent applied mathematics. That should be enough for the entire profession.

But no. The new ideal is that every mathematics department in the United States should have applied mathematicians. These should be people who interact with researchers in other departments, people who can teach the students how mathematics is applied to the study of nature. This new mission, backed by the power of funding (or potential loss thereof), caused quite a tumult within the professoriate. Where were we to find all these applied mathematicians? Who were they? And how does one recognize excellence in applied mathematics? What are the important problems in applied mathematics? How does one study them? And how can one tell when he/she has reached a solution?

It is safe to say that, for a period of fifteen or twenty years, most all of the mathematics departments in this country struggled with the issues just described. Certainly the Courant Institute played a significant role in supplying the needed properly trained applied mathematicians. Great Britain, long a bastion of practical science,[19] was also a great source for applied mathematical scientists. But mathematics departments had to re-think the way that they did things. For a tenure case in applied mathematics does not necessarily look like a tenure case in pure mathematics. The work is published in different journals, and new standards are applied to evaluate the work. An applied mathematician does not necessarily enunciate and prove crisp, new

[18] In his charming autobiographical memoir [HAR], British mathematican G. H. Hardy virtually crows that he has never done anything useful and never will.

[19] Ernest Rutherford (1871–1937) was a role model for the down-to-earth British scientist. He would never accept relativity theory, for example.

theorems. Instead he/she might be an expert in the analysis of numerical data, or in producing graphical images of physical phenomena. The applied mathematician might create a new high-level computer language (such as John Kemeny participating in the creation of `BASIC`). He might collaborate with engineers or physicists or medical researchers or workers in the school of social research.

Today, after a protracted struggle, it is safe to say that the American mathematical community has embraced applied mathematics. Some few universities have separate pure and applied mathematics departments. Thus the pure mathematicians can remain pure and the applied mathematicians can do what they want to do. But most universities have just one mathematics department and the pure and applied mathematicians co-exist. This author's university College of Arts and Sciences has just one math department, and almost all of its denizens are classically trained pure mathematicians. But a significant number of us has developed interests in applied mathematics. Two of us, who were originally trained in group theory and harmonic analysis, now study statistics. They collaborate with members of the Medical School and the School of Social Work. One of us, originally trained in harmonic analysis, is now an expert in wavelet algorithms for image compression and signal processing. He consults with many engineering firms and with professors in engineering and the Medical School. One of us (namely, this author) collaborates with plastic surgeons. Another works with chemical engineers.

This is exactly the sort of symbiosis that the government, and the university administrations, were endeavoring to foster thirty years ago. And it has come to pass. And the good news is that we are creating new courses, and new curricula, to validate the change. So students today are being exposed not just to pure, traditional mathematics but also to the manifold ways in which mathematics is used. We—the mathematics departments, the students, the government, and the university administration—can point with pride to the ways in which mathematics has affected our world:

- Mathematicians designed the carburation system in the Volvo automobile.

- Mathematical theory underlies the design of the cellular telephone.

- Mathematics is the basis for America's pre-eminence in radar and scanning technologies.

- Mathematical theory underlies the technology for CD music discs and DVD movie discs.
- Mathematics is the underpinning for queuing theory, coding theory, and the ideas behind Internet routing and security.
- The entire theoretical basis for cryptography is mathematical.
- Mathematics is very much in the public eye because of the Olympiad (the international mathematics contest), because of the movies *Good Will Hunting* and *A Beautiful Mind*, because of the television show *Numb3rs*, because of the play *Proof*.

The list could go on and on.

It is safe to say that, today, pure mathematics and applied mathematics co-exist in a mutually nurturing environment. They do not simply tolerate each other; in fact they provide ideas and momentum for each other. It is a fruitful and rewarding atmosphere in which to work, and it continues to develop and grow.

It has been noted elsewhere that the tradition in mathematics has been for the mathematician to be a single combat warrior. He/she sat in an office, thought lone thoughts, and proved theorems. In 1960, and before, almost all mathematical published work had just one author. Today that has changed entirely. In fact in the past fifteen years the distinct majority of mathematical work is done collaboratively. Now the lone worker is the exception. What is the reason for this change?

First of all, the symbiosis between the pure world and the applied world has necessitated that people *talk* to each other. This author works on research projects with plastic surgeons. They do not have any expertise in mathematics and I do not have any expertise in plastic surgery. So collaboration is necessary. My colleagues who work with chemical engineers, or with physicists, must share skills and resources in the same way and for the same reason.

But it should be stressed that, even among pure mathematicians, collaboration has increased dramatically. The reason, it seems, is that mathematics as a whole has become more complex. In the past forty years we have learned of a great many synergies between different mathematical fields. So it makes much more sense for a topologist to talk to an analyst, or a geometer to talk to a differential equations expert. The consequence is a blooming of joint

work that has enriched our subject and dramatically increased its depth. Mathematical collaboration has sociological and psychological consequences as well. It is difficult and depressing to work on mathematics in a solitary fashion. The problems are difficult and discouraging, and it is rather easy to feel a profound sense of isolation, and ultimately of depression and failure. A good collaborator can keep one alive, and provide momentum and encouragement when they are needed. We as a profession have discovered the value—both professionally and emotionally—of having collaborators and of doing joint work. It has been good for all concerned.

0.9 Mathematical Uncertainty

In this section we explore another aspect in which we have not been entirely truthful. While it is the case that, in many respects, mathematics is definitely the most reliable, infallible, reproducible set of ideas ever devised, it also contains some pitfalls (see [KLN]). In particular, the twentieth century has given mathematics some kicks in the pants. We shall take a few moments to describe some of these.

First some background. When we write up a proof, so that it can be submitted to a journal and refereed and ultimately (we hope) published, then we are anticipating that the (mathematical) world at large will read it and appreciate it and validate it. This is an important part of the process that is mathematics: it is the mathematics profession, taken as a whole, that decides what is correct and valid, and also what is useful and is interesting and has value. The creator of the new mathematics has the responsibility of setting it before the mathematical community; but the community itself either makes the work part of the canon or it does not.

And the writing of a mathematical paper is the walking of a fine line. The rhetoric of modern mathematics is a very strict mode of discourse. On the one hand, the formal rules of logic must be followed. The paper must contain no "leaps of faith" or guesses or sleight of hand. On the other hand, if the writer *really* includes every step, *really* cites every rule of logic that is being used, *really* leaves no stone unturned, then even the simplest argument will drag on for pages. A really substantial mathematical theorem could take hundreds of pages to prove. This simply will not work. The mathematical journals cannot afford to publish so much material, and nobody will be able to read it. So what we do in practice is that we skip steps. Usually these

are fairly small steps (at least small in the mind of the writer), but it is not uncommon for a mathematician to spend a couple of hours working out a step in a paper that he/she is reading because the author left something out.[20] To summarize: we usually omit many steps in proofs. In principle, the reader can fill in the missing steps. Mathematicians generally find it inappropriate to leave a trail of hints so that the reader can see how the ideas were discovered. This is something that the reader is supposed to do for himself. In mathematics we exhibit the finished product, gleaming and elegant. We do not tell the reader how we figured it out.

Let us now introduce some terminology. In mathematics, a *set* is a collection of objects. This is an example of a mathematical definition—one which describes a new concept (namely "set") in everyday language. We usually denote a set with a capital roman letter, such as S or T or U. There is a whole branch of mathematics called "set theory", and it is the very foundation of most any subject area of mathematics. Georg Cantor (1845–1918) was arguably the father of modern set theory. Thus the late nineteenth and early twentieth centuries were the heydays for laying the foundations of set theory.

This book is not the place to engage in a treatment of basic set theory. But we shall introduce one auxiliary piece of terminology that will be useful in the ensuing discussion. Let S be a set. We say that x is an *element* of S, and we write $x \in S$, if x is one of the objects in S. As an example, let S be the set of positive whole numbers. So

$$S = \{1, 2, 3, \ldots\}.$$

Then 1 is an element of S. And 2 is an element of S. And 3 is an element of S, and so forth. We write $1 \in S$ and $2 \in S$ and $3 \in S$, etc. But note that π is *not* an element of S. For $\pi = 3.14159265\ldots$. It is not a whole number, so it is not one of the objects in S. We write $\pi \notin S$.

Now let us return to our discussion of the saga of set theory. In 1902, G. Frege (1848–1925) was enjoying the fact that the second volume of his definitive work *The Basic Laws of Arithmetic* [FRE2] was at the printer

[20] What we are describing here is not entirely different from what is done in the laboratory sciences. When a chemist performs an important experiment, and writes it up for publication, he/she does so in a rather telegraphic style. The idea is to give the reader enough of an idea of how the experiment was done so that he/she can repeat it if so desired.

when he received a polite and modest letter from Bertrand Russell offering the following paradox:[21]

> Let S be the collection of all sets that are not elements of themselves. Can S be an element of S?

Why is this a paradox?[22]

Here is the problem. If $S \in S$ then, by the way that we defined S, S is *not* an element of S. And if S is *not* an element of S, then, by the way that we defined S, it follows that S *is* an element of S. Thus we have a contradiction no matter what.

Of course we are invoking Archimedes's law of the excluded middle. It *must* be the case that either $S \in S$ or $S \notin S$, but in fact either situation leads to a contradiction. And that is Russell's Paradox. Frege had to rethink his book, and make notable revisions, in order to address the issues raised by Russell's paradox.[23]

[21]This paradox was quite a shock to Frege. After considerable correspondence with Russell, he modified one of his axioms and added an Appendix explaining how the modification addresses Russell's concerns. Unfortunately this modification nullified several of the results in Frege's already-published Volume 1. Frege's second volume *did* ultimately appear (see [FRE2]). Frege was somewhat disheartened by his experience, and his research productivity definitely went into a decline. His planned third volume never appeared.

Lesniewski proved, after Frege's death, that the resulting axiom system that appears in print is inconsistent. Frege is nonetheless remembered as one of the most important figures in the foundations of mathematics. He was one of the first to formalize the rules by which mathematicians operate, and in that sense he was a true pioneer. Many scholars hold that his earlier work, *Begriffsschrift und andere Aufsätze* [FRE1],is the most important single work ever written in logic. It lays the very foundations of modern logic. A more recent 1995 paper of G. Boolos [BOO] makes considerable strides in rescuing much of Frege's original program that is represented in the two-volume work [FRE2].

[22]The thoughtful reader may well wonder whether it is actually possible for a set to be an element of itself. This sounds like a form of mental contortionism that is implausible at best. But consider the set S described by

> The collection of all sets that can be described in fewer than fifty words.

Notice that S is a set, and S can be described in fewer than fifty words. Aha! So S is certainly an element of itself.

[23]A popular version of Russell's Paradox goes like this. A barber in a certain town agrees to shave every man in the town who does not shave himself. He will not touch any man who ever deigns to shave himself. Who shaves the barber? If the barber shaves himself, then he himself is someone whom the barber has agreed not to shave. If instead the barber does not shave himself, then he himself is someone whom the barber must shave. A contradiction either way.

Now that we have had about a century to think about Russell's paradox, we realize that what it teaches us is that we cannot allow sets that are *too large*. The set S described in Russell's paradox is unallowably large. In a rigorous development of set theory, there are very specific rules for which sets we are allowed to consider and which not. In particular, modern set theory does not allow us to consider a set that is an element of itself. We cannot indulge in the details here.

It turns out that Russell's paradox is only the tip of the iceberg. Nobody anticipated what Kurt Gödel (1906–1978) would teach us thirty years later. In informal language, what Gödel showed us is that—in any sufficiently complex logical system (i.e., at least as complex as arithmetic)—there will be a sensible statement that we can neither prove nor disprove within that system.[24] This is Gödel's incompleteness theorem. It came as an unanticipated bombshell, and has completely altered the way that we think of our subject. Note here that the statement that Gödel created will definitely have a truth value—it will be either true or false. But we will not be able to prove it with a sequence of logical steps *within the given logical system*. It should be stressed however that the statement that Gödel found is not *completely unprovable*. If one transcends the specified logical system, and works instead in a larger and more powerful system, then one *can* create a proof.

In fact what Gödel did was extraordinarily powerful and elegant. He found a way to assign a number (i.e., a positive whole number) to each statement in the given logical system. This number has come to be known as the *Gödel number*. It turns out then that statements in the logical system about the natural numbers are actually statements about the statements themselves. Thus Gödel is able to formulate a statement U within the logical system that says, in effect, "U asserts that it is unprovable." This is a problem. Because if U is false then U is provable. And if U is true then U is unprovable. So we have a true statement that cannot be proved *within the system*. The books [SMU1], [SMU2] provide entertaining and accessible discussions of Gödel's ideas.

Just as quantum mechanics taught us that nature is not completely deterministic—we cannot know everything about a given physical system, even if we have a complete list of all the initial conditions—so it is the case

[24]Gödel even went so far as to show that the consistency of arithmetic itself is not provable. Certainly arithmetic is the most fundamental and widely accepted part of basic mathematics. The notion that we can never be certain that it will not lead to a contradiction is certainly unsettling.

that Gödel has taught us that there will always be statements in mathematics that are "undecidable".

It is safe to say that Gödel's ideas shook the very foundations of mathematics. They have profound implications for the logical basis for our subject, and for what we can expect from it. There are also serious consequences for theoretical computer science—just because the computer scientist wants to know where any given programming language (which is certainly a logical system) will lead and what it can produce.

The good news is that the consequences of Gödel's incompleteness theorem rarely arise in everyday mathematics. The "Gödel statement" is more combinatorial than analytical. One does not encounter such a sentence in calculus and analysis. It can, however, arise sometimes in algebra and number theory and discrete mathematics.[25] There have been highly desirable and much-sought-after results in number theory that have been proved to be undecidable (see Ax-Kochen [AXK1], [AXK2], [AXK3]). And of course Gödel's ideas play a major role in mathematical logic.

0.10 The Publication and Dissemination of Mathematics

Five hundred years ago, scientists were often rather secretive. They tended to keep their results, and their scientific discoveries, to themselves. Even when asked by another scientist for a specific piece of data, or queried about a specific idea, these scholars were often evasive. Why would serious scientific researchers behave in this fashion?

We must understand that the world was different in those days. There were very few academic positions. Many of the great scientists did their research as a personal hobby. Or, if they were lucky, they could locate a wealthy patron who would subsidize their work. But one can see that various resentments and jealousies could easily develop. Certainly there was no National Science Foundation and no National Institutes of Health at the time of Johannes Kepler (1571–1630). Many a noted scientist would spend years struggling to find an academic position. The successful landing of a professorship certainly involved various patronage issues and a variety of

[25] One must note, however, that the last decade has seen a great cross-fertilization of analysis with combinatorics. The statements we are making now may soon be out of date.

academic and non-academic politics. Even the great Riemann did not land a suitable professorship until he was lying on his death bed.

One of the most famous cases of a scientist who was secretive about his work was Isaac Newton. Newton was the greatest scientist who ever lived, and he produced myriad ideas that revolutionized scientific thought. But he was an irrascible, moody, temperamental individual with few friends. On one occasion a paper that he had submitted for publication was subjected to criticisms by Robert Hooke (1635–1703), and Newton took this badly. He published nothing for quite a time after that. Certainly Newton's reticence to publish meant that many of the key ideas of the calculus were kept under wraps. Meanwhile, Gottfried Wilhelm von Leibniz was independently developing the ideas of calculus in his own language. Leibniz did not suffer from any particular reticence to publish. And of course the publication of Leibniz's ideas caused considerable consternation among Newton and his adherents. For they felt that Leibniz was attempting to abscond with ideas that were first created by Newton. Of course one could argue in the other direction that Newton should have published his ideas in a timely manner. That would have removed all doubt about who first discovered calculus.

In the mid-seventeenth century, Henry Oldenburg was active in the scientific societies. Because of his personality, and his connections, Oldenburg became something of a go-between among scientists of the day. If he knew that scientist A needed some ideas of scientist B, then he would arrange to approach B to ask him to share his ideas. Usually Oldenburg could arrange to offer a quid for this largesse. In those days books were rare and expensive, and Oldenburg could sometimes arrange to offer a scientific book in exchange for some ideas.

After some years of these activities, Oldenburg and his politicking became something of an institution. This led Henry Oldenburg to create the first peer-reviewed scientific journal in 1665. He was the founding editor of *The Philosophical Transactions of the Royal Society of London*. At the time it was a daring but much-needed invention that supplanted a semi-secret informal method of scientific communication that was both counterproductive and unreliable. Today journals are part of the fabric of our professional life. Most scientific research is published in journals of some sort.

In modern times journals are the means of our professional survival. Any scientist who wants to establish his/her reputation must publish his/her ideas in scientific journals. If an Assistant Professor wants to get tenure then it must be established that he/she has a substantial scholarly *Gestalt*. This

means that the individual will have created some substantial new ideas in his/her field, and will have lectured about them and published some write-up of the development of this thought. This entire circle of considerations has led to the popular admonition of "Publish or perish." The notion is perhaps worthy of detailed discussion.

For the past many years, and especially since the advent of NSF (National Science Foundation) grants, we have been living under the specter of "publish or perish." The meaning of this aphorism is that, if you are an academic, and if you want to get tenured or promoted, or you want to get a grant, or you want an invitation to a conference, or you want a raise, or you want the respect and admiration of your colleagues, then you had better publish original work in recognized, refereed journals or books. Otherwise you're outta here. Who coined the phrase "publish or perish"?

One might think that it was a President of Harvard. Or perhaps a high-ranking officer at the NSF. Or some Dean at Cal Tech. One self-proclaimed expert on quotations suggested to me that it was Benjamin Franklin! But, no, it was sociologist Logan Wilson in his 1942 book *The Academic Man, A Study in the Sociology of a Profession* [WILS]. He said, "The prevailing pragmatism forced upon the academic group is that one must write something and get it into print. Situational imperatives dictate a 'publish or perish' credo within the ranks."

Wilson was President of the University of Texas and (earlier) a student at Harvard of the distinguished sociologist Robert K. Merton. So he no doubt knew whereof he spoke.

The estimable Marshall McLuhan has sometimes been credited with the phrase "publish or perish," and it is arguable that it was he who popularized it. In a June 22, 1951 letter to Ezra Pound he wrote (using Pound's favorite moniker "beaneries" to refer to the universities)

> The beaneries are on their knees to these gents (foundation administrators). They regard them as Santa Claus. They will do 'research on anything' that Santa Claus approves. They will think his thoughts as long as he will pay the bill for getting them before the public signed by the profesorry-rat. 'Publish or perish' is the beanery motto.

0.11 Closing Thoughts

The purpose of this chapter has been to bring the reader up to speed, to acquaint him or her with the basic tenets of mathematical thinking. The remainder of the book is an *analysis* of that mathematical thinking. Just what does a mathematician do all day? What is he or she trying to achieve? And what is the method for doing so?

For the theoretical mathematician, that method is *proof*. The mathematician discovers new ideas or theories, finds a way to formulate them, and then must verify them. The vehicle for doing so is the classical notion of proof. In the ensuing chapters we explore how the concept of proof came about, how it developed, how it became the established methodology in the subject, and how it has been developing and changing over the years.

Chapter 1

The Ancients

Man propounds negotiations, man accepts the compromise. Very rarely will he squarely push the logic of a fact to its ultimate conclusion in unmitigated act.

Rudyard Kipling

No man, for any considerable period, can wear one face to himself, and another to the multitude, without finally getting bewildered as to which may be the true.

Nathaniel Hawthorne

For a charm of powerful trouble,
Like a hell-broth boil and bubble.
Double, double toil and trouble;
Fire burn and caldron bubble.

William Shakespeare, *Macbeth*

Since the earliest times, all critical revision of the principles of mathematics as a whole, or of any branch of it, have almost invariably followed periods of uncertainty, where contradictions did appear and had to be resolved . . . There are now twenty-five centuries during which the mathematicians have had the practice of correcting their errors, and thereby seeing their science enriched, not impoverished; this gives them the right to view the future with serenity.

Nicholas Bourbaki

. . . If you retort with some very simple case which makes me out a stupid animal, I think I must do as the Sphynx did . . .

August De Morgan

One normally thinks that everything that is true is true for a reason. I've found mathematical truths that are true for no reason at all. These mathematical truths are beyond the power of mathematical reasoning because they are accidental and random.

G. J. Chaitin

One may say today that absolute rigor has been attained.

Henri Poincaré

Most of the research described as experimental is Baconian in nature, but also one can argue that all of mathematics proceeds out of Baconian experiments. One tries out a transformation here, and identity there, examines what happens when one weakens this condition or strengthens that one. Even the application of probabilistic arguments in number theory can be seen as a Baconian experiment. The experiments may be well thought out and very likely to succeed, but the criterion of inclusion of the result in the literature is success or failure. If the "messing about" works (e.g., the theorem is proved, the counterexample found), the material is kept; otherwise, it is relegated to the scrap heap.

J. Borwein, P. Borwein, R. Girgensohn, and S. Parnes

1.1 Eudoxus and the Concept of Theorem

Perhaps the first mathematical proof in recorded history is due to the Babylonians. They seem (along with the Chinese) to have been aware of the Pythagorean theorem (discussed in detail below) well before Pythagoras.[1] The Babylonians had certain diagrams that indicate why the Pythagorean theorem is true, and tablets have been found to validate this fact. They also had methods for calculating Pythagorean triples—that is, triples of integers (or whole numbers) a, b, c that satisfy

$$a^2 + b^2 = c^2$$

as in the Pythagorean theorem.

The Babylonians were remarkably sophisticated in a number of ways. As early as 1200 B.C.E. they had calculated $\sqrt{2}$ to (what we would call) six decimal places.[2] They did not "prove" theorems as today we conceive of the activity, but they certainly had well-developed ideas about mathematics (not just arithmetic).

It was Eudoxus (408 B.C.E–355 B.C.E.) who began the grand tradition of organizing mathematics into theorems. The word "theorem" comes from

[1] Although it must be stressed that they did not have Pythagoras's sense of the structure of mathematics, of the importance of rigor, or of the nature of formal proof.

[2] Of course the Babylonians did *not* have decimal notation.

the Greek root *theorema*, meaning "speculation". Eudoxus was one of the first to use this word in the context of mathematics.

What Eudoxus gained in the rigor and precision of his mathematical formulations, he lost because he did not prove anything. Formal proof was not yet the tradition in mathematics. As we have noted elsewhere, mathematics in its early days was a largely heuristic and empirical subject. It had never occurred to anyone that there was any need to prove anything. When you asked yourself whether a certain table would fit in your dining room, you did not prove a theorem; you just checked it out.[3] When you wondered whether a certain amount of fence would surround your pasture, you did not seek a rigorous argument; you simply unrolled the fence and determined whether it did the job. In its earliest days, mathematics was intimately bound up with questions precisely like these. Thus mathematical thinking was almost inextricable from practical thinking. And that is how its adherents viewed mathematical facts. They were just practical information, and their assimilation and verification was a strictly pragmatic affair.

1.2 Euclid the Geometer

Euclid (325 B.C.E.–265 B.C.E.) is hailed as the first scholar to systematically organize mathematics (i.e., a substantial portion of the mathematics that went before him), formulate definitions and axioms, and prove theorems. This was a monumental achievement, and a highly original one.

Although Euclid is not known so much (as were Archimedes and Pythagoras) for his original and profound insights, and although there are not many theorems named after Euclid, he has had an incisive effect on human thought. After all, Euclid wrote a treatise (consisting of thirteen Books)—now known as Euclid's *Elements*—which has been continuously available for over 2000 years and has been through a large number of editions. It is still studied in

[3]William (Willy) Feller (1906–1970) was a prominent mathematician at Princeton University. He was one of the fathers of modern probability theory. Feller and his wife were once trying to move a large circular table from their living room into the dining room. They pushed and pulled and rotated and maneuvered, but try as they might they could not get the table through the door. It seemed to be inextricably stuck. Frustrated and tired, Feller sat down with a pencil and paper and devised a mathematical model of the situation. After several minutes he was able to *prove* that what they were trying to do was impossible. While Willy was engaged in these machinations, his wife had continued struggling with the table, and she managed to get it into the dining room.

detail today, and continues to have a substantial influence over the way that we think about mathematics.

Not a great deal is known about Euclid's life, although it is fairly certain that he had a school in Alexandria. In fact "Euclid" was quite a common name in his day, and various accounts of Euclid the mathematician's life confuse him with other Euclids (one a prominent philosopher). One appreciation of Euclid comes from Proclus, one of the last of the ancient Greek philosophers:

> Not much younger than these [pupils of Plato] is Euclid, who put together the *Elements*, arranging in order many of Eudoxus's theorems, perfecting many of Theaetetus's, and also bringing to irrefutable demonstration the things which had been only loosely proved by his predecessors. This man lived in the time of the first Ptolemy; for Archimedes, who followed closely upon the first Ptolemy makes mention of Euclid, and further they say that Ptolemy once asked him if there were a shortened way to study geometry than the *Elements*, to which he replied that "there is no royal road to geometry." He is therefore younger than Plato's circle, but older than Eratosthenes and Archimedes; for these were contemporaries, as Eratosthenes somewhere says. In his aim he was a Platonist, being in sympathy with this philosophy, whence he made the end of the whole *Elements* the construction of the so-called Platonic figures.

As often happens with scientists and artists and scholars of immense accomplishment, there is disagreement, and some debate, over exactly who or what Euclid actually was. The three schools of thought are these:

- Euclid was an historical character—a single individual—who in fact wrote the *Elements* and the other scholarly works that are commonly attributed to him.

- Euclid was the leader of a team of mathematicians working in Alexandria. They all contributed to the creation of the complete works that we now attribute to Euclid. They even continued to write and disseminate books under Euclid's name after his death.

- Euclid was not an historical character at all. In fact "Euclid" was a *nom de plume* adopted by a group of mathematicians working in Alexandria.

They took their inspiration from Euclid of Megara (who *was* in fact an historical figure), a prominent philosopher who lived about 100 years before Euclid the mathematician is thought to have lived.

Most scholars today subscribe to the first theory—that Euclid was certainly a unique person who created the *Elements*. But we acknowledge that there is evidence for the other two scenarios. Certainly Euclid had a vigorous school of mathematics in Alexandria, and there is little doubt that his students participated in his projects.

It is thought that Euclid must have studied in Plato's (430 B.C.E.–349 B.C.E.) Academy in Athens, for it is unlikely that there would have been another place where he could have learned the geometry of Eudoxus and Theaetetus on which the *Elements* is based.

Another famous story[4] and quotation about Euclid is this. A certain pupil of Euclid, at his school in Alexandria, came to Euclid after learning just the first proposition in the geometry of the *Elements*. He wanted to know what he would gain by putting in all this study, doing all the necessary work, and learning the theorems of geometry. At this, Euclid called over his slave and said, "Give him three drachmas since he must needs make gain by what he learns."

What is important about Euclid's *Elements* is the paradigm it provides for the way that mathematics should be studied and recorded. He begins with several definitions of terminology and ideas for geometry, and then he records five important postulates (or axioms) of geometry. A version of these postulates is as follows:

P1 Through any pair of distinct points there passes a line.

P2 For each segment $\overline{AB}$ and each segment $\overline{CD}$ there is a unique point E (on the line determined by A and B) such that B is between A and E and the segment $\overline{CD}$ is congruent to $\overline{BE}$ (Figure 1.1).

P3 For each point C and each point A distinct from C there exists a circle with center C and radius CA.

P4 All right angles are congruent.

These are the standard four axioms which give our Euclidean conception of geometry. The fifth axiom, a topic of intense study for two

[4] A similar story is told of Plato.

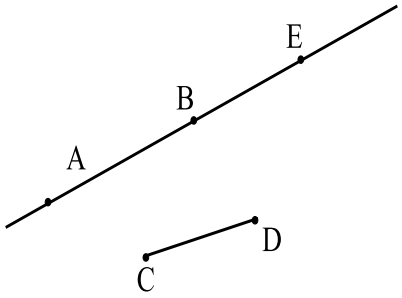


Figure 1.1

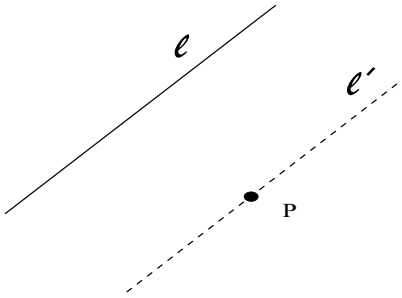


Figure 1.2

thousand years, is the so-called parallel postulate (in Playfair's formulation):

P5 For each line ℓ and each point P that does not lie on ℓ there is a unique line ℓ' through P such that ℓ' is parallel to ℓ (Figure 1.2).

Of course, prior to this enunciation of his celebrated five axioms, Euclid had defined "point", "line", "circle", and the other terms that he uses. Although Euclid borrowed freely from mathematicians both earlier and contemporaneous with himself, it is generally believed that the famous "Parallel Postulate", that is Postulate **P5**, is of Euclid's own creation.

It should be stressed that the book the *Elements* is not simply about plane geometry. In fact Books VII–IX deal with number theory. It is here that Euclid proves his famous result that there are infinitely many primes (treated elsewhere in this book) and also his celebrated "Euclidean algorithm" for long division. Book X deals with irrational numbers, and books XI–XIII treat three-dimensional geometry. In short, Euclid's *Elements* are an exhaustive treatment of a good deal of the mathematics that was known at the time. And

it is presented in a strictly rigorous and axiomatic manner that has set the tone for the way that mathematics is recorded and studied today. Euclid's *Elements* is perhaps most notable for the clarity with which theorems are formulated and proved. The standard of rigor that Euclid set was to be a model for the inventors of calculus nearly 2000 years later.

Noted algebraist B. L. van der Waerden (1903 C.E.–1996 C.E.) assesses the impact of Euclid's elements in this way:

> Almost from the time of its writing and lasting almost to the present, the *Elements* has exerted a continuous and major influence on human affairs. It was the primary source of geometric reasoning, theorems, and methods at least until the advent of non-Euclidean geometry in the 19th century. It is sometimes said that, next to the Bible, the *Elements* may be the most translated, published, and studied of all the books produced in the Western world.

Indeed, there have been more than 1000 editions of Euclid's *Elements*. It is arguable that Euclid was and still is the most important and most influential mathematics teacher of all time. It may be added that a number of other books by Euclid survive until now. These include *Data* (which studies geometric properties of figures), *On Divisions* (which studies the division of geometric regions into subregions having areas of a given ratio), *Optics* (which is the first Greek work on perspective), and *Phaenomena* (which is an elementary introduction to mathematical astronomy). Several other books of Euclid—including *Surface Loci*, *Porisms*, *Conics*, *Book of Fallacies*, and *Elements of Music*—have all been lost.

1.2.1 Euclid the Number Theorist

Most of us remember Euclid's *Elements* as a work on geometry. And it is perhaps that portion of the book that has been most influential. After all, it was this writing that laid the foundations for the axiomatic method that has stood now for over two millenia.

But in fact Euclid's *Elements* was a record of a good deal of the mathematics that was known at the time. And Books VII–IX of the *Elements* deal with number theory. Certainly one of the particular results presented there has stood the test of time, and the proof is taught today to every mathematics student. We shall discuss it now.

Recall that a *prime number* is a positive whole number which has no divisors except for 1 and itself. By tradition we do not consider 1 to be a prime. So the prime numbers are

$$2\ ,\ 3\ ,\ 5\ ,\ 7\ ,\ 11\ ,\ 13\ ,\ 17\ ,\ 19\ ,\ 23\ ,\ 29\ ,\ 31\ ,\ 37\ ,\ \ldots\ .$$

A number that is not prime is called *composite*. For example, 126 is composite. Notice that

$$126 = 2 \cdot 3^2 \cdot 7\,.$$

It is certainly not prime. Any composite number can be factored in a unique fashion into prime factors—that is the fundamental theorem of arithmetic.

The question that Euclid considered (and, unlike many of the other results in the *Elements*, this result seems to have originated with Euclid himself) is whether or not there are infinitely many prime numbers. And Euclid's dramatic answer is "yes".

Theorem: *There are infinitely many prime integers.*

For the proof, assume the contrary. So there are only finitely many primes. Call them $p_1, p_2, \ldots, p_N$. Now consider the number $P = (p_1 \cdot p_2 \cdots p_N) + 1$. What kind of number is P? Notice that if we divide P by p_1 then we get a remainder of 1 (since p_1 goes evenly into $p_1 \cdot p_2 \cdots p_N$). Also if we divide p_2 into P then we get a remainder of 1. And it is just the same if we divide any of p_3 through p_N into P.

Now, if P were a composite number then it would have to be evenly divisible by some prime. But we have just shown that it is not: We have divided every known prime number into P and obtained a nonzero remainder in each instance. The only possible conclusion is that P is another prime, obviously greater than any of the primes on the original list. That is a contradiction. So there cannot be finitely many primes. There must be infinitely many.

Euclid's argument is one of the first known instances of proof by contradiction.[5] This important method of formal reasoning has actually been quite controversial over the years. We shall discuss it in considerable detail as the book develops.

[5] There are many other ways to prove Euclid's result, including direct proofs and proofs by induction. In other words, it is not *necessary* to use proof by contradiction.

1.3 Pythagoras

Pythagoras (569–500 B.C.E.) was both a person and a society (i.e., the *Pythagoreans*). He was also a political figure and a mystic. He was special in his time, among other reasons, because he involved women as equals in his activities. One critic characterized the man as "one tenth of him genius, nine-tenths sheer fudge." Pythagoras died, according to legend, in the flames of his own school fired by political and religious bigots who stirred up the masses to protest against the enlightenment which Pythagoras sought to bring them.

The Pythagorean society was intensely mathematical in nature, but it was also quasi-religious. Among its tenets (according to [RUS]) were:

- To abstain from beans.
- Not to pick up what has fallen.
- Not to touch a white cock.
- Not to break bread.
- Not to step over a crossbar.
- Not to stir the fire with iron.
- Not to eat from a whole loaf.
- Not to pluck a garland.
- Not to sit on a quart measure.
- Not to eat the heart.
- Not to walk on highways.
- Not to let swallows share one's roof.
- When the pot is taken off the fire, not to leave the mark of it in the ashes, but to stir them together.
- Not to look in a mirror beside a light.

- When you rise from the bedclothes, roll them together and smooth out the impress of the body.

The Pythagoreans embodied a passionate spirit that is remarkable to our eyes:

> Bless us, divine Number, thou who generatest gods and men.

and

> Number rules the universe.

The Pythagoreans are remembered for two monumental contributions to mathematics. The first of these was establishing the importance of, and the necessity for, *proofs* in mathematics: that mathematical statements, especially geometric statements, must be verified by way of rigorous proof. Prior to Pythagoras, the ideas of geometry were generally rules of thumb that were derived empirically, merely from observation and (occasionally) measurement. Pythagoras also introduced the idea that a great body of mathematics (such as geometry) could be derived from a small number of postulates. The second great contribution was the discovery of, and proof of, the fact that not all numbers are commensurate. More precisely, the Greeks prior to Pythagoras believed with a profound and deeply held passion that everything was built on the whole numbers. Fractions arise in a concrete manner: as ratios of the sides of triangles with integer length (and are thus *commensurable*—this antiquated terminology has today been replaced by the word "rational")—see Figure 1.3.

Pythagoras proved the result that we now call *the Pythagorean theorem*. It says that the legs a, b and hypotenuse c of a right triangle (Figure 1.4) are related by the formula

$$a^2 + b^2 = c^2 \,. \qquad (\star)$$

This theorem has perhaps more proofs than any other result in mathematics—well over fifty altogether. And in fact it is one of the most ancient mathematical results. There is evidence that the Babylonians and the Chinese knew this theorem at least 500 years before Pythagoras.

It is remarkable that one proof of the Pythagorean theorem was devised by U.S. President James Garfield (1831–1881). We now provide one of the simplest and most classical arguments.

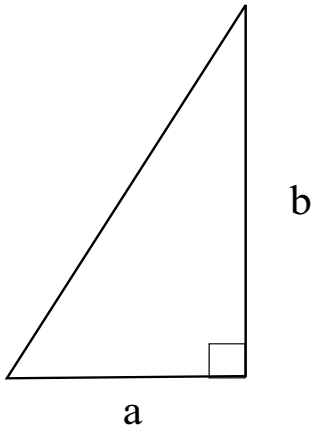


Figure 1.3. The fraction $\frac{b}{a}$.

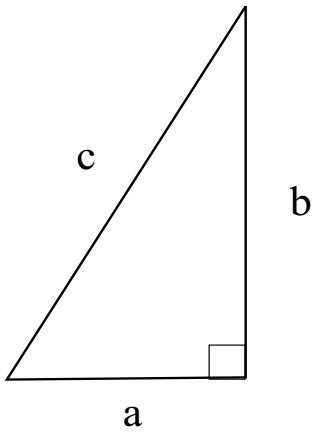


Figure 1.4. The Pythagorean theorem.

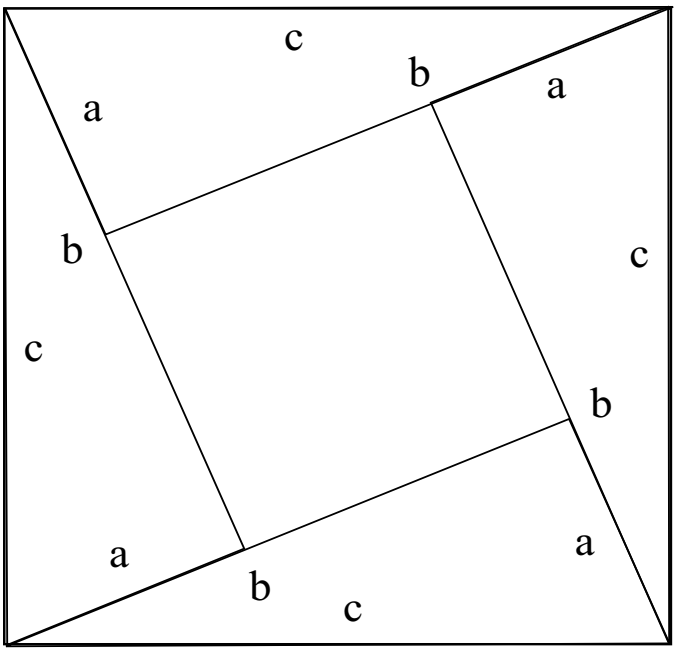


Figure 1.5

Proof of the Pythagorean Theorem:

Refer to Figure 1.5. Observe that we have four right triangles and a square packed into a large square. Each of the triangles has legs a and b and hypotenuse c, just as in the Pythagorean theorem. Of course, on the one hand, the area of the large square is c^2. On the other hand, the area of the large square is the sum of the areas of its component pieces.

Thus we calculate that

$$\begin{aligned} c^2 &= \text{(area of large square)} \\ &= \text{(area of triangle)} + \text{(area of triangle)} + \\ &\quad \text{(area of triangle)} + \text{(area of triangle)} + \\ &\quad \text{(area of small square)} \\ &= \frac{1}{2} \cdot ab + \frac{1}{2} \cdot ab + \frac{1}{2} \cdot ab + \frac{1}{2} \cdot ab + (b-a)^2 \\ &= 2ab + [a^2 - 2ab + b^2] \\ &= a^2 + b^2 \, . \end{aligned}$$

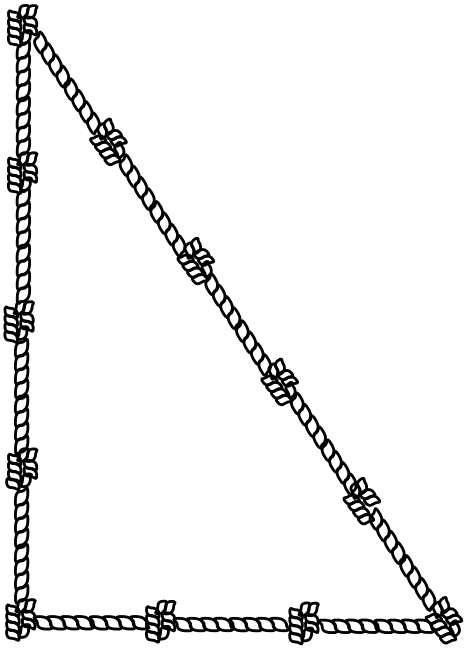


Figure 1.6

That proves the Pythagorean theorem. □

It is amusing to note that (according to legend) the Egyptians had, as one of their standard tools, a rope with twelve equally spaced knots. They used this rope to form a triangle with sides 3, 4, 5—see Figure 1.6. In this way they took advantage of the Pythagorean theorem to construct right angles.

Now Pythagoras noticed that, if $a = 1$ and $b = 1$, then $c^2 = 2$. He wondered whether there was a rational number c that satisfied this last identity. His stunning conclusion was this:

Theorem: *There is no rational number c such that $c^2 = 2$.*

Proof: Suppose that the conclusion is false. Then there *is* a rational number $c = \alpha/\beta$, expressed in lowest terms (i.e., α and β are integers with no factors in common) such that $c^2 = 2$. This translates to

$$\frac{\alpha^2}{\beta^2} = 2$$

or

$$\alpha^2 = 2\beta^2 \, .$$

We conclude that the righthand side is even, hence so is the lefthand side. Therefore α is even so $\alpha = 2m$ for some integer m.

But then

$$(2m)^2 = 2\beta^2$$

or

$$2m^2 = \beta^2.$$

We see that the lefthand side is even, so β^2 is even. Hence β is even.

But now both α and β are even—the two numbers have a common factor of 2. This statement contradicts the hypothesis that α and β have no common factors. Thus it cannot be that c is a rational number. Instead, c must be irrational. □

The Pythagoreans realized the profundity and the potential social importance of this discovery. It was ingrained in the ancient Greek consciousness that all numbers were rational. To claim the contrary would have been virtually heretical. For a time the Pythagoreans kept this new fact a secret. Ultimately, so legend has it, the Pythagoreans were destroyed by (ignorant) maurauding hordes.

Chapter 2

The Middle Ages and an Emphasis on Calculation

In any particular theory there is only as much real science as there is mathematics.

Immanuel Kant

This is the essence of what be termed formal understanding. We know that the results are true because we have gone through the crucible of the mathematical process and what remains is the essence of the truth.

J. Borwein, P. Borwein, R. Girgensohn, and S. Parnes

For, compared with the immense expanse of modern mathematics, what would the wretched remnants mean, the few isolated results incomplete and unrelated, that the intuitionists have obtained?

David Hilbert

Many would argue that the tricky concept of shared insight rather than logical precision is what mathematical communication is about. But since there are no absolute canons of rigour, and it is impossible to insist that every paper be written so that a (remarkably) patient graduate student can follow it, some mistakes are inevitably published. This observation does not render the proposed remedy nugatory, but it suggests that we shall still be working in an imperfect world.

Jeremy J. Gray

The sequence for the understanding of mathematics may be:

intuition, trial, error, speculation, conjecture, proof.

The mixture and the sequence of these events differ widely in different domains, but there is general agreement that the end product is rigorous proof—which we know and can recognize, without the formal advice of the logicians.

Saunders Mac Lane

Euclidean Methodology has developed a certain obligatory style of presentation. I shall refer to this as "deductivist style." This style starts with a painstakingly stated list of axioms, lemmas and/or definitions. The axioms and definitions frequently look artificial and mystifyingly complicated. One is never told how these complications arose. The list of axioms and definitions is followed by the carefully worded theorems. These are loaded with heavy-going conditions; it seems impossible that anyone should ever have guessed them. The theorem is followed by the proof.

Imre Lakatos

2.1 The Arabs and Algebra

In the early seventh century, the Muslims formed a small and persecuted sect. But by the end of that century, under the inspiration of Mohammed's leadership, they had conquered lands from India to Spain—including parts of North Africa and southern Italy. It is believed that, when Arab conquerors settled in new towns, they would contract diseases which had been unknown to them in desert life. In those days the study of medicine was confined mainly to Greeks and Jews. Encouraged by the caliphs (the local Arab leaders), these doctors settled in Baghdad, Damascus, and other cities. Thus we see that a purely social situation led to the contact between two different cultures which ultimately led to the transmission of mathematical knowledge.

Around the year 800, the caliph Haroun Al Raschid ordered many of the works of Hippocrates, Aristotle, and Galen to be translated into Arabic. Much later, in the twelfth century, these Arab translations were further translated into Latin so as to make them accessible to the Europeans. Today we credit the Arabs with preserving the grand Greek tradition in mathematics and science. Without their efforts, much of this classical work would have been lost.

2.2 The Development of Algebra

2.2.1 Al-Khwarizmi and the Basics of Algebra

There is general agreement that the rudiments of algebra found their genesis with the Hindus. Particularly Arya-Bhata in the fifth century and Brah-

magupta in the sixth and seventh centuries played a major role in the development of these ideas. Notable among the developments due to these men is the summation of the first N positive integers, and also the sum of their squares and their cubes (see our discussion of these matters in Chapter 9).

But the Arab expansion two hundred years later caused the transfer of these ideas to the Arab empire, and a number of new talents exerted considerable influence on the development of these concepts. Perhaps the most illustrious and most famous of the ancient Arab mathematicians was Abu Ja'far Muhammad ibn Musa Al-Khwarizmi (780 C.E.–850 C.E.). In 830 this scholar wrote an algebra text that became the definitive work in the subject. Called *Kitab fi al-jabr wa'l-mugabala*, it introduced the now commonly used term "algebra" (from "al-jabr"). The word "jabr" referred to the balance maintained in an equation when the same quantity is added to both sides (curiously, the phrase "al-jabr" also came to mean "bonesetter"); the word "mugabala" refers to cancelling like amounts from both sides of an equation.

Al-Khwarizmi's book *Art of Hindu Reckoning* introduced the notational system that we now call Arabic numerals: 1, 2, 3, 4, ... Our modern word "algorithm" was derived from a version of Al-Khwarizmi's name, and certainly some of his ideas contributed to the concept.

It is worth noting that good mathematical notation can make the difference between an idea that is clear and one that is obscure. The Arabs, like those who came before them, were hindered by lack of notation. When they performed their algebraic operations and solved their problems, they referred to everything with *words*. The scholars of this period are fond of saying that the Arabic notation was "rhetorical", with no symbolism of any kind. As an instance, we commonly denote the unknown in an algebraic equation by x; the Arabs would call that unknown "thing".

Moreover, the Arabs would typically exhibit their solutions to algebraic problems using geometric figures. They did not have an efficient method for simply writing the solution as we would today. It is clear that Al-Khwarizmi had very clearly formulated ideas about algorithms for solving polynomial equations. But he did not have the notation to write the solutions down as we now do. He certainly did not have the intellectual equipment (i.e., the formalism and the language) to formulate and prove theorems.

2.2.2 The Life of Al-Khwarizmi

Abu Ja'far Muhammad ibn Musa Al-Khwarizmi (780–850) was likely born in Baghdad, now part of Iraq. The little that we know about his life is based in part on surmise, and interpretation of evidence.

The "Al-Khwarizmi" in his name suggests that he came from Khwarizm, south of the Aral Sea in central Asia. But we also have this from an historian (Toomer) of the period:

> But the historian al-Tabari gives him the additional epithet "al-Qutrubbulli", indicating that he came from Qutrubbull, a district between the Tigris and Euphrates not far from Baghdad, so perhaps his ancestors, rather than he himself, came from Khwarizm ... Another epithet given to him by al-Tabari, "al-Majusi", would seem to indicate that he was an adherent of the old Zoroastrian religion. ... the pious preface to Al-Khwarizmi's "Algebra" shows that he was an orthodox Muslim, so Al-Tabari's epithet could mean no more than that his forebears, and perhaps he in his youth, had been Zoroastrians.

We begin our tale of Al-Khwarizmi's life by describing the context in which he developed. Harun al-Rashid became the fifth Caliph of the Abbasid dynasty on 14 September 786, at about the time that Al-Khwarizmi was born. Harun ruled in Baghdad over the Islam empire—which stretched from the Mediterranean to India. He brought culture to his court and tried to establish the intellectual disciplines which at that time were not flourishing in the Arabic world. He had two sons, al-Amin the eldest and al-Mamun the youngest. Harun died in 809, thus engendering a war between the two sons.

Al-Mamun won the armed struggle and al-Amin was defeated and killed in 813. Thus al-Mamun became Caliph and ruled the empire. He continued the patronage of learning started by his father and founded an academy called the House of Wisdom where Greek philosophical and scientific works were translated. He also built up a library of manuscripts, the first major library to be set up since that at Alexandria (which, according to one version of the story, was destroyed by invading hordes). His mission was to collect important works from Byzantium. In addition to the House of Wisdom, al-Mamun set up observatories in which Muslim astronomers could build on the knowledge acquired in the past.

Al-Khwarizmi and his colleagues, collectively called the Banu Musa, were scholars at the House of Wisdom in Baghdad. Their tasks there involved the translation of Greek scientific manuscripts; they also studied, and wrote on, algebra, geometry and astronomy. Certainly Abu Ja'far Al-Khwarizmi worked with the patronage of Al-Mamun; he dedicated two of his texts to the Caliph. These were his treatise on algebra and his treatise on astronomy. The algebra treatise *Kitab fi al-jabr wa'l-mugabala* was the most famous and significant of all of Al-Khwarizmi's works. It is, in an important historical sense, the very first book on algebra.

Al-Khwarizmi himself tells us that the significance of his book is:

> ... what is easiest and most useful in arithmetic, such as men constantly require in cases of inheritance, legacies, partition, lawsuits, and trade, and in all their dealings with one another, or where the measuring of lands, the digging of canals, geometrical computations, and other objects of various sorts and kinds are concerned.

It should be remembered that it was typical of early mathematics tracts that they concentrated on, and found their motivation in, practical problems. Al-Khwarizmi's work was no exception. His motivations and his interests may have been abstract, but his presentation was very practical. Al-Khwarizmi had a very good sense of his audience.

Early in the book Al-Khwarizmi describes the natural numbers in terms that are somewhat clumsy to us today. But we must acknowledge the new abstraction and profundity of what he was doing:

> When I consider what people generally want in calculating, I found that it always is a number. I also observed that every number is composed of units, and that any number may be divided into units. Moreover, I found that every number which may be expressed from one to ten, surpasses the preceding by one unit: afterwards the ten is doubled or tripled just as before the units were: thus arise twenty, thirty, etc. until a hundred: then the hundred is doubled and tripled in the same manner as the units and the tens, up to a thousand; ... so forth to the utmost limit of numeration.

A careful reading of this passage reveals the foundations of the base-10 arithmetic (i.e., the decimal system, with the arabic numerals) that we commonly

use today.

We should bear in mind that, for many centuries, the motivation for the study of algebra was the solution of equations. In Al-Khwarizmi's day these were linear and quadratic equations. His equations were composed of units, roots and squares. However, it is both astonishing and significant to bear in mind that Al-Khwarizmi did his algebra with no symbols—only words.

Al-Khwarizmi first reduces an equation (linear or quadratic) to one of six standard forms (which we describe both in words and in modern notation):

1. Squares equal to roots.

2. Squares equal to numbers.

3. Roots equal to numbers.

4. Squares and roots equal to numbers; e.g., $x^2 + 10x = 39$.

5. Squares and numbers equal to roots; e.g., $x^2 + 21 = 10x$.

6. Roots and numbers equal to squares; e.g., $3x + 4 = x^2$.

The reduction is carried out using the two operations of "al-jabr" and "al-muqabala". Here "al-jabr" means "completion" and is the process of removing negative terms from an equation. For example, using one of Al-Khwarizmi's own examples, "al-jabr" transforms $x^2 = 40x - 4x^2$ (in modern notation) into $5x^2 = 40x$ (in modern notation). The term "al-muqabala" means "balancing" and is the process of reducing positive terms of the same power when they occur on both sides of an equation. For example, two applications of "al-muqabala" reduce $50 + 3x + x^2 = 29 + 10x$ (in modern notation) to $21 + x^2 = 7x$ (one application to deal with the numbers and a second to deal with the roots).

Al-Khwarizmi then shows how to solve the six types of equations indicated above. He uses both algebraic methods of solution and geometric methods. We shall treat his geometric methodology, as well as some of his algebraic ideas, in detail below. Al-Khwarizmi continues his study of algebra in *Kitab fi al-jabr wa'l-mugabala* by considering how the laws of arithmetic extrapolate to an algebraic context. For example, he shows how to multiply out expressions such as

$$(a + bx)(c + dx)\,.$$

Again we stress that Al-Khwarizmi uses only words to describe his expressions; no symbols are in evidence.

There seems to be little doubt, from our modern perspective, that Al-Khwarizmi was one of the greatest mathematicians of all time. His algebra was original, incisive, and profound. It truly changed the way that we think about mathematics.

The next part of Al-Khwarizmi's Algebra consists of applications and worked examples. He then goes on to look at rules for finding the area of figures such as the circle and also finding the volume of solids such as the sphere, cone, and pyramid. His section on mensuration certainly has more in common with Hindu and Hebrew texts than it does with any Greek work. The final part of the book deals with the complicated Islamic rules for inheritance but requires little from the earlier algebra beyond techniques for solving linear equations. Again, in all these aspects of the book, we see the over-arching need to justify the mathematics with practical considerations.

Al-Khwarizmi also wrote a treatise on Hindu-Arabic numerals. The Arabic text is lost but a Latin translation, *Algoritmi de numero Indorum* (rendered in English, the title is *Al-Khwarizmi on the Hindu Art of Reckoning*) gave rise to the word "algorithm", deriving from his name in the title. The work describes the Hindu place-value system of numerals based on 1, 2, 3, 4, 5, 6, 7, 8, 9, and 0. The first use of zero as a place holder in positional base notation was probably due to Al-Khwarizmi in this work. Methods for arithmetical calculation are given, and a method for finding square roots is known to have been in the Arabic original although it is missing from the Latin version.

Another important work by Al-Khwarizmi was his book *Sindhind zij* on astronomy. The tome is based on Indian astronomical tracts. Most later Islamic astronomical handbooks, by contrast, utilized the Greek planetary models laid out in Ptolemy's *Almagest*.

The Indian text on which Al-Khwarizmi based his treatise was one which had been given to the court in Baghdad around 770 as a gift from an Indian political mission. There are two versions of Al-Khwarizmi's work which he wrote in Arabic but both are lost. In the tenth century al-Majriti made a critical revision of the shorter version and this was translated into Latin by Abelard of Bath. The main topics covered by Al-Khwarizmi in the *Sindhind zij* are: calendars; calculating true positions of the sun, moon and planets; tables of sines and tangents; spherical astronomy; astrological tables; parallax and eclipse calculations; and visibility of the moon. A related manuscript,

attributed to Al-Khwarizmi, concerns spherical trigonometry.

Although his astronomical work is based on that of the Indians, and most of the values from which he constructed his tables came from Hindu astronomers, Al-Khwarizmi must have been influenced by Ptolemy's work too.

Al-Khwarizmi wrote a major work on geography which gives latitudes and longitudes for 2402 localities as a basis for a world map. The book, which is based on Ptolemy's *Geography*, lists—with latitudes and longitudes—cities, mountains, seas, islands, geographical regions, and rivers. The manuscript does include maps which on the whole are more accurate than those of Ptolemy. In particular it is clear that, where more local knowledge was available to Al-Khwarizmi such as the regions of Islam, Africa and the Far East, his work is considerably more accurate than that of Ptolemy, but for Europe Al-Khwarizmi seems to have used Ptolemy's data.

Several minor works were written by Al-Khwarizmi on the astrolabe (on which he wrote two works), on the sundial, and on the Jewish calendar. He also wrote a political history containing horoscopes of prominent persons.

We have already discussed the varying views of the importance of Al-Khwarizmi's algebra which was his most important contribution to mathematics. Al-Khwarizmi is perhaps best remembered by Mohammad Kahn:

> In the foremost rank of mathematicians of all time stands Al-Khwarizmi. He composed the oldest works on arithmetic and algebra. They were the principal source of mathematical knowledge for centuries to come in the East and the West. The work on arithmetic first introduced the Hindu numbers to Europe, as the very name algorithm signifies; and the work on algebra ... gave the name to this important branch of mathematics in the European world ...

2.2.3 The Ideas of Al-Khwarizmi

All these ideas are perhaps best illustrated by some examples.

EXAMPLE 2.2.1 Solve this problem of Al-Khwarizmi:

> A square and ten roots equal thirty-nine dirhems.

Solution: It requires some effort to determine what is being asked. First, a *dirhem* is a unit of money in medieval Arabic times. In modern English (we shall introduce some mathematical *notation* later), what Al-Khwarizmi is telling us is that a certain number squared plus ten times that number (by "root" he means the number that was squared) equals 39. If we call this unknown number x, then what is being said is that

$$x^2 + 10x = 39$$

or

$$x^2 + 10x - 39 = 0\,.$$

Of course the quadratic formula quickly tells us that

$$x = \frac{-10 \pm \sqrt{10^2 - 4 \cdot (-39) \cdot 1}}{2} = \frac{-10 \pm \sqrt{256}}{2} = \frac{-10 \pm 16}{2}\,.$$

Alternatively one could factor the polynomial equation as

$$(x - 3)(x + 13) = 0\,.$$

Thus the roots are 3 and -13.

Now the Arabs could not deal with negative numbers, and in any event Al-Khwarizmi was thinking of his unknown as the side of a square. So we take the solution

$$x = \frac{-10 + 16}{2} = 3\,.$$

Thus, from our modern perspective, this is a straightforward problem. We introduce a variable, write down the correct equation, and solve it using a standard formula.

Matters were different for the Arabs. They did not have notation, and certainly did not yet know the quadratic formula. Their method was to deal with these matters geometrically. Consider Figure 2.1. This shows the "square" mentioned in the original problem, with unknown side length that we now call x. In Figure 2.2, we attach to each side of the square a rectangle of length x and width 2.5. The reasoning here is that Al-Khwarizmi tells us to add on 10 times the square's side length. We divide 10 into four pieces and thus add four times "2.5 times the side length". The quantity "2.5 times the side length" is represented by an appropriate rectangle in Figure 2.2.

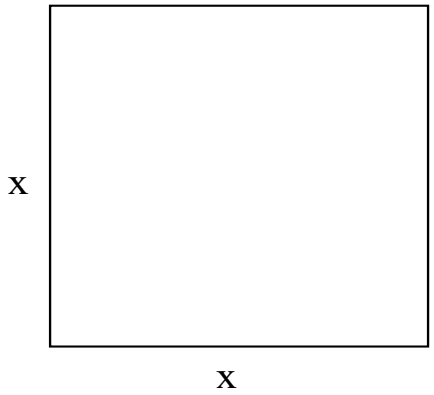


Figure 2.1

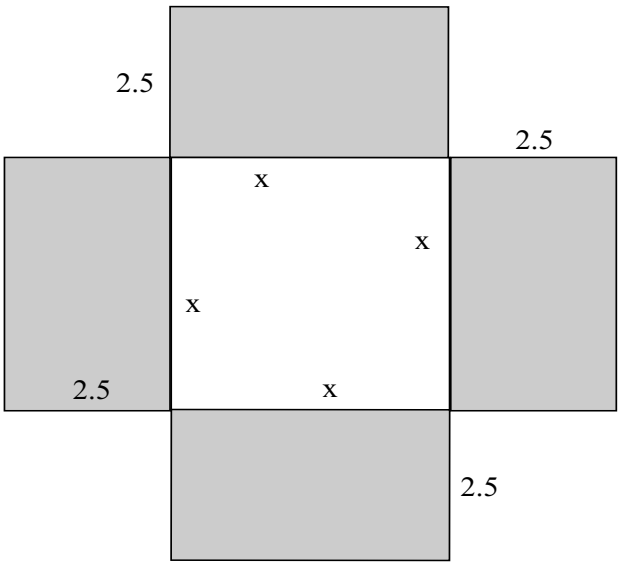


Figure 2.2. Sum of shaded areas is $10 \times x$.

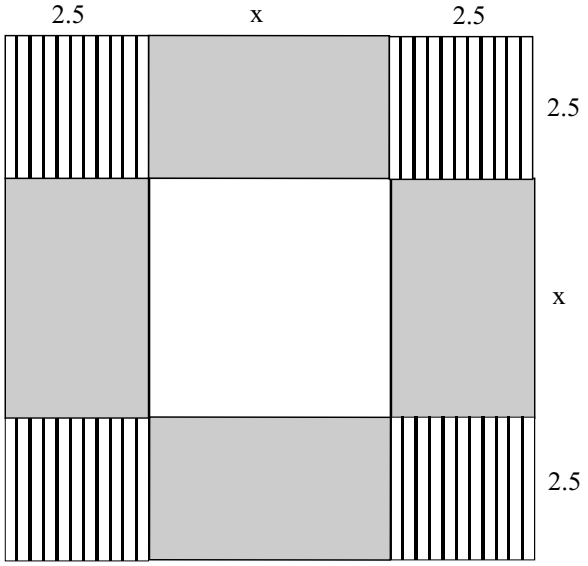


Figure 2.3. Area of large, inclusive square is 64.

Now we know, according to the statement of the problem, that the sum of the areas of the square in the middle and the four rectangles around the sides is 39. We handle this situation by filling in four squares in the corners—see Figure 2.3. Now the resulting large square plainly has area equal to $39 + 2.5^2 + 2.5^2 + 2.5^2 + 2.5^2 = 64$. Since the large square has area 64, it must have side length 8. But we know that each of the squares in the four corners has side length 2.5. It must follow then that $x = 8 - 2.5 - 2.5 = 3$. And that is the correct answer. □

In fact the method of this last example can be used to solve any quadratic equation with positive, real roots.

Now we examine another algebra problem of Al-Khwarizmi. This is in the format of a familiar sort of word problem. It has interesting social as well as mathematical content. We shall present the solution both in modern garb and in the argot of Al-Khwarizmi's time.

EXAMPLE 2.2.2 Solve this problem of Al-Khwarizmi:

> A man dies leaving two sons behind him, and bequeathing one-fifth of his property and one dirhem to a friend. He leaves ten

> dirhems in property and one of the sons owes him ten dirhems. How much does each legatee receive?

Solution: We already know that a dirhem is a unit of currency. It is curious that, in Al-Khwarizmi's time, there was no concept of "estate". A legacy could only be left to a person or people, not to an abstraction like an "estate".

However *we do* understand what an estate is, and it helps us to solve the problem in modern language. Our solution goes as follows. The dead man's estate consists of 20 dirhems: the 10 dirhems that he has in hand and the 10 dirhems owed to him by his son. The friend receives 1/5 of that estate plus one dirhem. Thus the friend receives $4 + 1 = 5$ dirhems. That leaves the estate with 5 dirhems in hand (the one son owing another 10 dirhems to the estate) and 10 dirhems owed to it, for a total of 15 dirhems. Thus each son is owed 7.5 dirhems. That means that the son who owes 10 dirhems should pay the estate 2.5 dirhems. Now the estate has 7.5 dirhems cash in hand. And that amount is paid to the other son.

Since Al-Khwarizmi did not have the abstraction of "estate" to aid his reckoning, he solved the problem with the following reasoning:

> Call the amount taken out of the debt *thing*. Add this to the property. The sum is 10 dirhems plus *thing*. Subtract 1/5 of this, since he has bequeathed 1/5 of his property to the friend. The remainder is 8 dirhems plus 4/5 of *thing*. Then subtract the 1 dirhem extra that is bequeathed to the friend. There remain 7 dirhems and 4/5 of *thing*. Divide this between the two sons. The portion of each of them is three and one half dirhems plus 2/5 of *thing*. Then you have 3/5 of *thing* equal to three and one half dirhems. Form a complete *thing* by adding to this quantity 2/3 of itself. Now 2/3 of three and one half dirhems is two and one third dirhems. Conclude that *thing* is five and five sixths dirhems.

□

2.2.4 Concluding Thoughts about the Arabs

This section has given an incisive picture of Arab mathematics. The Arab contributions have been extremely influential, even to modern times. But

one of the conclusions that one must draw is that medieval Arab algebra was *not* theoretical.[1] The Arabs did not consider proofs, nor did they treat any type of mathematics that would have required proofs. It may be argued that Al-Khwarizmi had a good theoretical grasp of quadratic equations. He certainly could solve any quadratic equation with positive, real roots. But he did not have the language or the notation to cogently express his ideas (at least from a modern perspective). Al-Khwarizmi did not know about negative numbers, and he had no conception of complex numbers. These are both rather theoretical notions, and beyond what mathematicians could do in his day. So there were many quadratic equations that Al-Khwarizmi could not solve, nor even consider. But his work was systematic and complete in the context that he defined for himself.

Certainly modern algebra is a highly recondite subject in which everything is proved. In fact abstract algebraists at the best universities today tend to shy away from concrete examples, and deal only in theory and proofs.[2] As our examples have illustrated, the basic algebra of the medieval Arabs was grounded in practical problems; the main goal was to arrive at a specific answer. It was left for modern times—the past two hundred years—to explore the theory behind algebra. Certainly the Arabs *might* have asked how many solutions a polynomial equation of degree n will have. But they did not. They might have asked whether every polynomial equation has a root (in the complex numbers) but they did not. These questions were ultimately resolved by the fundamental theorem of algebra (proved by Gauss in his doctoral dissertation) and the Euclidean algorithm.

2.3 Investigations of Zero

The concept of zero has a long and colorful history. As early as 600 B.C.E., the Babylonians had a symbol in their arithmetic for zero. But they made it very clear that there were doubts about what "zero" actually meant. For how could a definite and explicit symbol stand for nothing?

[1] On the other hand, the Arab ideas about geometry were rather more sophisticated. They contributed some quite perspicacious analyses of Euclid' *Elements*, for example.

[2] However Gröbner bases are enjoying something of a vogue, and their study is quite concrete. The software `Macauley` (which relies heavily on Gröbner bases) is a powerful calculational tool, and has changed the way that many mathematicians explore abstract mathematical ideas.

The ancient Greeks, somewhat later, were obsessed with questions of the existence of matter, and the essence of being. The concept of zero actually offended the Greeks.

In the middle ages zero took on religious overtones. The concept of nothing seemed to have connections to the soul, and to spirituality. In many contexts it was forbidden to discuss zero. People feared committing heresy. It was not until the sixteenth century that zero began to play a useful role in commerce. Obviously it will happen that a merchant will sell all his units of product A. It is thus helpful to be able to write that "zero units of A remain in stock." Gradually, the idea of zero was incorporated into the standard mathematical argot. The objections faded away.

As one might imagine, there were similar but perhaps more vigorous objections to the negative numbers. Over time, it was realized that negative numbers could play a useful role in commerce (as when a merchant is in debt), so negative numbers were gradually incorporated into the mathematical canon.

Today we use zero and the negative numbers with comfort and ease. We can actually *construct* number systems containing 0 and -5, so there is no longer any mystery of where these numbers come from. Also the religious questions have faded into the background. So there remain few objections to these extended concepts of number.

It may be said that mathematical abstraction, and *proof*, have actually helped in the acceptance of the concepts of zero and the negative numbers. In the middle ages, when we could not say how these mysterious numbers arose, or precisely what they meant, these numbers had an extra air of mystery attached to them. Much thought of the time was influenced by religion, and people feared that things that they did not understand were the works of the devil. Also the notion of a symbol that stood for nothing raised specters of evil signs and works of Satan. At various times, and by various people, it was actually *forbidden* to give explicit mention to zero or negative numbers. They were sometimes referred to explicitly in print as "forbidden" or "evil". The book [KAP1] gives a colorful history of the concept of zero. The Kaplans demonstrate artfully that zero is an important and sometimes controversial concept, and one that has had considerable influence over human thought.

As we have indicated, it is mathematical theory that puts zero on a solid footing. It is possible to use the theory of equivalence classes (see [KRA5]) to *construct* the integer number system—in effect to *prove* that it exists, and has all the expected properties. Thus it becomes clear that this number system is

a construct of human intellect, not of the devil or of some evil force. Since the construction involves only pure thought, and follows the time-tested rigorous lines of reasoning that are typical of the mathematical method, there is a cleansing process that makes the result palatable and non-threatening. It is a triumph for mathematics and its culture.

2.4 The Idea of Infinity

The historical struggles with infinity were perhaps more strident and fear-laden than those for zero. Again, the reasons were religious. For people felt that thinking about infinity was tantamount to thinking about the Creator. If one actually endeavored to *manipulate* infinity as a mathematical object—just as we routinely manipulate numbers and variables and other constructs in mathematics—then one was showing the utmost disrespect, and exhibiting a cavalier attitude towards the Deity. One feared being guilty of heresy or sacrilege.

Many prominent nineteenth-century mathematicians strictly forbade any discussion of infinity. Of course discussions of infinity were fraught with paradoxes and apparent contradictions. These suggested deep flaws in the foundations of mathematics. They only exacerbated people's fears and uncertainties. Today we have a much more sophisticated view of our discipline. We understand much more about the logical foundations of mathematics, and we have the tools for addressing (apparent) paradoxes and inconsistencies. In the nineteenth century mathematicians were not so equipped. They felt quite helpless when mathematics revealed confusions or paradoxes or lacunae; so they shied away from such disturbing phenomena. Certainly infinity was the source of much misunderstanding and confusion. It was a subject best avoided. Fear of religious fallout gave a convenient rationale for pursuing such a course.

It was the late nineteenth-century mathematician Georg Cantor (1845–1918) who finally determined the way to tame the infinity concept. He actually showed that there are many different "levels" or "magnitudes" of infinity. And he was able to prove some strikingly dramatic results—about classical topics like the transcendental numbers[3]—using his ideas about infinity.

[3]A real number is *algebraic* if it is the root of a polynomial equation with integer coefficients. A real number is *transcendental* if it is not algebraic. For example, the number $\sqrt{2}$ is algebraic because it is a root of the polynomial equation $x^2 - 2 = 0$. The

Cantor suffered vehement and mean-spirited attacks—even from his senior associate Leopold Kronecker (1823–1891)—over his ideas about infinity. Cantor in fact had a very complicated relationship with Kronecker, and it is not entirely clear how great was Kronecker's role in Cantor's problems. Today there is evidence that Cantor suffered from bipolar depression. Also Kronecker died 27 years before Cantor.

Cantor spent considerable time in asylums in an effort to cope with this calumny, and to deal with his subsequent depression. Near the end of his life, Cantor became disillusioned with mathematics. He spent a good portion of his final days in an effort to prove that Francis Bacon wrote the works of Shakespeare.

It is safe to say that Cantor's notions about infinity—his concept of *cardinality*, and his means of stratifying infinite sets of different orders or magnitudes—have been among the most profound and original ideas ever to be created in mathematics. These ideas have been universally embraced today; they are part of the bedrock of modern mathematics. But in Cantor's day the ideas were most controversial. Originality comes at a price, and Cantor was flying in the face of severe prejudice and fears founded in religious dogma. People capitalized on the perception that Cantor was a Jew, and many of their attacks were shamelessly antisemitic—even though Cantor was in fact *not* a Jew. It is easy to see how Cantor might have become depressed, discouraged, and unwilling to go on.

Towards the end of Cantor's life, Kronecker came around and began to support Cantor and his ideas. Other influential mathematicians also began to see the light, and to support the program for the study of infinity. But by then it was virtually too late (at least for poor Georg Cantor). Cantor was a broken man. He received full recognition and credit for his ideas after his death; but in life he was a tormented soul.

---

number π is *not* algebraic—it is transcendental—because it is not the root of any such polynomial equation. This last assertion is quite difficult to prove.

Chapter 3

The Dawn of the Modern Age

Every age has its myths and calls them higher truths.

Anonymous

I have resolved to quit only abstract geometry, that is to say, the consideration of questions that serve only to exercise the mind, and this, in order to study another kind of geometry, which has for its object the explanation of the phenomena of nature.

René Descartes

All men naturally desire knowledge.

Aristotle

There is nothing new to be discovered in physics now. All that remains is more and more precise measurement.

Lord Kelvin (1900)

. . . Atiyah and Mac Lane fell into a discussion, suited for the occasion, about how mathematical research is done. For Mac Lane it meant getting and understanding the needed definitions, working with them to see what could be calculated and what might be true, to finally come up with new "structure" theorems. For Atiyah, it meant thinking hard about a somewhat vague and uncertain situation, trying to guess what might be found out, and only then finally reaching definitions and the definitive theorems and proofs.

Saunders Mac Lane

It is now apparent that the concept of a universally accepted, infallible body of reasoning—the majestic mathematics of 1800 and the pride of man—is a grand illusion. Uncertainty and doubt concerning the future of mathematics have replaced the certainties and complacency of the past. The disagreement about the foundations of the "most certain" science are both surprising and, to put it mildly, disconcerting. The present state of mathematics is a mockery of the hitherto deep-rooted and widely reputed truth and logical perfection of mathematics.

Morris Kline

Standards in mathematics change, and later generations may regard as pedan-

tic and unnecessary those subtleties in arguments carried out by earlier generations with painstaking care. Once they have become accepted, they may later be taken for granted without careful examination of hypotheses. The later generation mathematician may put energy into understanding a different aspect of the problem.

Arthur Jaffe

Most ... explorations into the mathematical wilderness remain isolated illustrations. Heuristic conventions, pictures, and diagrams developing in one subfield often have little content for another. In each subfield, unproven results proliferate but remain conjectures, strongly held beliefs, or perhaps mere curiosities passed like folktales across the Internet.

J. Borwein, P. Borwein, R. Girgensohn, and S. Parnes

The truth is lies, and they jail all our prophets.

Charles Manson

3.1 Euler and the Profundity of Intuition

Leonhard Euler (1707–1783) was one of the greatest mathematicians who ever lived. He was also one of the most prolific. His collected works comprise more than seventy volumes, and they are still being edited today. Euler worked in all parts of mathematics, as well as mechanics, physics, and many other parts of science. He did his mathematics almost effortlessly, often while dandling a grandchild on his knee. Late in life he went partially blind, but he declared that this would help him to concentrate more effectively, and his scientific output actually *increased*.

Euler exhibited a remarkable combination of mathematical precision and mathematical intuition. He had some of the deepest insights in all of mathematics, and he also committed some of the most famous blunders. One of these involved calculating the infinite sum

$$1 - 1 + 1 - 1 + 1 - + \cdots .$$

In fact the *correct* way to analyze this sum is to add the first several terms and see whether a pattern emerges. Thus

$$\begin{aligned} 1 - 1 &= 0 \\ 1 - 1 + 1 &= 1 \end{aligned}$$

$$\begin{aligned} 1-1+1-1 &= 0 \\ 1-1+1-1+1 &= 1 \end{aligned}$$

and so forth. These "initial sums" are called *partial sums*. If the partial sums fall into a pattern, and tend to some limit, then we say that the original infinite sum (or *series*) converges. Otherwise we say that it diverges. What we see is that the partial sums for this particular series alternate between 0 and 1. They do not tend to any single, unique limit. So the series diverges.

But Euler did not know the correct way to analyze series (nor did anyone else in his day). His analysis went like this:

> If we group the terms of the series as
>
> $$(1-1)+(1-1)+(1-1)+\cdots$$
>
> then the sum is clearly $0+0+0+\cdots=0$. But if we group the terms of the series as
>
> $$1+(-1+1)+(-1+1)+(-1+1)+\cdots$$
>
> then the sum is clearly $1+0+0+0+\cdots=1$. Euler's conclusion was that this is a paradox.[1]

Euler is remembered today for profound contributions to number theory, geometry, the calculus of variations, and complex analysis. His blunders are all but forgotten. It was his willingness to take risks, and to make mistakes, that made him such an effective mathematician.

3.2 Dirichlet and the Heuristic Basis for Rigorous Proof

Peter Gustav Lejeune Dirichlet (1805–1859) was one of the great number theorists of the nineteenth century. His father's first name was Lejeune, coming from "Le jeune de Richelet". This means "young from Richelet." The Dirichlet family came from the town of Liège in Belgium. The father

[1] It may be noted that Euler also used specious reasoning to demonstrate that $1+2+4+8+\cdots=-1$.

was postmaster of Düren, the town where young Peter was born. At a young age Dirichlet developed a passion for mathematics; he spent his pocket money on mathematics books. He entered the Gymnasium in Bonn at the age of 12. There he was a model pupil. He exhibited an interest in history as well as mathematics.

After two years at the Gymnasium Dirichlet's parents decided that they would rather have him at the Jesuit College in Cologne. There he fell under the tutelage of the distinguished scientist Georg Ohm (1787–1854). By age 16, Dirichlet had completed his school work and was ready for the university. German universities were not very good, nor did they have very high standards, at the time. Hence Peter Dirichlet decided to study in Paris. It is worth noting that several years later the German universities would set the worldwide standard for excellence; Dirichlet himself would play a significant role in establishing this pre-eminence.

Dirichlet always carried with him a copy of Gauss's *Disquisitiones arithmeticae* (Gauss's masterpiece on number theory), a work that he revered and kept at his side much as other people might keep the Bible. Thus he came equipped for his studies in Paris. Dirichlet became ill soon after his arrival in Paris. But he would not let this deter him from attending lectures at the Collége de France and the Faculté des Sciences. He enjoyed the teaching of some of the leading scientists of the time, including Biot, Fourier, Francoeur, Hachette, Laplace, Lacroix, Legendre, and Poisson.

Beginning in the summer of 1823, Dirichlet lived in the house of the retired General Maximilien Sébastian Foy. He taught German to General Foy's wife and children. Dirichlet was treated very well by the Foy family, and he had time to study his mathematics.[2] It was at this time that he published his first paper, and it brought him instant fame. For it dealt with Fermat's last theorem. The problem is to show that the Diophantine[3] equation

$$x^n + y^n = z^n$$

has no integer solutions x, y, z when n is a positive integer greater than 2. The cases $n = 3, 4$ had already been handled by Euler and by Fermat himself. Dirichlet decided to attack the case $n = 5$. This case divides into

[2] Also the Foys introduced Dirichlet to Fourier, thus sparking Dirichlet's lifelong passion for Fourier series.

[3] A *Diophantine equation* is an algebraic equation for which we seek whole number solutions.

two subcases, and Dirichlet was able to dispatch Subcase 1. Legendre was a referee of the paper, and he was able, after reading Dirichlet's work, to treat Subcase 2. Thus a paper was published in 1825 that completely settled the case $n = 5$ of Fermat's last theorem. Dirichlet himself was subsequently able to develop his own proof of Subcase 2 using an extension of his techniques for Subcase 1. Later on Dirichlet was also able to treat the case $n = 14$.

General Foy died in November of 1825 and Dirichlet decided to return to Germany. However, in spite of support from Alexander von Humboldt, he could not assume a position in a German university since he had not submitted an Habilitation thesis. Dirichlet's mathematical achievements were certainly adequate for such a thesis, but he was not allowed to submit because **(i)** he did not hold a doctorate and **(ii)** he did not speak Latin.

The University of Cologne interceded and awarded Dirichlet an honorary doctorate. He submitted his Habilitation on polynomials with prime divisors and obtained a position at the University of Breslau. Dirichlet's appointment was still considered to be controversial, and there was much discussion among the faculty of the merits of the case.

Standards at the University of Breslau were still rather low, and Dirichlet was not satisfied with his position. He arranged to transfer, again with Humboldt's help, to the Military College in Berlin. He also had an agreement that he could teach at the University of Berlin, which was really one of the premiere institutions of the time. Eventually, in 1828, he obtained a regular professorship at the University of Berlin. He taught there until 1855. Since he retained his position at the Military College, he was saddled with an unusual amount of teaching and administrative duties.

Dirichlet also earned an appointment at the Berlin Academy in 1831. His improved financial circumstances then allowed him to marry Rebecca Mendelssohn, the sister of the noted composer Felix Mendelssohn. Dirichlet obtained an eighteen-month leave from the University of Berlin to spend time in Italy with Carl Jacobi (who was there for reasons of his health).

In 1845, Dirichlet returned to his duties at the University of Berlin and the Military College. He continued to find his duties at both schools to be a considerable burden, and complained to his student Kronecker. It was quite a relief when, on Gauss's death in 1855, Dirichlet was offered Gauss's distinguished chair at the University in Göttingen.

Dirichlet endeavored to use the new offer as leverage to obtain better conditions in Berlin. But that was not to be, and he moved to Göttingen directly. There he enjoyed a quieter life with some outstanding research

students. Unfortunately the new blissful conditions were not to be enjoyed for long. Dirichlet suffered a heart attack in 1858, and his wife died of a stroke shortly thereafter.

Dirichlet's contributions to mathematics were monumental. We have already described some of his work on Fermat's last problem. He also made contributions to the study of Gauss's quadratic reciprocity law. It can be said that Dirichlet was the father of the subject of analytic number theory. In particular, he proved foundational results about prime numbers occurring in arithmetic progression.[4]

Dirichlet had a powerful intuition, and it guided his thoughts decisively as he developed his mathematics. But he was widely recognized for the precision of his work. No less an eminence than Carl Jacobi (1804–1851) said

> If Gauss says he has proved something, it seems very probable to me; if Cauchy says so, it is about as likely as not; if Dirichlet says so, it is certain. I would gladly not get involved in such delicacies.

Dirichlet did further work on what was later to become (in the hands of Emmy Noether) the theory of *ideals*. He created *Dirichlet series*, which are today a powerful tool for analytic number theorists. And he laid some of the foundations for the theory of *class numbers* (later to be developed by Emil Artin).

Dirichlet is remembered for giving one of the first rigorous definitions of the concept of function. He was also among the first to define—at least in the context of Fourier series—precisely what it means for a series to *converge*

[4]Only recently, in [GRT], Benjamin Green and Terence Tao proved that there are arbitrarily long arithmetic progressions of prime numbers. An arithmetic progression is a sequence of numbers that are equally spaced apart. For example,

3 5 7

is a sequence of three primes, spaced 2 apart. This is an arithmetic progression of primes of *length 3*. As a second example,

11 17 23 29

is a sequence of four primes spaced 6 apart. This is an arithmetic progression of *length 4*. See whether you can find an arithmetic progression of primes of length 5. Suffice it to say that Green and Tao used very abstract methods to establish that there are arithmetic progressions of primes of any length.

(Cauchy dealt with this issue somewhat earlier). He is remembered as one of the fathers of the theory of Fourier series.

Dirichlet had a number of historically important students, including Kronecker and Riemann. Riemann went on to make seminal contributions to complex variables, Fourier series, and geometry.

3.3 The Pigeonhole Principle

Today combinatorics and number theory and finite mathematics are thriving enterprises. Cryptography, coding theory, queuing theory, and theoretical computer science all make use of counting techniques. But the idea of "counting", as a science, is relatively new.

Certainly Peter Lejeune Dirichlet was one of the greatest workers in number theory to ever live. Many theorems and ideas in that subject are named after him. But he was in fact loathe to spend time finding rigorous proofs of his new discoveries. He generally proceeded with a keen intuition and a profound grasp of the main ideas. But he left it to others, and to future generations, to establish the results rigorously.

Dirichlet was one of the first masters of the theory of counting. And one of his principal counting techniques, the one for which he is most vividly remembered, is that which was originally called the "Dirichletscher Schubfachschluss" (Dirichlet's drawer inference principle). Today we call it the "pigeonhole principle". It is a remarkably simple idea that has profound consequences. The statement is this:

> If you put $n + 1$ letters into n mailboxes then some mailbox will contain two letters.

This is quite a simple idea (though it *can* be given a rigorous mathematical proof—see Chapter 11). It says, for instance, that if you put 101 letters into 100 mailboxes then some mailbox will contain two letters. The assertion makes good intuitive sense. This pigeonhole principle turns out to be a terrifically useful mathematical tool. As a simple instance, if you have thirteen people in a room, then two of them will have their birthdays in the same month. Why? Think of the months as mailboxes and the birthdays as letters.

Dirichlet in fact applied his pigeonhole principle to prove deep and significant facts about number theory. Here is an important result that bears his name, and is frequently cited today:

> **Theorem:** Let ξ be a real number. If $n > 0$ is an integer, then there are integers p, q such that $0 < q \leq n$ and
>
> $$\left| \frac{p}{q} - \xi \right| < \frac{1}{q^2} \,. \qquad (\dagger)$$

It is a key idea in number theory that irrational numbers, and transcendental numbers, may be characterized by the rate at which they can be approximated by rational numbers. This entire circle of ideas originates with Dirichlet's theorem, and the root of that theorem is the simple but profound pigeonhole principle.

3.4 The Golden Age of the Nineteenth Century

Nineteenth-century Europe was a haven for brilliant mathematics. So many of the important ideas in mathematics today grew out of ideas that were developed at that time. We list just a few of these:

- Jean Baptiste Joseph Fourier (1768–1830) developed the seminal ideas for Fourier series and created the first formula for the expansion of an arbitrary function into a trigonometric series. He developed applications to the theory of heat.

- Evariste Galois (1812–1832) and Augustin Louis-Cauchy (1789–1857) laid the foundations for abstract algebra by inventing group theory.

- Bernhard Riemann (1826–1866) established the subject of differential geometry, defined the version of the integral (from calculus) that we use today, and made profound contributions to complex variable theory and Fourier analysis.

- Augustin-Louis Cauchy laid the foundations of complex variable theory and partial differential equations. He also did seminal work in geometric analysis.

- Carl Jacobi (1804–1851), Ernst Kummer (1810–1893), Niels Henrick Abel (1802–1829), and numerous other mathematicians from many countries developed number theory.

- Joseph Louis Lagrange (1736–1813), Cauchy and others were laying the foundations of the calculus of variations, classical mechanics, the implicit function theorem, and many other important ideas in modern geometric analysis.

- Karl Weierstrass (1815–1897) laid the foundations for rigorous analysis with numerous examples and theorems. He made seminal contributions both to real and to complex analysis.

This list could be expanded considerably. The nineteenth century was a fecund time for European mathematics, and communication among mathematicians was at an all-time high. There were several prominent mathematics journals, and important work was widely disseminated. The many great universities in Italy, France, Germany, and England (England's was driven by physics) had vigorous mathematics programs and many students. This was an age when the foundations for modern mathematics were laid.

And certainly the seeds of rigorous discourse were being sown at this time. The language and terminology and notation of mathematics was not quite yet universal, the definitions were not well established, and even the methods of proof were in development. But the basic methodology was in place and the mathematics of that time traveled reasonably well among countries and to the twentieth century and beyond. As we shall see below, Bourbaki and Hilbert set the tone for rigorous mathematics in the twentieth century. But the work of the many nineteenth-century geniuses paved the way for those pioneers.

Chapter 4

Hilbert and the Twentieth Century

In all questions of logical analysis, our chief debt is to Frege.

Bertrand Russell and Alfred North Whitehead

There is no religious denomination in which the misuse of metaphysical expressions has been responsible for so much sin as it has in mathematics.

Ludwig Wittgenstein

Mathematics is not necessarily characterized by rigorous proofs. Many examples of heuristic papers written by prominent mathematicians are given in [JAQ]; one can list many more papers of this kind. All these papers are dealing with mathematical objects that have a rigorous definition.

Albert Schwarz

Specifically, speculative theoretical reasoning in physics is usually strongly constrained by experimental data. If mathematics is going to contemplate a serious expansion in the amount of speculation which it supports (which could have positive consequences), it will have a serious and complementary need for the admission of new objective sources of data, going beyond rigorously proven theorems, and including computer experiments, laboratory experiments and field data. Put differently, the absolute standard of logically correct reasoning was developed and tested in the crucible of history. This standard is a unique contribution of mathematics to the culture of science. We should be careful to preserve it, even (or especially) while expanding our horizons.

James Glimm

The most vexing of issues ... is the communication of insights. Unlike most experimentalized fields, Mathematics does not have a vocabulary tailored to the transmission of condensed data and insight. As in most physics experiments, the amount of raw data obtained from mathematical experiment will, in general, be too large for anyone to grasp. The collected data need to be compressed and compartmentalized.

J. Borwein, P. Borwein, R. Girgensohn, and S. Parnes

The most fundamental precept of the mathematical faith is "thou shalt prove everything rigorously." While the practitioners of mathematics differ on their views of what constitutes rigorous proof, and there are fundamentalists who insist on even a more rigorous rigor than the one practiced by the mainstream, the belief in this principle could be taken as the "defining property" of "mathematician".

Doron Zeilberger

It is by logic we prove, it is by intuition that we invent.

Henri Poincaré

Logic, therefore, remains barren unless fertilised by intuition.

Henri Poincaré

4.1 David Hilbert

Along with Henri Poincaré (1854–1912) of France, David Hilbert (1862–1943) of Germany was the spokesman for early twentieth century mathematics. Hilbert is said to have been one of the last mathematicians to be conversant with the entire subject—from differential equations to geometry to logic to algebra. He exerted considerable influence over all parts of mathematics, and he wrote seminal texts in many of them. Hilbert had an important and profound vision for the rigorization of mathematics (one that was later dashed by work of Bertrand Russell, Kurt Gödel, and others), and he set the tone for the way that mathematics was to be practiced and recorded in our time.

Hilbert had many important students, ranging from Richard Courant (1888–1972) to Theodore von Kármán (1881–1963) (the father of modern aeronautical engineering) to Hugo Steinhaus (1887–1972) to Hermann Weyl (1885-1955). His influence was felt widely, not just in Germany but around the world. He certainly helped to establish Göttingen as one of the world centers for mathematics, and it continues to be so today.

One of Hilbert's real coups was to study the subject of algebraic invariants and to prove that there was a basis for these invariants. For several decades people had sought to prove this result by constructive means—by actually *writing down* the basis.[1] Hilbert established the result nonconstructively,

[1] A "basis" is a minimal generating set for an algebraic system.

essentially with a proof by contradiction. This was quite controversial at the time—even though proof by contradiction had been around at least since the time of Euclid. Hilbert's work put a great many mathematicians out of business, and established him rather quickly as a force to be reckoned with. Certainly Hilbert is remembered today for a great many mathematical innovations, one of which was his *Nullstellensatz*—one of the key algebraic tools that he developed for the study of invariants.

4.2 G. D. Birkhoff, Norbert Wiener, and the Development of American Mathematics

In the late nineteenth century and early twentieth century, American mathematics had something of a complex. Not a lot of genuine (abstract, rigorous) mathematical research was done in this country, and the pre-eminent mathematicians of Europe—the leaders in the field—looked down their collective noses at the paltry American efforts. Of course we all must grow where we are planted, and the American intellectual life was a product of its context. America was famous then, even as it is now, for being practical, empirical, rough-and-ready, and eager to embrace the next development—whatever it may be. America prides itself on being a no-holds-barred society, in which there is great social mobility and few if any obstacles to progress. Intellectual life in the nineteenth-century United States reflected those values. University education was very concrete and practical, and grounded in particular problems that came from engineering or other societal issues.

As a particular instance of the point made in the last paragraph, William Chauvenet was one of the intellectual mathematical leaders of nineteenth century America. He founded the U. S. Naval Academy in Annapolis and then he moved to Washington University in St. Louis (home of this author). In those days most American universities did not have mathematics departments. In fact, generally speaking, mathematics was part of the Astronomy Department. This fit rather naturally, from a practical point of view, because Astronomy involves a good deal of calculation and mathematical reckoning. It was Chauvenet who convinced the administration of Washington University to establish a free-standing Mathematics Department. And he was its first Chairman. Chauvenet went on to become the Chancellor of Washington University, and he played a decisive role in the institution's early develop-

ment.

So what sort of mathematical research did a scholar like William Chauvenet do? Perhaps the thing he is most remembered for is that he did all the calculations connected with the construction of the Eads Bridge (which spans the Mississippi River in downtown St. Louis). The Eads Bridge is a great example of the classical arch style of bridge design, and it still stands and is in good use today—both as a footbridge and for automobile traffic.

Of course the premiere mathematicians of Europe—Riemann and Weierstrass and Dirichlet and Gauss in Germany as well as Cauchy and Liouville and Hadamard and Poincaré in France—were spending their time developing the foundations of real analysis, complex analysis, differential geometry, abstract algebra, and other fundamental parts of modern mathematics. You would not find them doing calculations for bridges.[2] There was a real disconnect between European mathematics and American mathematics.

One of the people who bridged the gap between the two mathematical cultures was J. J. Sylvester. It happened that Johns Hopkins University in Baltimore was seeking a new mathematical leader, and the British mathematician Sylvester was brought to their attention. In fact it came about this way:

When Harvard mathematician Benjamin Peirce (1809–1880) heard that the Johns Hopkins University was to be founded in Baltimore, he wrote to the new president Daniel Coit Gilman (1831–1908) in 1875 as follows:

> Hearing that you are in England, I take the liberty to write you concerning an appointment in your new university, which I think would be greatly for the benefit of our country and of American science if you could make it. It is that of one of the two greatest geometers of England, J. J. Sylvester. If you enquire about him, you will hear his genius universally recognized but his power of teaching will probably be said to be quite deficient. Now there is no man living who is more luminary in his language, to those who have the capacity to comprehend him than Sylvester, provided the hearer is in a lucid interval. But as the barn yard fowl

[2] To be fair, Gauss—especially later in his career—took a particular interest in a variety of practical problems. A number of other European mathematicians famous for their abstract intellectual work also did significant applied work. But it is safe to say that their roots, and their focus, were in pure mathematics. This is the work that has stood the test of time, and for which they are remembered today.

> cannot understand the flight of the eagle, so it is the eaglet only who will be nourished by his instruction .... Among your pupils, sooner or later, there must be one, who has a genius for geometry. He will be Sylvester's special pupil—the one pupil who will derive from his master, knowledge and enthusiasm—and that one pupil will give more reputation to your institution than the ten thousand, who will complain of the obscurity of Sylvester, and for whom you will provide another class of teachers .... I hope that you will find it in your heart to do for Sylvester—what his own country has failed to do—place him where he belongs—and the time will come, when all the world will applaud the wisdom of your selection.

J. J. Sylvester (1814–1897) was educated in England. When he was still young, he accepted a professorship at the University of Virginia. One day a young member of the chivalry whose classroom recitation Sylvester had criticized became quite piqued with the esteemed scholar. He prepared an ambush and fell upon Sylvester with a heavy walking stick. Sylvester speared the student with a sword cane, which he just happened to have handy. The damage to the student was slight, but the professor found it advisable to leave his post and take the earliest possible ship to England. Sylvester took a position there at a military academy. He served long and well, but subsequently retired and accepted a position at Johns Hopkins University when he was in his late fifties. He founded the *American Journal of Mathematics* the following year, and was certainly the leading light of American mathematics in his day.

H. F. Baker (1866–1956) recounts the following history of J. J. Sylvester and the Johns Hopkins University:

> In 1875 the Johns Hopkins University was founded at Baltimore. A letter to Sylvester from the celebrated Joseph Henry (1797–1878), of date 25 August 1875, seems to indicate that Sylvester had expressed at least a willingness to share in forming the tone of the young university; the authorities seem to have felt that a Professor of Mathematics and a Professor of Classics could inaugurate the work of a University without expensive buildings or elaborate apparatus. It was finally agreed that Sylvester should go, securing besides his travelling expenses, an annual stipend of 5000 dollars "paid in gold." And so, at the age of sixty-one, still

> full of fire and enthusiasm, ... he again crossed the Atlantic, and did not relinquish the post for eight years, until 1883. It was an experiment in educational method; Sylvester was free to teach whatever he wished in the way he thought best; so far as one can judge from the records, if the object of an University be to light a fire of intellectual interests, it was a triumphant success. His foibles no doubt caused amusement, his faults as a systematic lecturer must have been a sore grief to the students who hoped to carry away note-books of balanced records for future use; but the moral effect of such earnestness ... must have been enormous.

J. J. Sylvester once remarked that Arthur Cayley had been much more fortunate than himself: "that they both lived as bachelors in London, but that Cayley had married and settled down to a quiet and peaceful life at Cambridge; whereas he [Sylvester] had never married, and had been fighting the world all his days." Those in the know attest that this is a fair summary of their lives.

Teaching is an important pursuit, and has its own rewards. So is research. But the two can work symbiotically together, and the whole created thereby is often greater than the sum of its parts. J. J. Sylvester, who was an eccentric teacher at best, describes the process in this way:

> But for the persistence of a student of this university in urging upon me his desire to study with me the modern algebra I should never have been led into this investigation; and the new facts and principles which I have discovered in regard to it (important facts, I believe), would, so far as I am concerned, have remained still hidden in the womb of time. In vain I represented to this inquisitive student that he would do better to take up some other subject lying less off the beaten track of study, such as the higher parts of the calculus or elliptic functions, or the theory of substitutions, or I wot [know] not what besides. He stuck with perfect respectfulness, but with invincible pertinacity, to his point. He would have the new algebra (Heaven knows where he had heard about it, for it is almost unknown in this continent [America]), that or nothing. I was obliged to yield, and what was the consequence? In trying to throw light upon an obscure explanation in our text-book, my brain took fire, I plunged with re-quickened zeal into a subject which I had for years abandoned, and found

> food for thoughts which have engaged my attention for a considerable time past, and will probably occupy all my powers of contemplation advantageously for several months to come.

J. J. Sylvester once gave a commencement address at Johns Hopkins. He began by remarking that mathematicians were not any good at that sort of thing because the language of mathematics was antithetical to general communication. That is to say, mathematics is very concise: one can express pages of thought in just a few symbols. Thus, since he was accustomed to mathematical expression, his comments would be painfully brief. He finished three hours later.

J. J. Sylvester once sent a paper to the London Mathematical Society for publication. True to form, he included a cover letter asserting that this was the most important result in the subject for 20 years. The Secretary replied that he agreed entirely with Sylvester's assessment, but that Sylvester had actually published the result in the *Journal of the London Mathematical Society* five years earlier.

The next big event, from our chauvinistic point of view as Americans, is that G. D. Birkhoff (1884–1944) came along. Educated at Harvard, he ended up on the faculty at Harvard. And he was the first native-born American to prove a theorem that caught the attention and respect of the European mandarins in the field. This was his proof of Poincaré's last theorem. Also his proof of the general ergodic theorem attracted considerable interest; it was a problem that had received broad and intense interest for many years. Birkhoff cracked it, and he thereby put himself, Harvard, and American mathematics on the map. Now America was a real player in the great mathematical firmament.

Of course it took more than just one man to earn the undying respect and admiration of the rather stodgy European mathematicians. So we can be grateful that Norbert Wiener (1894–1964) came on the scene. Wiener was a child prodigy, tutored by his martinet father (who was himself an academic). Norbert was born on the campus of the University of Missouri in Columbia, where his father was a professor of languages. The elder Wiener was on the losing end of some pivotal political battle, and as a result was forced to leave his position in Missouri. The family ended up moving to the Boston area. After floundering around for a while, Wiener père ended up on the faculty at Harvard. He remained there for the rest of his career.

Young Norbert, who was given a running start in his education by his

father's diligent attentions, began attending Tufts University at the age of eleven. He was at the time the youngest student in America ever to attend college. There was considerable press coverage of this event, and Norbert was an instant celebrity. Norbert only learned when he was seventeen years old that he was Jewish (prior to that his father had represented that the family was gentile). Wiener took this news very badly, and the fact of being Jewish plagued him for the rest of his life. He felt that anti-Semitic forces (G. D. Birkhoff notable among them) were in power in American mathematics. Thus, at the start of his mathematical career, Wiener lived in England. It was only there that he felt he could get a fair shake. He was finally able, through some careful politicking, to land a job at MIT. Wiener was to spend the rest of his professional life in Cambridge, Massachusetts.

Wiener was short, rotund, and extremely myopic. He cut quite a figure as he strode around the MIT campus. But he was very famous for his intellectual prowess. His classes had to be held in great halls, because people attended from all over the Boston area. It was rumored that Wiener's salary was higher than that of the President of MIT.

And Wiener put MIT and American mathematics (and himself, of course) on the map yet again. His work in Fourier analysis and stochastic integrals and cybernetics (a term, and a subject, that he invented) was groundbreaking, and his theorems are still studied and cited today.

From the point of view of mathematical proof, Norbert Wiener was a classicist. He formulated theorems in the traditional way, and proved them rigorously with pen and paper. But Wiener also maintained a considerable interest in the *applications* of mathematics, and in the way that scientists interact with society. He was deeply troubled by the use of the atomic bomb to end World War II in Asia. He campaigned against scientists lending their intellects to support the military.

Norbert Wiener was one of the fathers of the modern theory of *stochastic processes*. This is a branch of probability theory that analytically describes random processes, such as Brownian motion. In the late 1940s and early 1950s, all of Norbert Wiener's many interests converged in such a way that he invented an entirely new avenue of human inquiry. He called it "cybernetics".[3]

Cybernetics is the study of how man interacts with machines (particularly,

[3]The word *cybernetics* derives from the Greek *kubernêtai*, which means "steersman" or "helmsman".

but not exclusively, computers). It considers questions such as whether a machine can "think". It turned out that these were not merely questions of philosophical speculation. Wiener could analyze them using his ideas from stochastic processes. He wrote copious papers on cybernetics, and traveled the world spreading his gospel. And these ideas enjoyed some real currency in the 1950s. Wiener had quite a following, and at MIT he had a cybernetics lab with some top scholars as his co-workers.

It is safe to say that G. D. Birkhoff and Norbert Wiener were two of the key players who helped to put American mathematics into a prominent position in the twentieth century. They both had some truly important and influential students: Birkhoff's included Marston Morse (1892–1977), Hassler Whitney (1907–1989), and Marshall Stone (1903–1989); Wiener's included Amar Bose (of Bose Stereo fame) and Norman Levinson and Abe Gelbart. Certainly the opening of the Institute for Advanced Study in Princeton in 1930, with its galaxy of world-class mathematicians (and of course with superstar physicist Albert Einstein) helped to put American mathematics into the fore. The University of Chicago, founded with money from John D. Rockefeller in 1890 (the first classes were held in 1892), also became a bastion of American mathematical strength. It was soon followed by Princeton and Harvard, and later by MIT (thanks to the role model and considerable effort of Norbert Wiener).

Today America is unquestionably one of the world leaders in mathematics. There are several reasons for this pre-eminence. First of all, America has a good many first-class universities. Secondly, there is considerable government subsidy (through the National Science Foundation, the Department of Defense, the Department of Energy, and many other agencies) for mathematical research. But it might be noted that the American way of doing business has played a notable role in the development of mathematics. In America, if you are a hard-working and successful mathematician, you can really move ahead. You might begin with a humble Ph.D. and a modest job. But if you prove important theorems, then you will get better job offers. It is definitely possible to move up through the ranks—to more and more prestigious positions and more elite universities—and to get a better salary and a fancy job with more perks. The really top mathematicians in this country have extremely good salaries, discretionary funds, a stable of assistants (postdocs and Assistant Professors and others who work with them), and many other benefits.

This is not how things are in most other countries. In many of the leading

European countries, all education is centralized. In Italy all the decisions come out of Rome and in France they all come from Paris. This has certain benefits—in terms of nationalized standards and uniformity of quality. But it also makes the system stodgy and inflexible. When Lars Hörmander, a Swede, won the Fields Medal in 1962 there was no job for him in Sweden and they were unable to create one for him. What does this mean? It is a fact that, in Sweden, the total number of full Professorships in mathematics is a fixed constant. In fact today it is twenty. But in 1960 it was nineteen. And all the nineteen positions were filled. And nobody was about to step down or abandon his professorship so that Hörmander might have a job. Recall that Isaac Newton's teacher Isaac Barrow did indeed give up his Lucasian Chair Professorship so that the brilliant young Newton could assume it.[4] But nothing like this was going to happen in socialistic Sweden in 1962. So Hörmander, never a wilting flower, quit Sweden and moved to the Institute for Advanced Study in Princeton, New Jersey. This is arguably the most prestigious mathematics job in the world, with a spectacular salary and many lovely perks. So Hörmander flourished in his new venue. After several years the Swedes became conscious of their loss. One day an insightful politician stood up in the national Parliament and said, "I think it's a tragedy that the most brilliant Swedish mathematician in history cannot find a suitable job in our country. He has left and moved to the United States." So in fact the Swedish Parliament created one more mathematics Professorship—raised the sacred total from nineteen to twenty—just so that Hörmander could have a Professorship in Sweden. And, loyal Swede that he was, Hörmander dutifully moved back to Sweden. He has been at the University of Lund for nearly forty years.

But it should be stressed that Hörmander's job in Lund (he is now retired) was nothing special. He took a more than 50% cut in pay to return to Sweden, and his salary was about the same as that of any other senior Professor. And this is so because such things as Professors' salaries are centrally regulated, and a system of central regulation does not take into account exceptional individuals like Lars Hörmander.

The point here is that the American system is more competitive—the academic world is really a marketplace—and this perhaps motivates people to strive harder and to seek greater heights.[5] By contrast, a German academic

[4]Barrow had other motivations for this largesse: he wanted to assume a position at Court. Nonetheless, his resignation gave Newton the opportunity he needed.

[5]It must be noted, however, that until recently the Russian system was highly im-

once told me that he knew what his salary would be in five years or ten years or twenty years because the government had published a book laying it all out. If he got an outside offer, or moved to another university, then nothing would change. Everything was centrally regulated.

One of the chief tenets of American education is "local control of schools". This has its upsides and its downsides; certainly education in rural Mississippi is quite different from education in Boston. But the American system contributes to a more competitive atmosphere. An American university (especially a private one) would have no trouble whatever creating a special Professorship for a scholar like Lars Hörmander. And it would not be a cookie-cutter position just like any other professorship at the university. The Chancellor and the Provost could tailor the position to fit the achievements and the prestige of the individual in question. They could cook up any salary they wanted, and assign any perks to the position that they thought appropriate. The distinguished Professor could have a private secretary, he could have his own limousine, or whatever they thought would keep him happy and loyal to the institution.

It is also the case that the American academic world is much less of a caste system than the European or Asian systems. Things are much less structured, and there is much more mobility. Certainly it is very common for American Assistant Professors and Associate Professors and full Professors to be friends. American Professors will joke around with the secretaries and have lunch with the graduate students. One sees much less of this type of behavior at foreign universities. Perhaps the free and open nature of our system contributes to its success.

But it must be said that one of the keys to the success of the American academic system is money, pure and simple. Higher education is expensive, and American universities have access to many more resources than do European universities. There are many government and private funding agencies, and also there is a great tradition at American universities of alumni giving (not so in England, for example). Harvard University has an endowment that exceeds 20 billion dollars, and much of that money comes from alumni gifts. It is quite common for a wealthy donor to approach an American university with the offer of $10 million to endow five chair professorships.[6] This is a

pressive, and produced a myriad of brilliant mathematicians and great theorems. And certainly, under the communist regime, the Russian system was *not* a marketplace. But Russian culture is special in many ways, and there were other forces at play.

[6]For example this type of situation happened at Penn State in 1986. Wealthy donor

special academic position, with a fine salary and many benefits, that is designed to attract and retain the best scholars. Nothing like this ever happens in Europe.

Many brilliant scholars leave Europe to accept positions at American universities just because the salaries are so much higher (and, concomitantly, the working conditions so much more conducive to productivity). At the same time, they often return to their homeland when they retire.

4.3 L. E. J. Brouwer and Proof by Contradiction

L. E. J. Brouwer (1881–1966) was a bright young Dutch mathematician whose chief interest was in topology. Now topology was quite a new subject in those days (the early twentieth century). Affectionately dubbed "rubber sheet geometry", the subject concerns itself with geometric properties of surfaces and spaces that are preserved under continuous deformation (i.e., twisting and bending and stretching). In his studies of this burgeoning new subject, Brouwer came up with a daring new result, and he found a way to prove it.

Known as the "Brouwer Fixed-Point Theorem", the result can be described as follows. Consider the closed unit disc $\overline{D}$ in the plane, as depicted in Figure 4.1. This is a round, circular disc—including the boundary circle as shown in the picture. Now imagine a function $\varphi : \overline{D} \to \overline{D}$ that maps this disc continuously to itself, as shown in Figure 4.2. Brouwer's result is that the mapping φ must have a fixed point. That is to say, there is a point $P \in \overline{D}$ such that $\varphi(P) = P$. See Figure 4.3.

This is a technical mathematical result, and its rigorous proof uses profound ideas such as homotopy. But, serendipitously, it lends itself rather naturally to some nice heuristic explanations. Here is one popular interpretation. Imagine that you are eating a bowl of soup—Figure 4.4. You sprinkle grated cheese uniformly over the surface of the soup (see Figure 4.5). And then you stir up the soup. We assume that you stir the soup in a civilized manner so that all the cheese remains on the surface of the soup (refer to Fig-

Robert Eberly gave money to each of the science departments—biology, chemistry, astronomy, and physics—to establish an endowed chair professorship. He decilined to give money for mathematics because he had never had a math class at Penn State that he had liked. But the university was able to use its own funds to repair that situation.

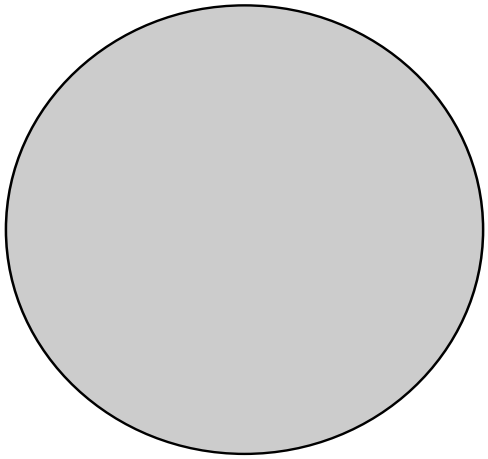


Figure 4.1

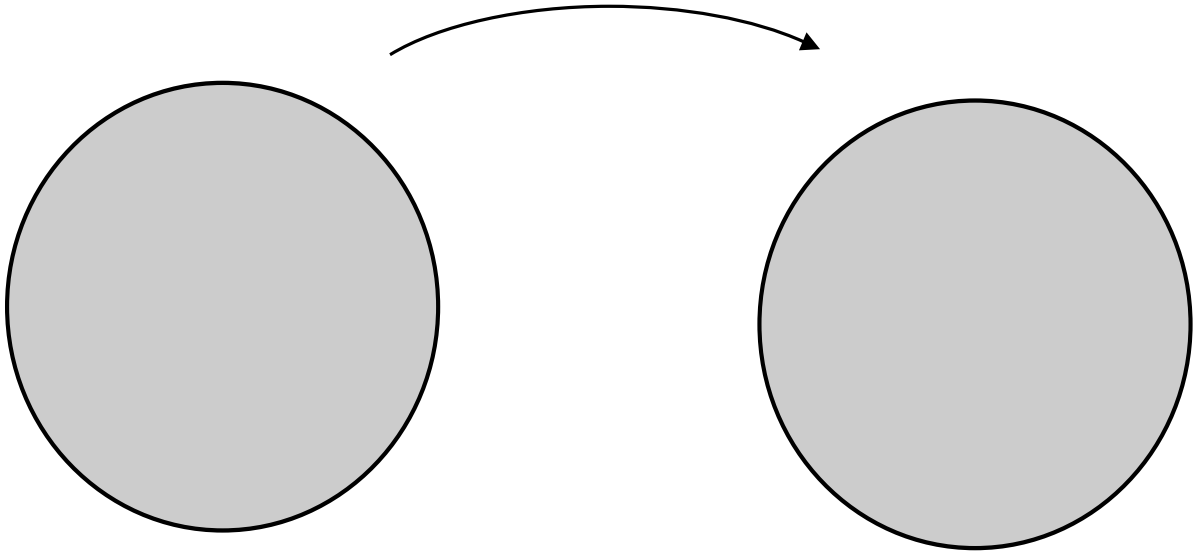


Figure 4.2

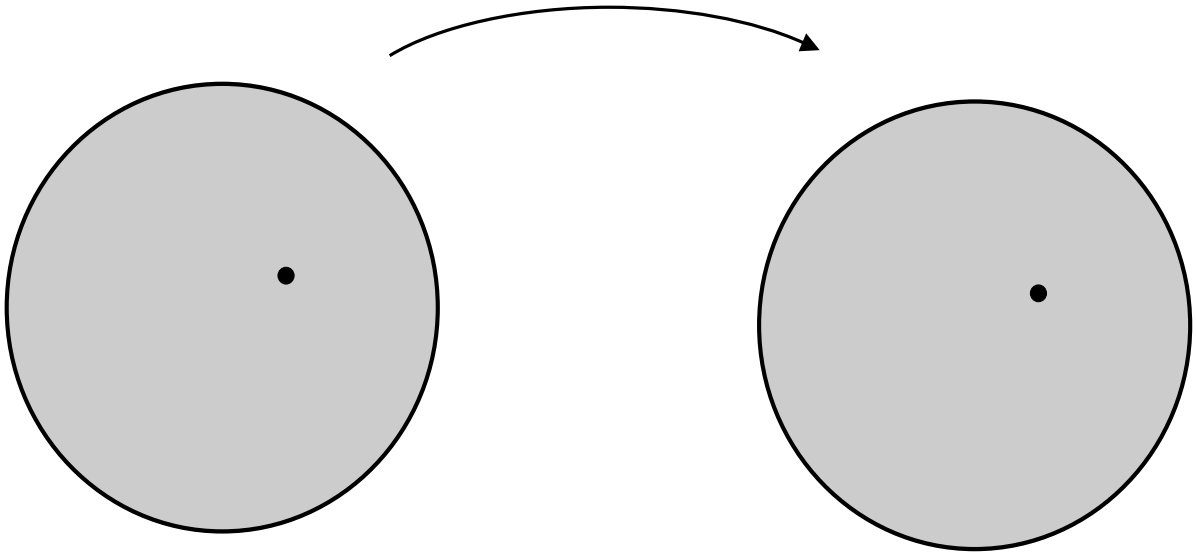


Figure 4.3

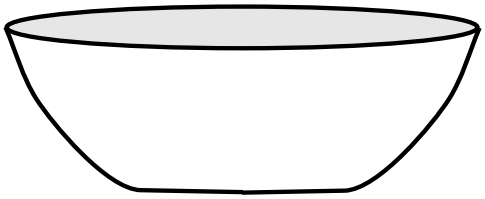


Figure 4.4

ure 4.6). Then some grain of cheese remains in its original position (Figure 4.7).

The soup analogy gives a visceral way to think about the Brouwer fixed-point theorem. Both the statement and the proof of this theorem—in the year 1909—were quite dramatic. In fact it is now known that the Brouwer fixed-point theorem is true in every dimension (Brouwer himself proved it only in dimension 2). We shall provide some discursive discussion of the result below.

The Brouwer fixed-point theorem is one of the most fascinating and important theorems of twentieth-century mathematics. Proving this theorem established Brouwer as one of the pre-eminent topologists of his day. But he refused to lecture on the subject, and in fact he ultimately rejected this (his

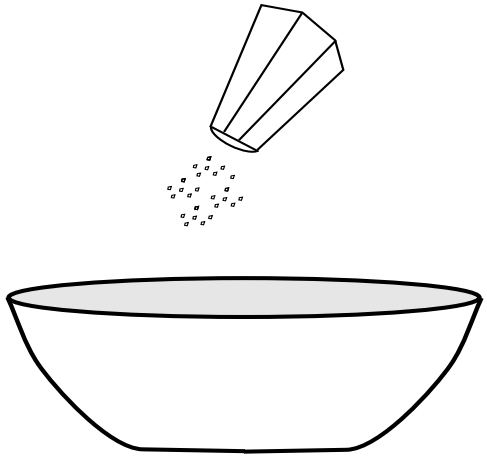


Figure 4.5

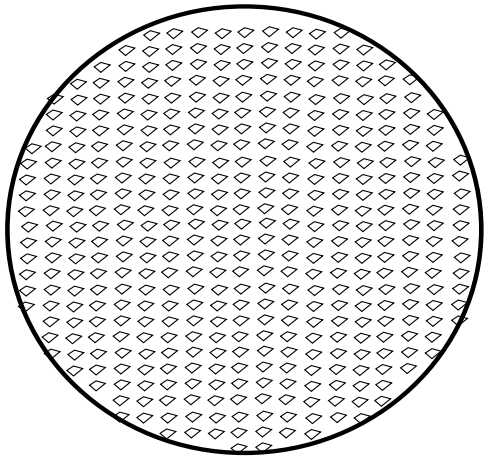


Figure 4.6

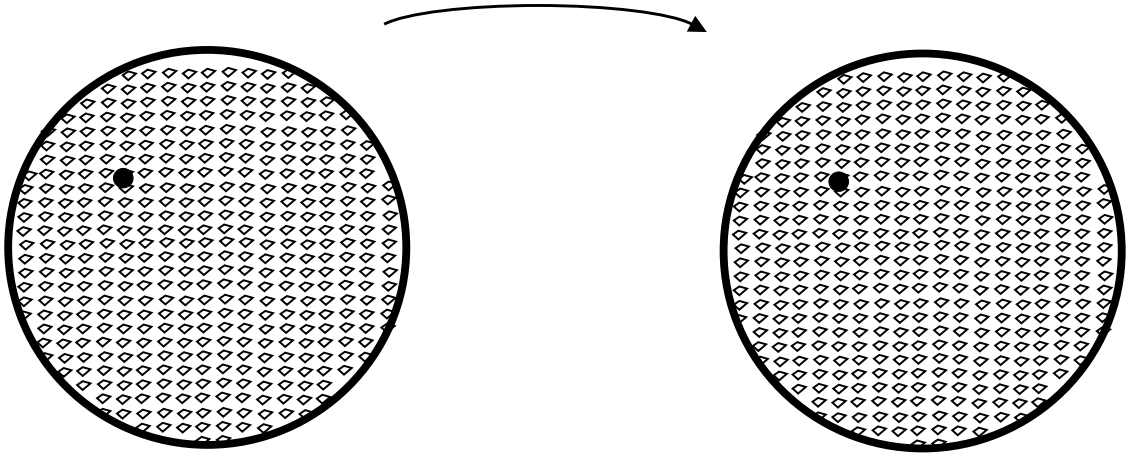


Figure 4.7

own!) work. The reason for this strange behavior is that L. E. J. Brouwer had become a convert to *constructivism* or *intuitionism*. He rejected the Aristotelian dialectic (that a statement is either true or false and there is no alternative), and therefore rejected the concept of "proof by contradiction". Brouwer had come to believe that the only valid proofs—at least when one is proving *existence* of some mathematical object (like a fixed point!) and when infinite sets are involved—are those in which we *construct* the asserted objects being discussed.[7] Brouwer's school of thought became known as "intuitionism", and it has made a definite mark on twentieth century mathematics.

As we shall see below, the Brouwer fixed-point theorem asserts the existence of a "fixed point" for a continuous mapping. We demonstrate that the fixed point exists by assuming that it does not exist and deriving thereby a contradiction. This is Brouwer's original method of proof, but the methodology flies in the face of the intuitionism that he later adopted.

Let us begin by discussing the general idea of the Brouwer fixed point theorem. We proceed by considering a "toy" version of the question in one dimension. Consider a continuous function f from the interval $[0, 1]$ to $[0, 1]$. Figure 4.8 exhibits the graph of such a function.

Note here that the word "continuous" refers to a function that has no breaks in its graph. Some like to say that the graph of a continuous function

[7] In fact, for the constructivists, the phrase "there exists" must take on a rigorous new meaning that exceeds the usual rules of formal logic.

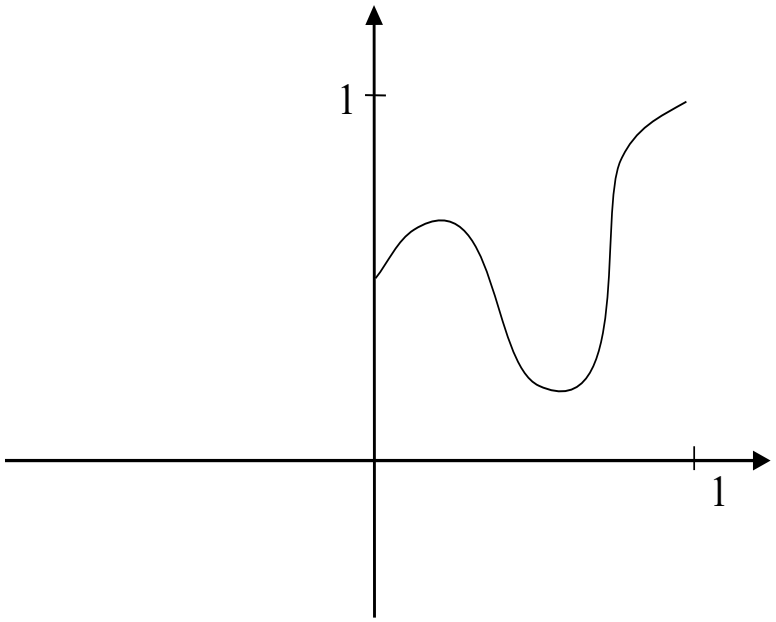


Figure 4.8

"can be drawn without lifting the pencil from the paper." Although there are more mathematically rigorous definitions of continuity, this one will suffice for our purposes. The question is whether there is a point $p \in [0,1]$ such that $f(p) = p$. Such a point p is called a *fixed point* for the function f. Figure 4.9 shows how complicated a continuous function from $[0,1]$ to $[0,1]$ can be. In each instance it is not completely obvious whether there is a fixed point or not. But in fact Figure 4.10 exhibits the fixed point in each case.

Of course it is one thing to draw a few pictures and quite another to establish once and for all that, no matter what the choice of the continuous function $f : [0,1] \to [0,1]$, there is a fixed point p. What is required now is a *mathematical proof*. Now here is a formal enunciation and proof of our result:

Theorem 4.3.1 *Let $f : [0,1] \to [0,1]$ be a continuous function. Then there is a point $p \in [0,1]$ such that $f(p) = p$.*

Proof: We may as well suppose that $f(0) \neq 0$ (otherwise 0 is our fixed point and we are done). Thus $f(0) > 0$. We also may as well suppose that $f(1) \neq 1$ (otherwise 1 is our fixed point and we are done). Thus $f(1) < 1$.

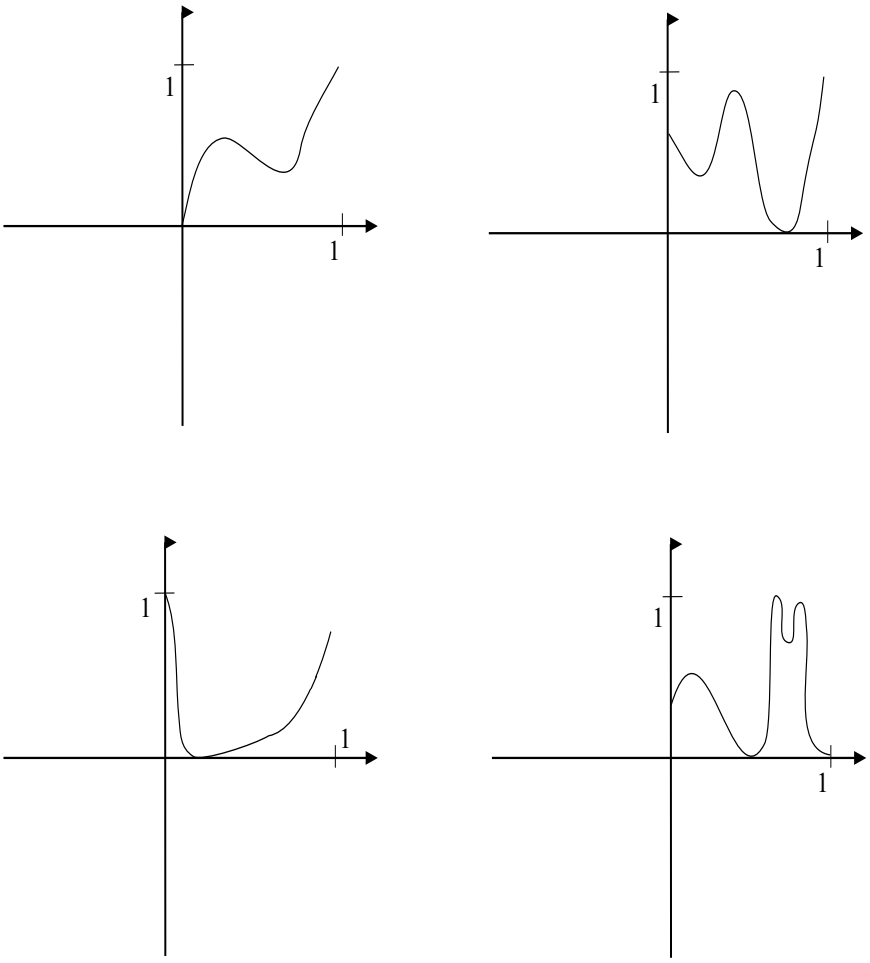


Figure 4.9

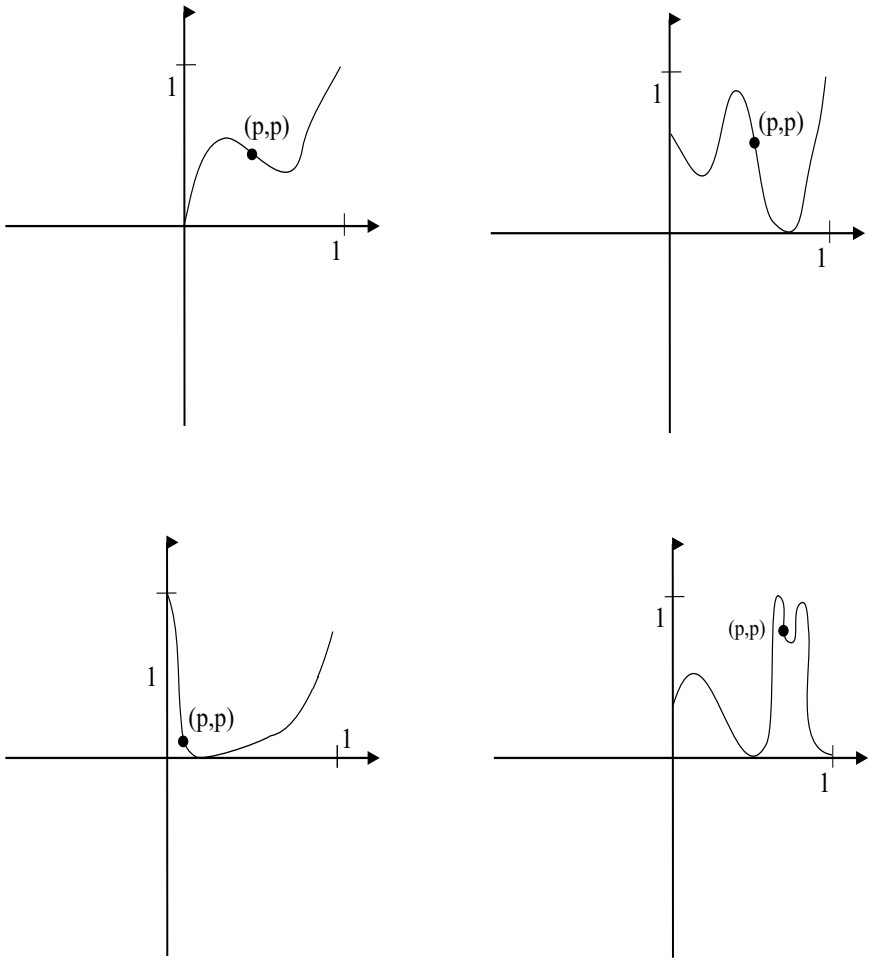


Figure 4.10

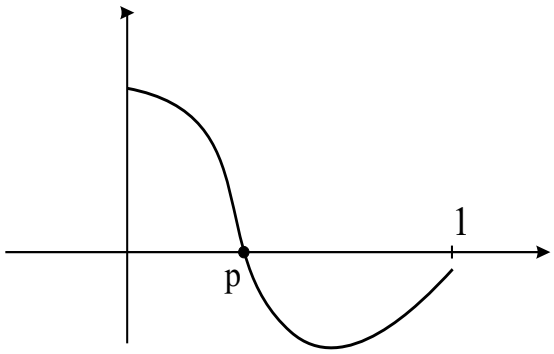


Figure 4.11

Consider the auxiliary function $g(x) = f(x) - x$. By the observations in the last paragraph, $g(0) > 0$ and $g(1) < 0$. Look at Figure 4.11. We see that a continuous function with these properties must have a point p in between 0 and 1 such that $g(p) = 0$. But this just says that $f(p) = p$. □

Now we turn to the higher-dimensional, particularly the 2-dimensional, version of the Brouwer fixed-point theorem. This is the formulation that caused such interest and excitement when Brouwer first proved the result over ninety years ago. Before we proceed, we must establish an auxiliary topological fact. And it is for this purpose that we are going to use Poincaré's homotopy theory.

Lemma 4.3.2 *Let U, V be geometric figures and $g : U \to V$ be a continuous function. If γ is a closed curve in U that can be continuously deformed to a point, then $g(\gamma)$ is a subset of V that is also a closed curve that can be continuously deformed to a point.*

This statement makes good sense. Obviously a continuous function will not take a closed curve and open it up into a *non*-closed curve; that is antithetical to the notion of continuity. And if we imagine a flow of curves, beginning with γ, that merge to a point in U, then of course their images under f will be a flow of curves in V that merge to a point in V.

Definition 4.3.3 Let $\overline{D}$ be the closed unit disc (i.e., the unit disc together with its boundary) as shown in Figure 4.12. Let C denote the boundary

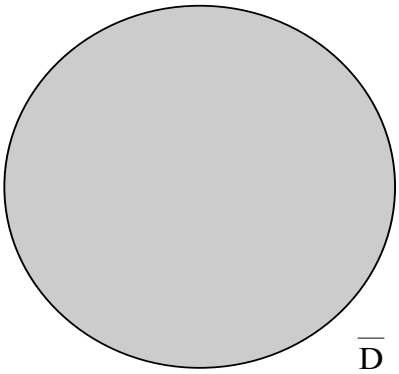


Figure 4.12

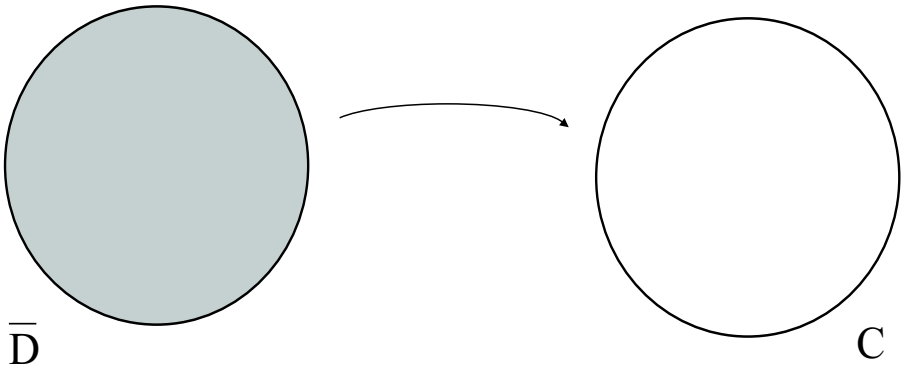


Figure 4.13

circle of $\overline{D}$. A continuous function $h : \overline{D} \to C$ that fixes each point of C is called a *retraction* of $\overline{D}$ onto C. See Figure 4.13.

Proposition 4.3.4 *There does not exist any retraction from $\overline{D}$ onto C.*

For the reasoning, consider Figure 4.14. Seeking a contradiction, we assume that there *is* a retraction $r : \overline{D} \to C$. The function u is just the inclusion map from C into $\overline{D}$. Now let γ be the curve in C that just wraps once around the circle—Figure 4.15. Then the composition $r \circ u$ (which means the operation u followed by the operation r) obviously just takes γ onto itself. On the other hand, u must take γ to a curve $u(\gamma) \subseteq \overline{D}$ that is shrinkable to a point—since all curves in $\overline{D}$ shrink to a point. And, by the lemma, r in turn must take $u(\gamma)$ to another curve that shrinks to a point.

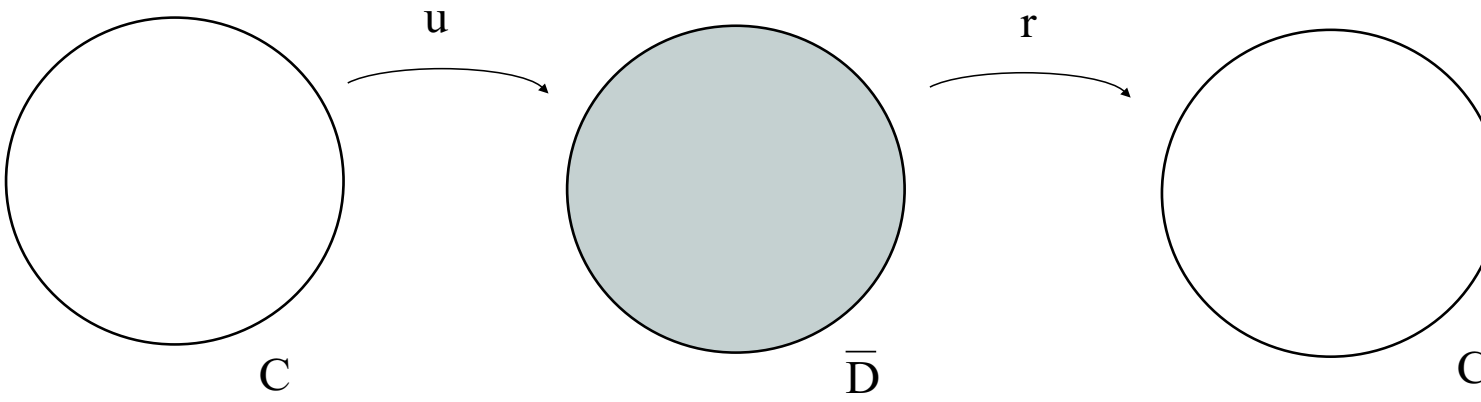


Figure 4.14

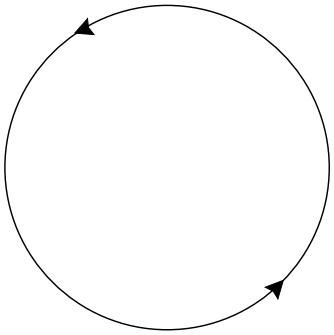


Figure 4.15

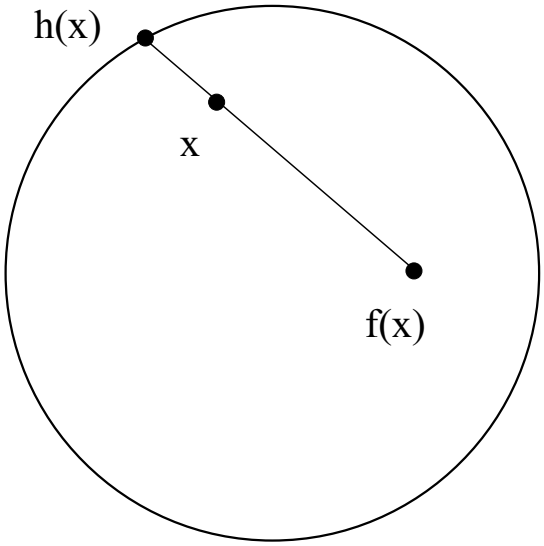


Figure 4.16

But now we have a problem: On the one hand, $r \circ u$ takes γ onto itself, and thus $r \circ u(\gamma)$ *cannot* be shrunk to a point. On the other hand, we just argued that $r \circ u(\gamma)$ *can be shrunk to a point*. It is impossible to have both statements be true. That is our contradiction. So the retraction cannot exist.

And now here is Brouwer's famous theorem:

Theorem 4.3.5 *Let $\overline{D}$ be the closed unit disc. Let $f : \overline{D} \to \overline{D}$ be a continuous function. Then there is a point $P \in \overline{D}$ such that $f(P) = P$.*

At the risk of offending Brouwer himself, we provide a proof by contradiction. Suppose that there is such a map f that does *not* possess a fixed point. Then, for each point $x \in \overline{D}$, $f(x) \neq x$. But then we can use f to construct a retraction of $\overline{D}$ onto C as follows. Examine Figure 4.16. You can see that the segment that begins at $f(x)$, passes through x, and ends at a point $h(x)$ in C gives us a mapping

$$x \longmapsto h(x) \in C = \partial D\,.$$

Notice that the domain of this mapping—the collection of points x that we plug into the mapping—is just $\overline{D}$, the disc together with its boundary. And the image of the mapping—the set of values—is the circle C, or the

boundary of the disc. This mapping is evidently continuous, as a small perturbation in x will result in a small perturbation in $f(x)$ and hence a small perturbation in $h(x)$. Furthermore, each element of C is mapped, under h, to itself. So in fact *h is a retraction of $\overline{D}$ to C*. Note that the reason that we can construct this retraction is that $f(x) \neq x$; it is because of that inequality that we know how to draw the segment that defines $h(x)$. But we know, by the proposition, that this is impossible. Thus it cannot be that $f(x) \neq x$ for all x. As a result, some point P is fixed by f. And that is the end of our proof. We have established Brouwer's fixed point theorem using Poincaré's homotopy theory. □

4.4 The Generalized Ham-Sandwich Theorem

4.4.1 Classical Ham Sandwiches

In this section we are going to discuss a far-reaching generalization of the Brouwer fixed-point theorem. Our treatment will be almost entirely intuitive, as it must be. But it serves to show that mathematical ideas are not stagnant. Any good insight gives rise to further investigation and further discoveries. The "generalized ham-sandwich theorem" is one of these.

First, let us define a *classical ham sandwich*. Such a sandwich consists of two square pieces of bread and a square slice of ham (assuming that we are using packaged ham) and a square slice of cheese (assuming that we are using packaged cheese). See Figure 4.17.

Now it is easy to see that, with a single slice of the knife, it is possible to cut the sandwich in such a way that

- the aggregate of bread (both slices) is sliced in half,
- the cheese is sliced in half,
- the ham is sliced in half.

Figure 4.18 illustrates one such cut. Figure 4.19 illustrates another.

In fact there are infinitely many ways to perform a planar cut of the classical ham sandwich that will bisect each of the bread, the cheese, and the ham.

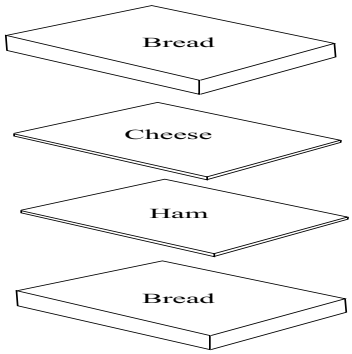


Figure 4.17. A classical ham sandwich.

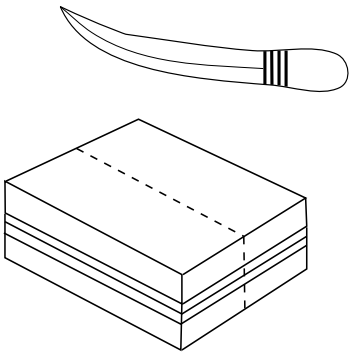


Figure 4.18

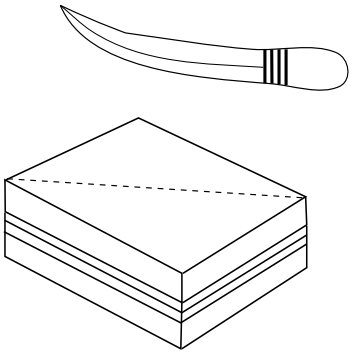


Figure 4.19

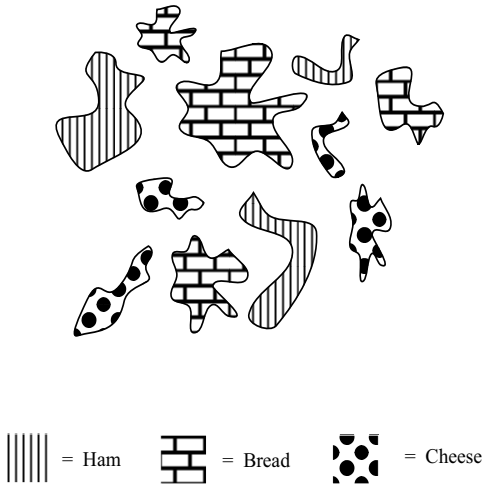


Figure 4.20. A generalized ham sandwich.

In the next subsection we shall define a "generalized ham sandwich" and discuss an analogous but considerably more surprising result.

4.4.2 Generalized Ham Sandwiches

A generalized ham sandwich consists of *some ham*, *some cheese*, and *some bread*. But the ham could be in several pieces, and in quite arbitrary shapes. Similarly for the cheese and the bread. Figure 4.20 illustrates a generalized ham sandwich.

Please remember that these ham sandwiches live in 3-dimensional space. The generalized ham sandwich shown in Figure 4.20 is a 3-dimensional ham sandwich. Each of the ham, the cheese, and the bread is a solid, 3-dimensional object.

Now we have the following astonishing theorem:

Theorem 4.4.1 *Let $\mathcal{S}$ be a generalized ham sandwich in 3-dimensional space. Then there is a single planar knife cut that*

- *bisects the bread,*
- *bisects the cheese,*
- *bisects the ham.*

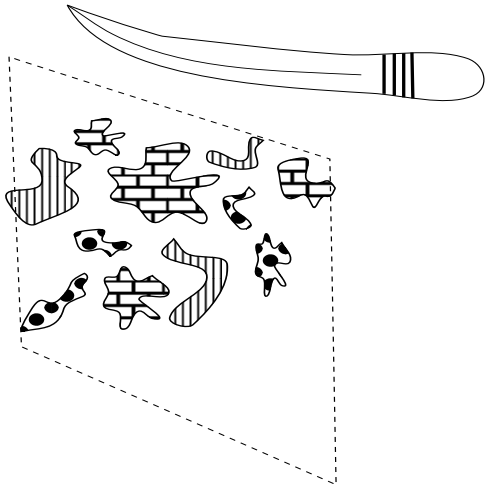


Figure 4.21

See Figure 4.21. The proof, which is too complicated to present here, is a generalization of the Intermediate Value Property that we used to prove the fixed-point theorem in dimension 1.

In fact it is worth pondering this matter a bit further. Let us consider the generalized ham-sandwich theorem in dimension 2. In this situation we cannot allow the generalized ham sandwich to have three ingredients. In fact, in dimension 2, the generalized ham sandwich has only bread and ham. No cheese. Then the same result is true: a single linear cut will bisect the ham and bisect the bread. Examine Figure 4.22 and convince yourself that, with ham and cheese and bread configured as shown in dimension 2 there is no linear cut that will bisect all three quantities. But *any two* of the ham, bread, and cheese may be bisected by a single linear cut.

In dimension 4, we can add a fourth ingredient to the generalized ham sandwich—such as turkey. And then there is a single hyper-planar slice that will bisect each of the four quantities: turkey, ham, cheese, and bread. This is all pretty abstract, and we cannot discuss the details here (see [GAR] for a nice discussion).

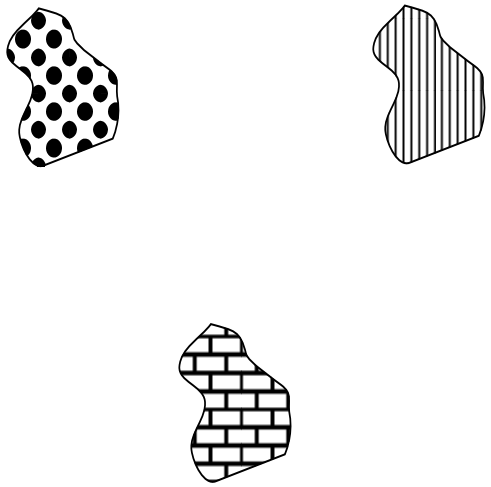


Figure 4.22

4.5 Much Ado About Proofs by Contradiction

As we have seen, L. E. J. Brouwer used a "proof by contradiction" to establish his celebrated fixed-point theorem. But he later repudiated this methodology, claiming that all mathematical existence theorems should be established constructively. Today in fact there *are* constructive methods for proving versions of the Brouwer fixed-point theorem (see [BIS]). We shall discuss one such method—due to Isaac Newton—below.[8]

At first Brouwer was a lone wolf, preaching his doctrine of constructivism all by himself. But, over time, he gained adherents. Certainly theoretical computer scientists have an interest in constructivism, as a computer is nothing but a mechanical device for carrying our mathematical operations in a constructive manner. Proofs by contradiction have their place in the theory of computers, yet the computer itself is a constructivist tool.

The world was somewhat taken by surprise when, in 1968, the distinguished mathematician Errett Bishop (1928–1983) came out in favor of constructivism. He was another researcher who had made his reputation by giving a number of dazzling proofs using the method of proof by contradic-

[8]For completeness, we must note that Newton's method predates Brouwer by a couple of hundred years!

tion. But then he rejected the methodology. We shall say more about Bishop later on.

It is perhaps worth examining the 1-dimensional version of the Brouwer fixed-point theorem to see whether we can think about it in a constructivist manner. Refer to Section 4.2 to review the key ideas of the proof. The main step is that we had a function $g(x) = f(x) - x$, and we had to find a place where the function vanishes. We knew that $g(0) > 0$ and $g(1) < 0$, so we could invoke the Intermediate Value Property of continuous functions to conclude that there is a point ξ between 0 and 1 such that $g(\xi) = 0$. It follows that $f(\xi) - \xi = 0$ or $f(\xi) = \xi$.

All well and good. This is a plausible and compelling proof. But it is certainly not constructive. The existence of ξ is just that: an abstract existence proof. In general we cannot say what ξ is, nor how to find it. In case g is a *smooth* function—i.e., no corners on its graph—then there is a technique of Isaac Newton that often gives a constructive method for finding ξ.[9]

We describe it briefly now.

So imagine our smooth function g that is positive at 0 and negative at 1—see Figure 4.23. In order to invoke Newton's method, we must start with a first guess ξ_1 for ξ. See Figure 4.24. Now the idea is to take the tangent line[10] to the graph at the point corresponding to ξ_1—see Figure 4.25—and see where *it* intersects the x-axis. That will be our second approximation ξ_2 to the number ξ that we seek. Figure 4.26 shows how the second guess is a considerable improvement over the first guess. In fact calculations show that one usually doubles the number of decimal places of accuracy with each iteration of Newton's method.

Of course one may apply this reasoning once again—taking a tangent line to the graph at ξ_2 and seeing where it intersects the x-axis. This will result in another notable improvement (called ξ_3) to our guess. Iterating Newton's method, one usually obtains a rapidly converging sequence of approximations

[9]We should stress that there are certainly examples in which Newton's method does *not* give the desired result—i.e., find the fixed point. Imposing smoothness does not address the fundamental epistemological issue raised by L. E. J. Brouwer—one still cannot in general find a fixed point constructively. But, as a practical matter, Newton's method usually works, and it gives a constructive procedure for producing a sequence that will converge to a fixed point. Newton's method is important, and is the basis for a big area of mathematics these days that is known as *numerical analysis*.

[10]This tangent line can be explicitly calculated using methods of calculus.

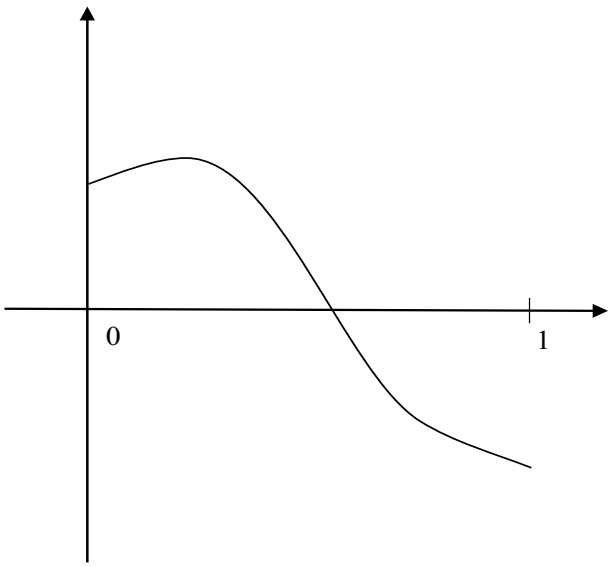


Figure 4.23

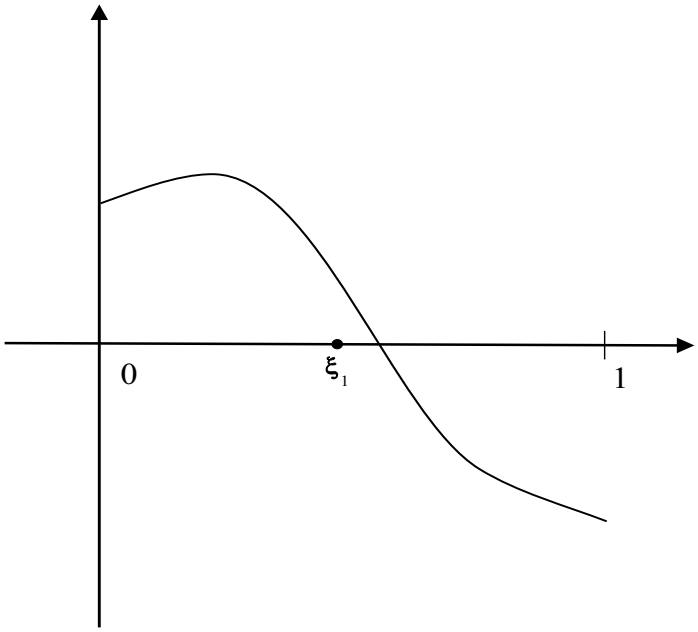


Figure 4.24

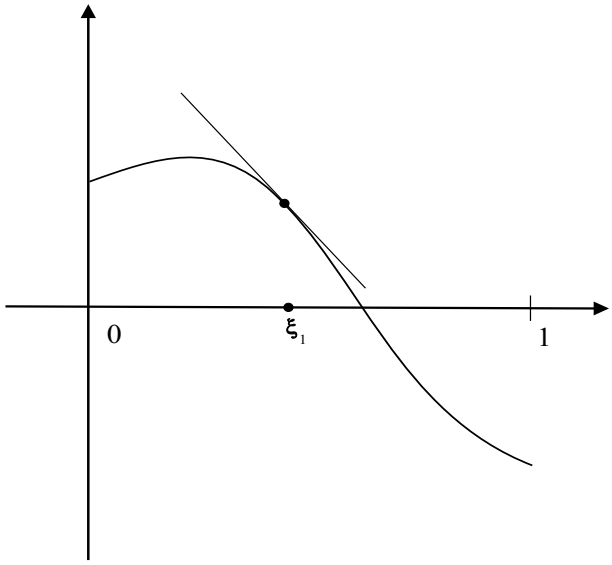


Figure 4.25

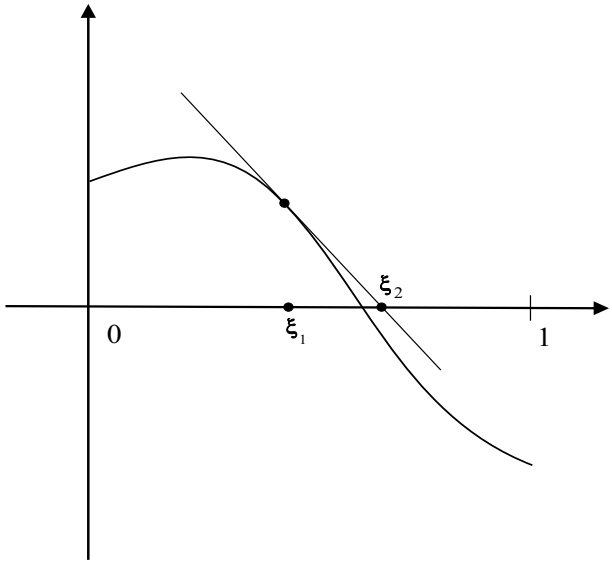


Figure 4.26

to the true root of the function.

As a simple illustration of Newton's method, suppose that we wish to calculate $\sqrt{2}$—the famous number that Pythagoras proved to be irrational. Of course this number is a root of the polynomial equation $p(x) = x^2 - 2 = 0$. Let us take our initial guess to be $\xi_1 = 1.5$. Of course $1.5^2 = 2.25$, so this guess is a very rough approximation to the true value that we seek. It turns out that if one carries out the calculations needed (as we described above) for Newton's method—and these are quite straightforward if one knows a little calculus—then

$$\xi_{n+1} = \frac{\xi_n}{2} + \frac{1}{\xi_n}\,.$$

Applying this simple formula with $n = 1$, we find that

$$\xi_2 = \frac{\xi_1}{2} + \frac{1}{\xi_1}\,.$$

Now plugging in the value of our initial guess $\xi_1 = 1.5$ we find that

$$\xi_2 = 1.416666\ldots\,. \tag{$*$}$$

Since the true value of $\sqrt{2}$, accurate to twelve decimal places, is $\sqrt{2} = 1.414213562373$, we find that just one iteration of Newton's method gives us $\sqrt{2}$ to two decimal places of accuracy (well, only one decimal place if we round it off).

Now let us apply Newton's method again. So

$$\xi_3 = \frac{\xi_2}{2} + \frac{1}{\xi_2}\,.$$

Plugging in the value of ξ_2 from $(*)$ we find that

$$\xi_3 = 1.414215\ldots\,. \tag{$\star$}$$

Now we have $\sqrt{2}$ to four decimal places of accuracy.

Let us apply Newton's method one more time. We know that

$$\xi_4 = \frac{\xi_3}{2} + \frac{1}{\xi_3}\,.$$

Plugging in the value of ξ_3 from $(\star)$ now yields

$$\xi_4 = 1.41421356237\ldots\,.$$

Thus, with three simple iterations of Newton's method, we have achieved eleven decimal places of accuracy. Notice that the degree of accuracy at least doubled with each application of the technique!

Newton's method is the most fundamental technique in that branch of mathematics known as *numerical analysis*. The idea of numerical analysis is to use approximation methods—and computer calculation—to obtain answers to otherwise intractable problems.[11] The use of numerical analysis methods involves another paradigm shift, for we are no longer finding precise solutions. Instead, we pre-specify a desired degree of accuracy (in terms of the number of decimal places, for example) and then we find an approximate solution that has that number of decimal places of accuracy. As we have mentioned, Newton's method typically doubles the number of decimal places of accuracy with each iteration of the method. So it is a very effective device when used in conjunction with a computer.

4.6 Errett Bishop and Constructive Analysis

Errett Bishop was one of the great geniuses of mathematical analysis in the 1950s and 1960s. He made his reputation by devising devilishly clever proofs about the structure of spaces of functions. Many of his proofs were indirect proofs—that is to say, proofs by contradiction.

Bishop underwent some personal changes in the mid- to late-1960s. He was a Professor of Mathematics at U. C. Berkeley and he was considerably troubled by all the political unrest on campus. After a time, he felt that he could no longer work in that atmosphere. So he arranged to transfer to U. C. San Diego. At roughly the same time, Bishop became convinced that proofs by contradiction were fraught with peril. He wrote a remarkable and rather poignant book [BIS] which touts the philosophy of constructivism—similar in spirit to L. E. J. Brouwer's ideas from fifty years before. Unlike Brouwer, Bishop really put his money where his mouth was. In the pages of his book, Bishop is able to actually develop most of the key ideas of mathematical analysis without resort to proofs by contradiction. Thus he created a new field of mathematics called "constructive analysis".

A quotation from Bishop's Preface to his book gives an indication of how that author himself viewed what he was doing:

[11] Of course we performed the calculations for $\sqrt{2}$ by hand, and this just took a few minutes. But a computer could do the calculations in a fraction of a second.

> Most mathematicians would find it hard to believe that there could be any serious controversy about the foundations of mathematics, any controversy whose outcome could significantly affect their own mathematical activity.

He goes on to say that

> It is no exaggeration to say that a straightforward realistic approach to mathematics has yet to be tried. It is time to make the attempt.

In a perhaps more puckish mood, Bishop elaborates:

> Mathematics belongs to man, not to God. We are not interested in properties of the positive integers that have no descriptive meaning for finite man. When a man proves a positive integer to exist, he should show how to find it. If God has mathematics of His own that needs to be done, let Him do it Himself.

But our favorite Errett Bishop quotation, and the one that bears most closely on the theme of this book, is

> A proof is any completely convincing argument.

Bishop's arguments in *Methods of Constructive Analysis* [BIS] were, as was characteristic of Bishop, devilishly clever. The book had a definite impact, and certainly caused people to reconsider the methodology of modern analysis. Bishop's acolyte and collaborator D. Bridges produced the revised and expanded version [BIB] of his work (published after Bishop's death), and there the ideas of constructivism are carried even further.

4.7 Nicolas Bourbaki

The turn of the twentieth century saw the dawn of the modern age of logic. Mathematicians realized that the intellectual structure of mathematics was rather chaotic. There were no universally accepted standards of rigor. Different people wrote up the proofs of their theorems in different ways. Some rather prominent mathematicians rarely proved anything rigorously.

From our current perspective of over one hundred years distance, it is not clear how much of the mathematical climate around 1900 is due to a lack of coherence in the subject and how much is a reflection of a large number of geniuses working to the limits of their capacity, and in relative intellectual isolation. In today's mathematical climate, any worker is instantly answerable (because of the Internet, and because of the worldwide, closely knit scholarly community) to mathematicians in Australia, Japan, France, Istanbul, and other points around the globe. It is virtually impossible to work in isolation. For the most part, a mathematician who does his/her work less than rigorously, who does not follow the well-established rules of the game, is quickly sidelined.[12] That was not the state of the art at the turn of the twentieth century.

One hundred or so years ago there was also no universally accepted language of mathematics. Different technical terms meant different things to different people. The foundations of geometry in France looked different from the foundations of geometry in England, and those in turn looked different (at least in emphasis) from the foundations of geometry in Germany. America was a fifth wheel in the mathematics game. From the point of view of the mandarins at the great world centers in Paris and Berlin and Göttingen, there had never been a great American mathematician. Which is to say that nobody in the United States had ever proved a great theorem—one that was recognized by the authorities in the great intellectual centers of Europe. Recognition of American mathematics would come somewhat later, through the work of G. D. Birkhoff and Norbert Wiener and others.

Certainly David Hilbert of Göttingen was considered to be one of the premiere intellectual leaders of European mathematics. Just as an indication of his pre-eminence, he was asked to give the keynote address at the second International Congress of Mathematicians that was held in Paris in 1900. What Hilbert did at that meeting was earthshaking—from a mathematical point of view. He formulated twenty-three problems that he thought should serve as beacons in the mathematical work of the twentieth century. On the advice of Hurwitz and Minkowski, Hilbert abbreviated his remarks and only presented ten of these problems in his lecture. But soon thereafter a more complete version of Hilbert's ideas was published in several countries.

[12]Of course there are exceptions. Fractal geometers have forged their own path, and developed a version of mathematics that is largely phenomenological. Chaoticists, numerical analysts, low dimensional topologists all do mathematics a bit differently.

For example, in 1902 the *Bulletin of the American Mathematical Society* published an authorized translation by Mary Winston Newson[13] (1869–1959). This version described all twenty-three of the unsolved mathematics problems that Hilbert considered to be of the first rank, and for which it was of the greatest importance to find a solution. In fact Hilbert's *eighteenth* problem contains the Kepler sphere-packing problem, which we discuss in detail in Chapter 8. The first person to solve a Hilbert problem was Max Dehn, in the year 1900. It was problem number 3.[14] Dehn was a Professor at the University of Frankfurt, and also spent time at the Illinois Institute of Technology.

It is often said that David Hilbert was the last mathematician to be conversant with all parts of mathematics. Certainly he wrote fundamental texts on all the basic parts of the mathematics of the day. So it was with ease that Hilbert could, in his lecture to the International Congress in 1900, survey the entire subject of mathematics and pick out certain areas that demanded attention and offered particularly tantalizing problems. What are now known as the "Hilbert problems" cover algebra, geometry, differential equations, combinatorics, real analysis, logic, complex analysis, and many other parts of mathematics as well. Hilbert's message was that these twenty-three were the problems that mathematicians of the twentieth century should concentrate their efforts on solving.

Of course Hilbert's name carried considerable clout, and the mathematicians in attendance paid careful attention to the great savant's admonitions. They took the problems home with them and in turn disseminated them to their peers and colleagues. We have noted that Hilbert's remarks were written up and published, and thereby found their way to universities all over the world. It rapidly became a matter of great interest to solve a Hilbert problem, and considerable praise and encomia were showered on anyone who did so. Today most of the Hilbert problems are solved, but there are a few particularly thorny ones that remain. The references [GRA] and [YAN] give a detailed historical accounting of the colorful history of the Hilbert problems.

One of Hilbert's overriding passions was logic, and he wrote an important treatise in the subject [HIA]. Since Hilbert had a universal and comprehensive knowledge of mathematics, he thought carefully about how the different parts

[13]Newson was the first American woman to earn the Ph.D. degree at the university in Göttingen.

[14]The book [GRA] recounts that in fact a significant special case of this third Hilbert problem was solved even before Hilbert posed it!

of the subject fit together. And he worried about the axiomatization of the subject. Hilbert believed fervently that there ought to be a universal (and rather small) set of axioms for mathematics, and that all mathematical theorems should be derivable from those axioms.[15] But Hilbert was also fully cognizant of the rather uneven history of mathematics. He knew all too well that much of the literature was riddled with errors and inaccuracies and inconsistencies.

Hilbert's geometry book [HIL], which may well be his most important work (though others may argue for his number theory, his functional analysis, or his invariant theory), corrected many errors in Euclid's *Elements* and refined Euclid's axioms to the form in which we know them today. And his logic book lays out his program for the twentieth century to formalize mathematics, to uniformize its language, and to put the subject on a solid logical footing. To repeat, Hilbert's hopes were rather dashed by the bold and ingenious work of Gödel that was to come thirty years later. Certainly Russell's Paradox (see Section 0.7) was already making people mighty nervous. But the fact remains that Hilbert was an extremely prominent and influential scholar. People attended carefully to his teachings, and they adopted his program enthusiastically. Thus David Hilbert had an enormous influence over the directions that mathematics was taking at the beginning of the twentieth century.

There had long been a friendly rivalry between French mathematics and German mathematics. Although united by a common subject that everyone loved, these two ethnic groups practiced mathematics with different styles and different emphases and different priorities. The French certainly took David Hilbert's program for mathematical rigor very seriously, but it was in their nature then to endeavor to create their own home-grown program. This project was ultimately initiated and carried out by a remarkable figure in the history of modern mathematics. His name is Nicolas Bourbaki.

Jean Dieudonné, the great raconteur of twentieth-century French mathematics, tells of a custom at the École Normale Supérieure in France to subject first-year students in mathematics to a rather bizarre rite of initiation. A senior student at the university would be disguised as an important visitor from abroad; he would give an elaborate and rather pompous lecture in which

[15] We now know, thanks to work of Kurt Gödel (see Chapter 1), that in fact Hilbert's dream cannot be fulfilled. At least not in the literal sense. But it is safe to say that most working mathematicians take Hilbert's program seriously, and most of us approach our subject with this ideal in mind.

several "well-known" theorems were cited and proved. Each of the theorems would bear the name of a famous or sometimes not-so-famous French general, and each was wrong in some very subtle and clever way. The object of this farce was for the first-year students to endeavor to spot the error in each theorem, or perhaps not to spot the error but to provide some comic relief.

In the mid-1930's, a cabal of French mathematicians—ones who were trained at the notorious École Normale Supérieure—was formed with the purpose of writing definitive texts in the basic subject areas of mathematics. They ultimately decided to publish their books under the *nom de plume* Nicolas Bourbaki. In fact the inspiration for their name was an obscure French general named Charles Denis Sauter Bourbaki. This general, so it is told, was once offered the chance to be King of Greece but (for unknown reasons) he declined the honor. Later, after suffering an embarrassing retreat in the Franco-Prussian War, Bourbaki tried to shoot himself in the head—but he missed. Certainly Bourbaki's name had been used in the tomfoolery at the École Normale. Bourbaki was quite the buffoon. When the young mathematicians André Weil (1906–1998), Jean Delsarte (1903–1968), Jean Dieudonné (1906–1992), Lucien de Possel (1905–1974), Claude Chevalley (1909–1984), and Henri Cartan (1904–), decided to form a secret organization (named Nicolas Bourbaki) that was dedicated to writing definitive texts in the basic subject areas of mathematics, they decided to name themselves after someone completely ludicrous. For what they were doing was of the utmost importance for their subject. So it seemed to make sense to give their work a thoroughly ridiculous byline.

Thus, through a sequence of accidents and coincidences, it came about that some of the former students of the École Normale banded together and decided to assemble an ongoing work that would systematically describe and develop all the key ideas in modern mathematics. Today it can safely be said that "Nicolas Bourbaki" is one of the most famous and celebrated mathematical names of modern times. He is the author of an extensive, powerful, and influential series of mathematics books. But "Nicolas Bourbaki" is actually an allonym for an anonymous group of distinguished French mathematicians.

There are other, not necessarily contradictory, stories of how Bourbaki came about. André Weil tells that, in 1934, he and Henri Cartan were constantly squabbling about how best to teach Stokes's theorem in their respective courses at Strasbourg. He says, "One winter day toward the end of 1934, I thought of a brilliant way of putting an end to my friend's [Cartan's] persistent questioning. We had several friends who were responsible for

teaching the same topics in various universities. 'Why don't we get together and settle such matters once and for all, and you won't plague me with your questions any more?' Little did I know that at that moment Bourbaki was born."

Weil claims that the name "Bourbaki" has an even longer history. In the early 1920's, when they were students at the École Polytechnique, one of the older students donned a false beard and a strange accent and gave a much-advertised talk under the *nom de plume* of a fictitious Scandinavian. The talk was balderdash from start to finish, and concluded with a "Bourbaki's theorem" which left the audience speechless. One of the École's students claimed afterward to have understood every word.

Jean Dieudonné describes the philosophy for what is proper grist for the Bourbaki books as follows:

> . . . those which Bourbaki proposes to set forth are generally mathematical theories almost completely worn out already, at least in their foundations. This is only a question of foundations, not details. These theories have arrived at the point where they can be outlined in an entirely rational way. It is certain that group theory (and still more analytical number theory) is just a succession of contrivances, each one more extraordinary than the last, and thus extremely anti-Bourbaki. I repeat, this absolutely does not mean that it is to be looked down upon. On the contrary, a mathematician's work is shown in what he is capable of inventing, even new stratagems. You know the old story—the first time it is a stratagem, the third time a method. Well, I believe that greater merit comes to the man who invents the stratagem for the first time than to the man who realizes after three or four times that he can make a method from it.

As we have noted, the founding mathematicians in this new Bourbaki group came from the tradition of the École Normale Supérieure. This is one of the most elite universities in all of France, but it also has a long-standing tradition of practical joking. Weil himself tells of one particularly delightful story. In 1916, Paul Painlevé (1863–1933) was a young and extremely brilliant Professor at the Sorbonne. He was also an examiner for admission to the École Normale Supérieure. Each candidate for admission had to undergo a rigorous oral exam, and Painlevé was on the committee. So the candidates came early in the morning and stood around the hall outside the examination

room awaiting their turn. On one particular day, some of the more advanced students of the École began to chat with the novices. They told the youngsters about the fine tradition of practical joking at the school. They said that one of the standard hoaxes was that some student would impersonate an examiner, and then ridicule and humiliate the student being examined. The students should be forewarned.

Armed with this information, one of the students went in to take the exam. He sat down before the extremely youthful-looking Painlevé and blurted out, "You can't put this over on me!" Painlevé, bewildered, replied, "What do you mean? What are you talking about?" So the candidate smirked and said, "Oh, I know the whole story, I understand the joke perfectly, you are an impostor." The student sat back with his arms folded and waited for a reply. And Painlevé said, "I'm Professor Painlevé, I'm the examiner, . . . "

Things went from bad to worse. Painlevé insisted that he was a professor, but the student would not back down. Finally Painlevé had to go ask the Director of the École Normale to come in and vouch for him.

When André Weil used to tell this story, he would virtually collapse in hysterics.

In later years, Weil was at a meeting in India and told his friend Kosambi the story of the incidents that led to the formation of Bourbaki. Kosambi then used the name "Bourbaki" in a parody that he passed off as a contribution to the proceedings of some provincial academy. On the strength of this development, the still-nascent Bourbaki group determined absolutely that this would be its name. Weil's wife Eveline became Bourbaki's godmother and baptized him Nicolas.

Weil concocted a biography of Bourbaki and imputed him to be of "Poldavian descent". Nicolas Bourbaki of Poldavia submitted a paper to the *Comptes-Rendus* of the French Academy to establish his *bona fides*. Élie Cartan (1869–1951) and André Weil exercised considerable political skill in getting the man recognized and the paper accepted. It turns out that "Poldavia" was another concoction of the practical jokers at the École Normale. Puckish students frequently wrote letters and gave speeches on behalf of the beleaguered Poldavians. One such demagogue gave a speech that ended by saying, "And thus I, the president of the Poldavian Parliament, live in exile, in such a state of poverty that I do not even own a pair of trousers." He climbed onto a chair and was seen to be in his undershorts.

In 1939, André Weil was living in Helsinki. On November 30 of that year,

the Russians conducted the first bomb attack on Helsinki. Shortly after the incident, Weil was wandering around the wrong place at the wrong time; his squinty stare and obviously foreign attire brought him to the attention of the police, and he was arrested. A few days later the authorities conducted a search of Weil's apartment in his presence. They found

- Several rolls of stenotypewritten paper at the bottom of a closet; Weil claimed that these were the pages of a Balzac novel.
- A letter, in Russian, from Pontryagin. It was arranging a visit of Weil to Leningrad.
- A packet of calling cards belonging to Nicolas Bourbaki, member of the Royal Academy of Poldavia.
- Some copies of Bourbaki's daughter Betti's wedding invitations.

In all, this was an incriminating collection of evidence. Weil was slammed into prison for good.

A few days later, on December 3, Rolf Nevanlinna (1895–1980)—at that time a reserve colonel on the general staff of the Finnish Army—was dining with the chief of police. [It should be noted that Nevanlinna was an extremely distinguished mathematician—a complex analyst—in his own right. He was teacher to future Fields Medalist Lars Ahlfors.] Over coffee, the chief allowed that, "Tomorrow we are executing a spy who claims to know you. Ordinarily I wouldn't have troubled you with such trivia, but since we're both here anyway, I'm glad to have the opportunity to consult you." "What is his name?" inquired Nevanlinna. "André Weil." You can imagine Nevanlinna's shock—for André Weil was a world-renowned mathematician of the first rank. But he maintained his composure and said, "I know him. Is it really necessary to execute him?" The police chief replied, "Well, what do you want us to do with him?" "Couldn't you just deport him?" asked Nevanlinna innocently. "Well, there's an idea; I hadn't thought of it," replied the chief of police. And so André Weil's fate was decided.

When Weil was deported from Finland he was taken in custody to England and then to France. In France he was a member of the army reserves, and he was immediately jailed for failure to report for duty. He liked to say many years later that jail was the perfect place to do mathematics: it was quiet, the food was not bad, and there were few interruptions. In fact

Weil wrote one of his most famous works—the book *Basic Number Theory* [WEIL1]—during his time in prison. In later years Hermann Weyl (1885–1955) threatened to have Weil put back in jail so that he would be more productive! In fact Weil himself wrote to his wife from his jail cell with the following sentiment:

> My mathematics work is proceeding beyond my wildest hopes, and I am even a bit worried—if it is only in prison that I works so well, will I have to arrange to spend two or three months locked up every year?

Reading his memoir [WEIL2], one gets the sense that the young Weil felt that the Bourbaki group was in effect re-inventing mathematics for the twentieth century. In particular, they endeavored to standardize much of the terminology and the notation. Weil in fact takes credit for inventing the notation $\emptyset$ for the empty set (i.e., the set with no elements). The way he tells it, the group was looking for a good way to denote this special set; Weil was the only person in the group who knew the Norwegian alphabet, and he suggested that they co-opt this particular letter. In fact this consideration went into the *very first* Bourbaki book—on set theory! It is remarkable—at least to someone with mathematical training—that the notation for the empty set was still being debated in the late 1930s.

The Nicolas Bourbaki group was formed in the 1930s. Each of the founding members of the organization was himself a prominent and accomplished mathematician. Each had a broad view of the subject, and a clear vision of what Bourbaki was meant to be and what it set out to accomplish. Even though the books of Bourbaki became well known and widely used throughout the world, the identity of the members of Bourbaki was a closely guarded secret. Their meetings, and the venues of those meetings, were kept under wraps. The inner workings of the group were not leaked by anyone.

Over time, new members were added to Bourbaki—for instance Alexandre Grothendieck of the Institut des Hautes Études Scientifiques and Hyman Bass of Columbia University and Serge Lang of Columbia and Yale worked with Bourbaki. And others dropped out. But the founding group consisted exclusively of French mathematicians, ones with a coherent vision for the needs of twentieth-century mathematics.

The membership of Bourbaki was dedicated to the writing of the fundamental texts—in all the basic subject areas—in modern mathematics. Bourbaki's method for producing a book was as follows:

- The first rule of Bourbaki is that they would not write about a mathematical subject unless **(i)** it was basic material that any mathematics graduate student should know and **(ii)** it was mathematically "dead". This second desideratum meant that the subject area must no longer be an active area of current research in mathematics. Considerable discussion was required among the Bourbaki group to determine which were the proper topics for the Bourbaki books.

- Next there would be extensive and prolonged discussion of the chosen subject area: what are the important components of this subject, how do they fit together, what are the milestone results, and so forth. If there were several different ways to approach the subject (and often in mathematics that will be the case), then due consideration was given to which approach the Bourbaki book would take. The discussions we are describing here often took a long weekend, or several long weekends. The meetings were punctuated by long and sumptuous meals at good French restaurants.

- Finally someone would be selected to write the first draft of the book. This of course was a protracted affair, and could take as long as a year or more. Jean Dieudonné, one of the founding members of Bourbaki, was famous for his skill and fluidity at writing. Of all the members of Bourbaki, he was perhaps the one who served most frequently as the scribe. Dieudonné was also a prolific mathematician and writer in his own right. [16]

- After a first draft had been written, copies would be made for the members of the Bourbaki group. And they would read every word—assiduously and critically. Then the group would have another meeting or series of meetings—punctuated as usual by sumptuous repasts at elegant French restaurants—in which they would go through the book page by page or even line by line. The members of Bourbaki were good friends, and had the highest regard for each other as scholars, but they would argue vehemently over particular words or particular sentences

[16] There is one particular incident which serves to delineate the two lines of his work. Dieudonné once published a paper under Bourbaki's name, and it turned out that this paper had a mistake in it. Some time later, a paper was published entitled, "On an error of M. Bourbaki", and the signed author was Jean Dieudonné.

in the Bourbaki text. It would take some time for the group to work together through the entire first draft of a future Bourbaki book.

- After the group got through that first draft, and amassed a copious collection of corrections and revisions and edits, then a second draft would be created. This task could be performed by the original author of the first draft, or by a different author. And then the entire cycle of work would repeat itself.

It would take several years, and many drafts, for a new Bourbaki book to be created. The first Bourbaki book, on set theory, was published in 1939; Bourbaki books, and new editions thereof, have appeared as recently as 2005. So far there are thirteen volumes in the monumental series *l'Éléments de Mathématique*. These compose a substantial library of modern mathematics at the level of a first or second year graduate student. Topics covered range from abstract algebra to point-set topology to Lie groups to real analysis. The writing in the Bourbaki books is crisp, clean, and precise. Bourbaki has a very strict notion of mathematical rigor. For example, *no Bourbaki books contain any pictures!* That's right. Bourbaki felt that pictures are an intuitive device, and have no place in a proper mathematics text. If the mathematics is written correctly then the ideas should be clear—at least after sufficient cogitation. The Bourbaki books are written in a strictly logical fashion, beginning with definitions and axioms and then proceeding with lemmas and propositions and theorems and corollaries. Everything is proved rigorously and precisely. There are few examples and little explanation. Mostly just theorems and proofs. There are no "proofs omitted", no "sketches of proofs", and no "exercises left for the reader". In every instance, Bourbaki gives us the whole enchilada.

The Bourbaki books have had a considerable influence in modern mathematics. For many years, other textbook writers sought to mimic the Bourbaki style. Walter Rudin was one of these, and he wrote a number of influential texts without pictures and adhering to a strict logical formalism. Certainly, in the 1950s and 1960s and 1970s, Bourbaki ruled the roost. This group of dedicated French mathematicians with the fictitious name had set a standard to which everyone aspired. It can safely be said that an entire generation of mathematics texts danced to the tune that was set by Bourbaki.

But fashions change. It is now a commonly held belief in France that Bourbaki caused considerable damage to the French mathematics enterprise. How could this be? Given the value system for mathematics that we have

been describing in this book, given the passion for rigor and logic that is part and parcel of the subject, it would seem that Bourbaki would be our hero for some time to come. But no. There are other forces at play.

One feature of Bourbaki is that the books were only about *pure* mathematics. There are no Bourbaki books about applied partial differential equations, or control theory, or systems science, or theoretical computer science, or cryptography, or any of the other myriad areas where mathematics is applied. Another feature of Bourbaki is that it rejects intuition of any kind.[17] Certainly one of the main messages of the present book is that we *record* mathematics for posterity in a strictly rigorous, axiomatic fashion. This is the mathematician's version of the reproducible experiment with control used by physicists and biologists and chemists. But we *learn* mathematics, we *discover* mathematics, we *create* mathematics using intuition and trial and error. Certainly we draw pictures. Certainly we try things and twist things around and bend things to try to make them work. Unfortunately, Bourbaki does not teach any part of this latter process.

Thus, even though Bourbaki has been a role model for what recorded mathematics ought to be, even though it is a shining model of rigor and the axiomatic method, it is not necessarily a good and effective teaching tool. So, in the end, Bourbaki has not necessarily completed its grand educational mission. Whereas, in the 1960s and 1970s, it was quite common for Bourbaki books to be used as texts in courses all over the world, now the Bourbaki books are rarely used anywhere in classes. They are still useful references, and helpful for self-study. But, generally speaking, there are much better texts written by other authors. We cannot avoid saying, however, that those "other authors" certainly learned from Bourbaki. Bourbaki's influence is still considerable.

One amusing legacy that Bourbaki has left us is a special symbol used to denote a tricky passage in one of the Bourbaki texts. The reader who has traveled to Europe may remember the universal road signs to indicate a curvy or dangerous road. See Figure 4.27. This is the sign that Bourbaki uses to denote mathematical danger. And this is perhaps the artifact of Bourbaki that lives on most universally. Many another textbook writer uses the "Bourbaki wiggly curve" to mark challenging passages in his text.

[17] In this sense Bourbaki follows a grand tradition. The master mathematician Carl Friedrich Gauss used to boast that an architect did not leave up the scaffolding so that people could see how he constructed a building. Just so, a mathematician does not leave clues as to how he constructed or found a proof.

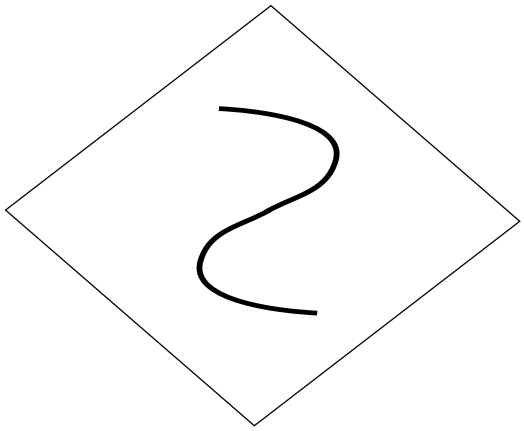


Figure 4.27

David Hilbert (discussed earlier in this section) and Nicolas Bourbaki have played an important role in the development of twentieth-century mathematics. It can safely be said that today mathematics is practiced in just the same way all over the world. Everyone uses the same terminology. Everyone embraces the same standards of rigor. Everyone uses the axiomatic method in just the same way. And we can thank David Hilbert and Nicolas Bourbaki for showing us the way and setting the lasting example.

Jean Dieudonné (1906–1992) waxes euphoric in describing how Bourbaki cuts to the quick of mathematics:

> What remains: the archiclassic structures (I don't speak of sets, of course), linear and multilinear algebra, a little general topology (the least possible), a little topological vector spaces (as little as possible), homological algebra, commutative algebra, noncommutative algebra, Lie groups, integration, differentiable manifolds, Riemannian geometry, differential topology, harmonic analysis and its prolongations, ordinary and partial differential equations, group representations in general, and in its widest sense, analytic geometry. (Here of course I mean in the sense of Serre, the only tolerable sense. It is absolutely intolerable to use *analytic geometry* for linear algebra with coordinates, still called analytical geometry in the elementary books. Analytical geometry in this sense has never existed. There are only people who do linear algebra badly, by taking coordinates and this they call analytical

> geometry. Out with them! Everyone knows that analytical geometry is the theory of analytical spaces, one of the deepest and most difficult theories of all mathematics.) Algebraic geometry, its twin sister, is also included, and finally the theory of algebraic numbers.

When Ralph Boas was the Executive Editor of the *Mathematical Reviews*, he made the mistake of printing the opinion that Bourbaki does not actually exist—in an article for the *Encyclopædia Britannica*, no less. The *Encyclopædia* subsequently received a scalding letter, signed by Nicolas Bourbaki, in which he declared that he would not tolerate the notion that anyone might question his right to exist. To avenge himself on Boas, Bourbaki began to circulate the rumor that Ralph Boas did not exist. In fact Bourbaki claimed that "Boas" was actually an acronym; the letters B.O.A.S. were actually a pseudonym for a group of the *Mathematical Reviews'* editors.

"Now, really, these French are going too far. They have already given us a dozen independent proofs that Nicolas Bourbaki is a flesh and blood human being. He writes papers, sends telegrams, has birthdays, suffers from colds, sends greetings. And now they want us to take part in their canard. They want him to become a member of the American Mathematical Society (AMS). My answer is 'No'." This was the reaction of J. R. Kline, secretary of the AMS, to an application from the legendary Nicolas Bourbaki.

4.8 Srinivasa Ramanujan and a New View of Proof

Srinivasa Ramanujan (1887–1920) was one of the most remarkable and talented mathematicians of the twentieth century. Born in poverty in the small village of Erode, India, Ramanujan attended a number of elementary schools. He showed considerable talent in all his subjects. At the age of 13 Ramanujan began to conduct his own independent mathematical investigations. He determined how to sum various kinds of series (arithmetical and geometric). At age 15 Ramanujan was shown how to solve a cubic or third-degree polynomial equation. He used that idea to devise his own method for solving the quartic or fourth-degre equation. His efforts to solve the quintic or fifth degree equation were of course doomed to failure, because Abel had shown many years before that this was impossible.

By age 17, Ramanujan had studied on old textbook of G. S. Carr and developed considerably in mathematical sophistication. He began his own investigations of Euler's constant and of Bernoulli numbers.

Ramanujan earned a scholarship to attend the Government College in Kumbakonam. He began his studies at age 17 in 1904, but was soon dismissed because he spent all his time on mathematics and let his other studies languish. He studied hypergeometric series and elliptic functions on his own.

In 1906 Ramanujan endeavored to pass the entrance exams to the University of Madras. But he could only pass the mathematical part of the test, so he failed.

During all this time Ramanujan's health was quite fragile. He suffered from smallpox when he was just 2 years old and other recurrent ailments of a serious nature throughout his young adult life. In 1909 Ramanujan had an arranged marriage with a 10-year-old girl named S. Janaki Ammal.

Ramanujan was able to publish some of his ideas about elliptic functions and Bernoulli numbers in the *Journal of the Indian Mathematical Society*. He gained thereby a considerable reputation as a young mathematical genius.

In order to support himself while he studied mathematics, Ramanujan occupied various low-paying clerkships. In 1912 and 1913, Ramanujan wrote to a number of prominent British mathematicians seeking advice and help with his work. These included M. J. M. Hill, E. W. Hobson, H. F. Baker, and G. H. Hardy. It was Hardy who truly understood Ramanujan's genius and who decided to take action. This was surely the turning point of Ramanujan's life, and of his mathematical career.

In 1914 Hardy was able to bring Ramanujan to Trinity College at Cambridge University. Ramanujan was an orthodox Brahmin and also a strict vegetarian. The former property made his travel very difficult, and the latter made it difficult for Ramanujan (a man of ill health) to find proper nourishment in World War I Britain. Nonetheless, Hardy and Ramanujan immediately began a most fruitful and dynamic mathematical collaboration.

But there were difficulties too, and that is really the point of the present discussion. Ramanujan had very little formal education. He did not see nor understand the importance of mathematical proof. He just "saw" things in an almost mystical fashion, and let others worry about how to prove them. Often Ramanujan was right in profound and important ways. But he made mistakes too. His view was that the gods spoke to him and gave him these mathematical insights. Hardy was one of the most accomplished and powerful mathematicians of his day, and he found Ramanujan's view

frustrating. Of course Hardy himself could sometimes provide the needed proofs; but sometimes not.

Ramanujan found his ill health to be too much of a burden by 1919, and he returned to India. He died within a year of tuberculosis.

Ramanujan left behind a number of remarkable notebooks filled with his formulas and his insights. But no proofs. Mathematicians even today are hard at work filling in the details so that we may understand the derivations of Ramanujan's astonishing relationships among numbers. We close with a famous story that illustrates Ramanujan's almost magical powers.

During Ramanujan's final days in Great Britain, he lay in a hospital ill with tuberculosis. Of course Hardy went to visit him frequently. The great professor was not much of a conversationalist, and he struggled to find things to say to his friend. One day Hardy walked in the door and mumbled, "I thought the number of my taxicab was 1729. It seemed rather a dull number." Ramanujan's instantaneous reaction was, "No, Hardy! No, Hardy! It is a very interesting number. It is the smallest number expressible as the sum of two cubes in two different ways."

Of course once Hardy knew this fact he could prove it without much difficulty. But how did Ramanujan see it?

4.9 Perplexities and Paradoxes

One upshot of the investigations into the foundations of mathematics that are the hallmark of twentieth-century research is a variety of rather troubling paradoxes that have come to the fore. We have already mentioned Russell's Paradox in Section 0.7. We take the opportunity in this section to indicate a few others.

The point of these paradoxes is that they are results that can be proved logically—and *correctly*—from the axioms of set theory as we know them. But these theorems are so counterintuitive, and so astonishing, that we cannot help but wonder whether mathematics is fraught with internal contradictions, or whether we are all on a fool's errand.

Of course we know better than this. All the paradoxes that are discussed here *can* be explained, and we shall certainly give an indication of what those explanations are. It takes some time, and some effort, to become inured to these paradoxes. We may consider this treatment to be your first exposure.

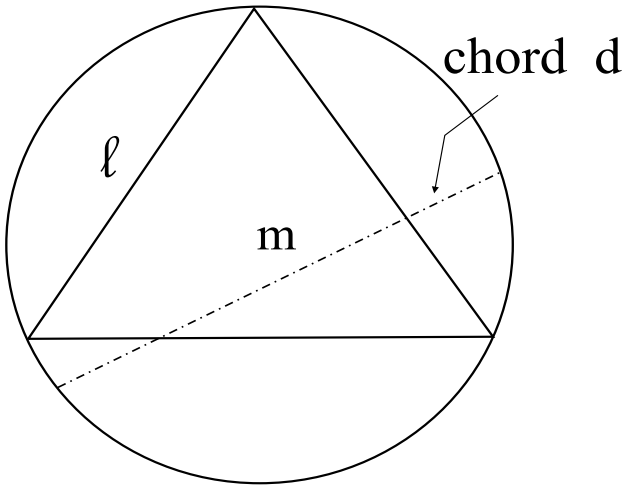


Figure 4.28

4.9.1 Bertrand's Paradox

This paradox was discovered a few hundred years ago. It is part of the reason that the subject of probability theory had such a rocky start. The proper resolution of this paradox, and why it really harbors no inherent contradiction, did not come about until the late 1930s. So it is proper grist for our mill here.

Fix a circle of radius 1. Draw the inscribed equilateral triangle as shown in Figure 4.28. We let ℓ denote the length of a side of this triangle. Suppose that a chord d (with length m) of the circle is chosen "at random." What is the probability that the length m of d exceeds the length ℓ of a side of the inscribed triangle?

The "paradox" is that this problem has three different but equally valid solutions. We now present these apparently contradictory solutions in sequence. At the end we shall explain why it is possible for a problem like this to have three distinct solutions.

Solution 1: Examine Figure 4.29. It shows a shaded, open disc whose boundary circle is internally tangent to the inscribed equilateral triangle. If the center of the random chord d lies *inside* that shaded disc, then $m > \ell$. If the center of the random chord d lies *outside* that shaded disc, then $m \leq \ell$.

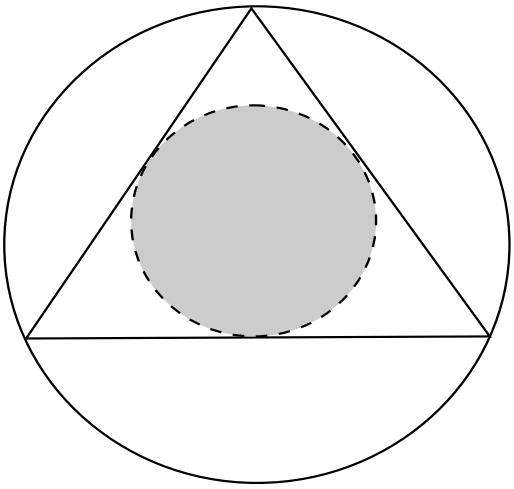


Figure 4.29

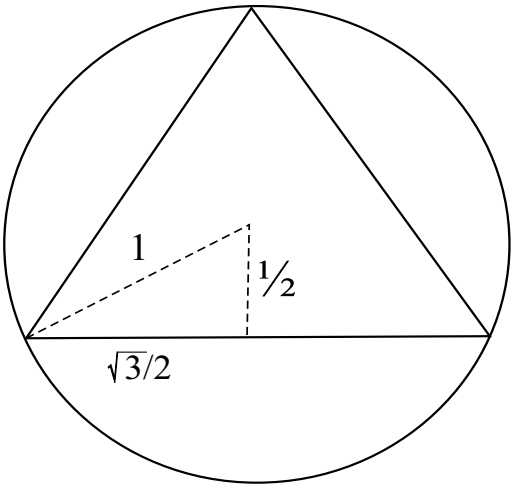


Figure 4.30

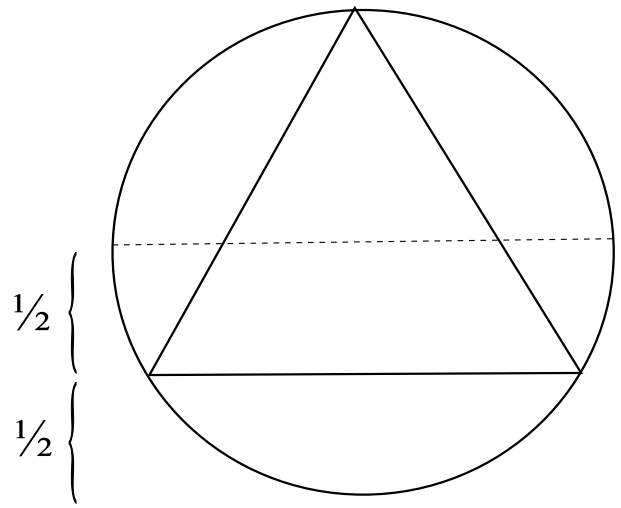


Figure 4.31

Thus the probability that the length d is greater than the length ℓ is

$$\frac{\text{area of shaded disc}}{\text{area of unit disc}}.$$

But an analysis of the equilateral triangle (Figure 4.30) shows that the shaded disc has radius $1/2$ hence area $\pi/4$. The larger unit disc has area π. The ratio of these areas is $1/4$. We conclude that the probability that the length of the randomly chosen chord exceeds ℓ is $1/4$.

Solution 2: Examine Figure 4.31. We may as well assume that our randomly chosen chord is horizontal (the equilateral triangle and the chord can both be rotated so that the chord is horizontal and one side of the triangle is horizontal). Notice that if the height, from the base of the triangle, of the chord d is less than or equal to $1/2$ then $m \leq \ell$ while if the height is greater than $1/2$ (and not more than 1) then $m > \ell$. We thus see that there is probability $1/2$ that the length m of d exceeds the length ℓ of a side of the equilateral triangle.

Solution 3: Examine Figure 4.32. We may as well assume that one vertex of our randomly chosen chord occurs at the lower left vertex A of the inscribed triangle (by rotating the triangle we may always arrange this to be the case). Now look at the angle θ that the chord subtends with the tangent line to the

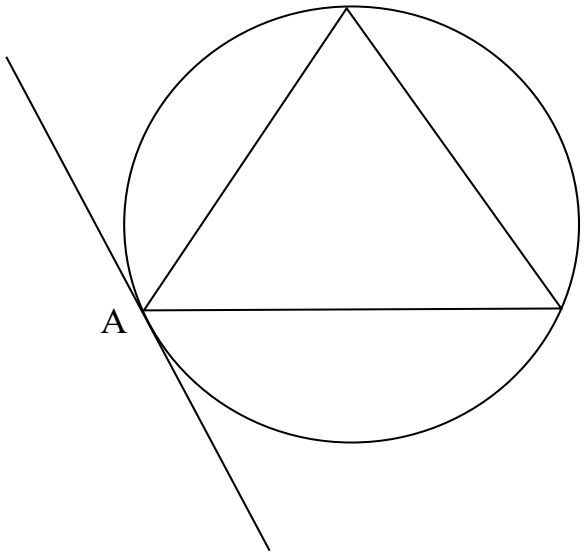


Figure 4.32

circle at the vertex A (shown in the Figure 4.33). If that angle is between $0°$ and $60°$ inclusive then the chord is shorter than or equal to ℓ. If the angle is strictly between $60°$ and $120°$ then the chord is longer than ℓ. Finally, if the angle is between $120°$ and $180°$ inclusive then the chord is shorter than ℓ. In sum we see that the probability is $60/180 = 1/3$ that the randomly chosen chord has length exceeding ℓ.

We have seen, then, three solutions to our problem. And they are different: we have found valid answers to be $1/4$, $1/2$, and $1/3$. How can a perfectly reasonable problem have three distinct solutions? And be assured that each of these solutions is correct! The answer is that, when one is dealing with a probability space having infinitely many elements (that is to say, a problem in which there are infinitely many outcomes—in this case there are infinitely many positions for the random chord), then there are infinitely many different ways to fairly assign probabilities to those different outcomes. Our three distinct solutions arise from three distinct ways to assign probabilities: notice that one of these is based on area, one is based on height, and one is based on angle.

For many years, because of paradoxes such as this one, the subject of probability theory was in ill repute. It was not until the invention of a

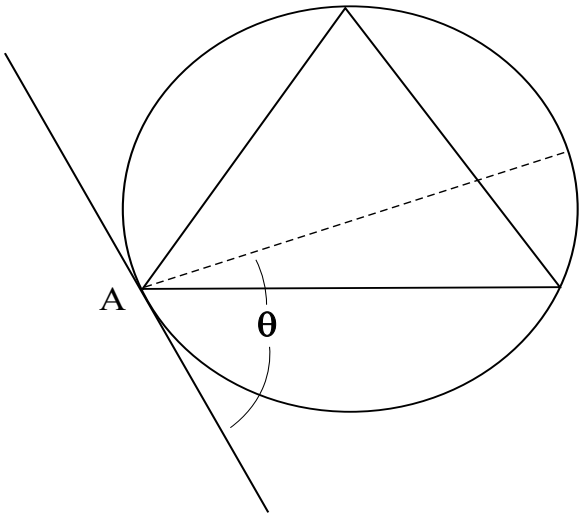


Figure 4.33

branch of mathematics called "measure theory" (Henri Lebesgue, 1901) that the tools became available to put probability theory on a rigorous footing. The celebrated Soviet mathematician A. Kolmogorov is credited with this development. These ideas are treated in advanced courses on measure theory and probability.

4.9.2 The Banach-Tarski Paradox

Alfred Tarski (1902–1983) was one of the pioneering logicians of the twentieth century. Stefan Banach (1892–1945) was one of the great mathematical analysts. In 1924 they published a joint paper presenting the remarkable paradox that we are about to describe. It is a byproduct of the intense examination of the axioms of set theory that was the hallmark of the first half of the twentieth century.

> **The Banach-Tarski Paradox:** *It is possible to take a solid ball of radius 1, break it up into 7 pieces, and reassemble those pieces into two solid balls of radius 1. Refer to Figure 4.34.*

How could this be? The statement seems to contradict fundamental principles of physics, and to gainsay the notion that we have any concept of volume or distance. After all, we are beginning with a ball of volume $4\pi/3$ and ending up with two balls having total volume $8\pi/3$.

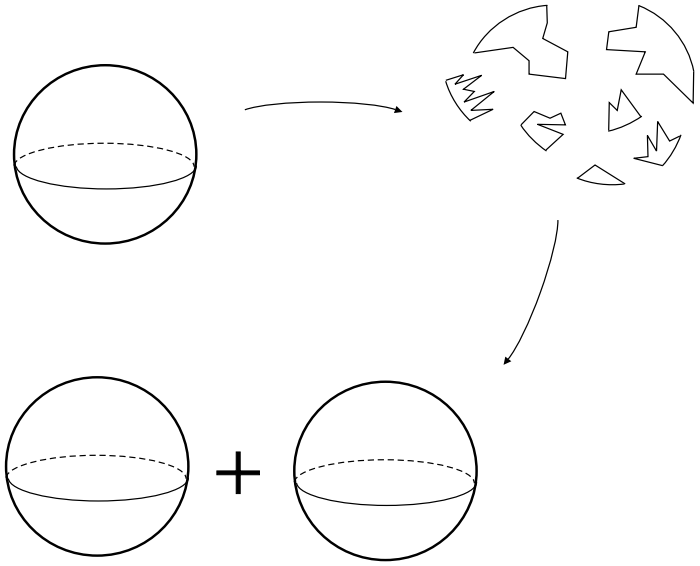


Figure 4.34

And in fact the Banach-Tarski paradox has even more dramatic formulations. It is actually possible to take a solid ball of radius 1, break it up into finitely many pieces, and reassemble those pieces into a full-sized replica of the Empire State Building. See Figure 4.35.

What could possibly be the explanation for this conundrum? It all has to do with measure theory, a subject pioneered by Henri Lebesgue (1875–1941) in 1901. The goal of measure theory is to assign a "measure" or size or volume to every set. It turns out—and this is a consequence of the Axiom of Choice, about which we shall say more below—that this goal is an impossible one. Any attempt to assign a measure to every set is doomed to failure. Thus there are certain sets, which we can identify explicitly by a concrete criterion, which are called *measurable*.[18] These are the sets to which we are allowed to assign a measure, and we can be assured (by the mathematical theory) that no contradiction will result thereby. The other sets are called *non-measurable*. And we are virtually guaranteed that any attempt to assign measures to the non-measurable sets will lead to trouble.

[18]Measurable sets have the intuitively appealing and attractive property that the measure, or volume, of the union of two disjoint sets is the sum of the measures or volumes of the two component sets. Thus, if we confine our attention to measurable sets, then the Banach-Tarski paradox is ruled out.

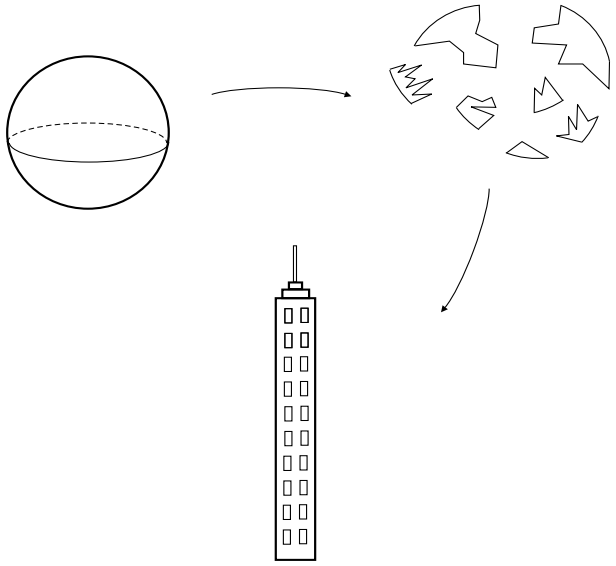


Figure 4.35

The long and the short of it is that the seven sets that are created in the Banach-Tarski Paradox are *non-measurable sets*. Thus they do not enjoy any of the intuitive properties that we expect a measure on sets to have. In particular, one of the desiderata for a measure is that if A and B are disjoint sets, then the measure of the union of A and B should equal the sum of the measures of A and B. For measurable sets this is true, and can be proved. But for non-measurable sets it is not.

Measure theory is now a highly developed subject, and is used routinely in real analysis, Fourier analysis, wavelets, image compression, signal processing, and many other parts of the mathematical sciences. We may do so blithely, because we confine our attention to measurable sets. We are all aware of the Banach-Tarski Paradox, but we avoid it like the plague. A nice treatment of the Banach-Tarski paradox appears in [WAG]. See also [JEC].

4.9.3 The Monty Hall Problem

Strictly speaking, this is not a paradox. It is not the sort of high-level assault on the foundations of our subject that the preceding two examples have described. But it is another situation where the product of mathematical discourse runs counter to intuition, and produces confusing results. And it

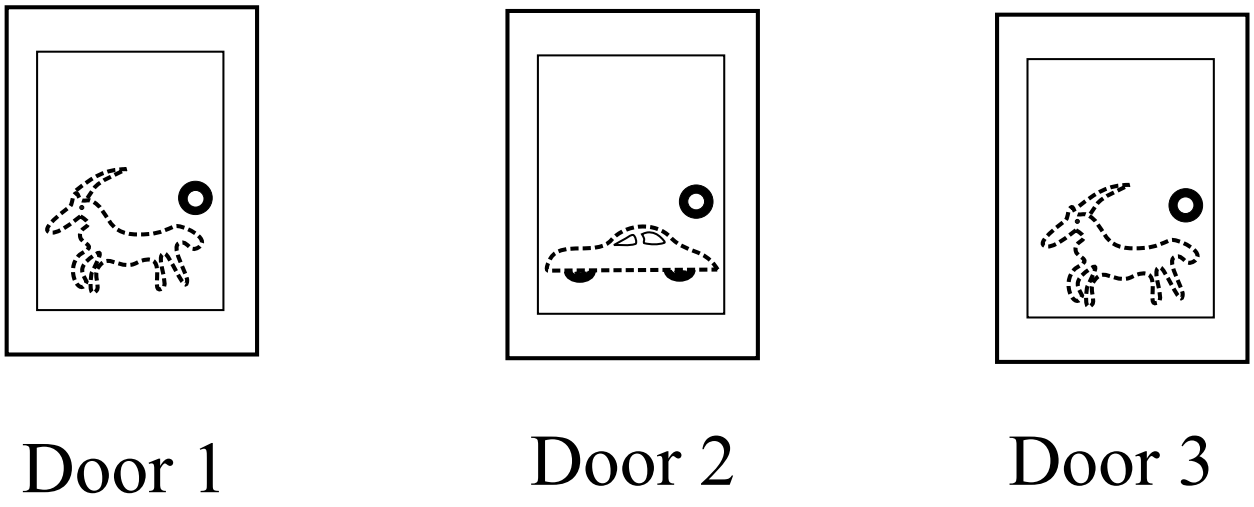


Figure 4.36

certainly made a major splash in American mathematics.

The Monty Hall problem has received a considerable amount of publicity in the last few years. It was inspired by the television game show LET'S MAKE A DEAL. The format of the game show (a bit over-simplified) is as follows. The contestant is faced with three doors. He/she knows that behind one door is a very desirable prize—say a fancy car. Behind the other two doors are rather pesky and undesirable items—say that a goat is behind each. See Figure 4.36. The contestant is to pick a door (blind), and is awarded the prize that is behind the door. But the game show host, Monty Hall, teases and cajoles and bribes the contestant, encouraging the contestant to change his/her mind and forcing the contestant to become confused over which is the most desirable door.

What has become known as the "Monty Hall" problem is this: The contestant picks a door. For the sake of argument, let us say that he/she has picked Door 3. Before the door is opened, revealing what is behind it, Monty Hall says "I will now reveal to you what is behind one of the other doors." A door is opened and there stands a goat. Then Monty Hall says "Would you like to change *your* door selection?" Very interesting.

Clearly the contestant will not pick the door that Monty Hall has already opened, since that has a goat behind it. So the issue is whether the contestant will switch from the currently selected door to the remaining door (the one

that the contestant has not chosen and Monty Hall did not open). A naive approach would be to say there is an equal probability for there to be a goat behind the remaining door and behind the door that the contestant has already selected—after all, one door has a goat and one has a car. What is the point of switching? However this naive approach does not take into account the fact that there are two distinct goats. A more careful analysis of cases occurs in our solution to the problem, and reveals a surprising answer.

Let us use a case-by-case analysis to solve the Monty Hall problem. We denote the goats by G_1 and G_2 (for goat one and goat two) and the car by C. For simplicity, we assume that the contestant will always select Door Three. We may not, however, assume that Monty Hall always reveals a goat behind Door One; for there may not be a goat behind Door One (it could be behind Door Two). Thus there are several cases to consider:

Door 1	Door 2	Door 3
G_1	G_2	C
G_2	G_1	C
G_1	C	G_2
G_2	C	G_1
C	G_1	G_2
C	G_2	G_1

In general there are $n! = n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot 3 \cdot 2 \cdot 1$ different ways to arrange n objects in order. Thus there are $6 = 3!$ possible permutations of three objects. That is why there are six rows in the array. The table represents all the different ways that two goats and a car could be distributed behind the doors.

1. In the first case, Monty Hall will reveal a goat behind either Door 1 or Door 2. It is *not* to the contestant's advantage to switch (for the contestant has selected Door 3, which has the fancy car behind it), so we record **N**.

2. The second case is similar to the first, it is not to the contestant's advantage to switch, and we record **N**.

3. In the third case, Monty Hall will reveal a goat behind Door 1, and it *is* to the contestant's advantage to switch. We record **Y**.

4. The fourth case is like the third, and it is to the contestant's advantage to switch. We record **Y**.

5. In the fifth case, Monty Hall will reveal a goat behind Door 2. It *is* to the contestant's advantage to switch, so we record **Y**.

6. The sixth case is like the fifth, it is to the contestant's advantage to switch, and we record **Y**.

Observe that the tally of our case-by-case analysis is four **Y**'s and just two **N**'s. Thus the odds are two against one in favor of switching after Monty Hall reveals the goat. Put in other words, the player has a 2/3 probability of improving his/her position by switching doors.

The modern history of the Monty Hall problem is both amusing and alarming. As the reader may know, Marilyn vos Savant (1946–) is a popular newspaper columnist. [In point of fact "vos Savant" is not her real name, but is borrowed from a favorite aunt.] She is supposed to have the highest IQ in history (although there is word that a child in China has a measurably higher IQ). Her syndicated column, *Ask Marilyn*, is premised on her high intellectual powers. She is very clever, and has a knack for answering difficult questions (she usually has the good sense to ask experts when confronted with a problem on which she has no expertise). She rarely makes a mistake, although there is a Web site called *Marilyn is Wrong!* [`http://www.wiskit.com/marilyn.html`] which claims to point out a number of her *faux pas*.

Actually Marilyn vos Savant was educated in St. Louis (my stamping ground). She attended Meramec Junior College, but did not graduate. She also attended Washington University (my institution), but did not graduate. She is married to Robert K. Jarvik (who invented the artificial heart), and makes her home in New York City.

Marilyn vos Savant gained particular celebrity when one of her readers wrote in to ask about the "Monty Hall Problem." She checked her facts and described in her column the correct solution, as we have discussed above. Woe is us!! Over two thousand academic mathematicians wrote to Marilyn vos Savant and told her that she was in error. And some of them were not too polite about it. This was quite a debacle for American mathematics.

I have to say that this whole set of circumstances went to Marilyn vos Savant's head. When Andrew Wiles published his solution of Fermat's last problem, Ms. Savant published a little book [SAV] claiming that his solution

is incorrect. The basis for her daring allegation is twofold: **(i)** she offers a proof that the complex numbers do not exist (therefore nullifying Wiles's use of said number system) and **(ii)** she observes that Wiles uses hyperbolic geometry, in which the circle can be squared—but everyone knows that the circle *cannot be squared*—and this is a contradiction.

I wrote to Ms. vos Savant, and to her publisher, pointing out the error of her ways. And she answered! She said, "My mathematician friends and I had a good laugh over your submission. Keep those cards and letters coming." So much for scholarly discourse.

Martin Gardner is particularly chagrined for having been thanked for checking the math in vos Savant's book. He asserts vehemently that he did not. Barry Mazur of Harvard was also asked to check the math, but he managed to dodge the bullet.

Chapter 5

The History of the Four-Color Theorem

The computer has given us the ability to look at new mathematical worlds that would have remained inaccessible to the unaided human mind; but this access has come at a price. Many of these worlds, at present, can only be known experimentally. The computer has allowed us to fly through the rarefied domains of hyperbolic spaces and examine more than a billion digits of π, but experiencing a world and understanding it are very different.

J. Borwein, P. Borwein, R. Girgensohn, and S. Parnes

... the miracle of the appropriateness of the language of mathematics for the formulation of the laws of physics is a wonderful gift which we neither understand nor deserve. We should be grateful for it and hope that it will remain valid in the future and that it will extend ... to wide branches of learning.

Eugene Wigner

Mathematicians often underestimate the reliability of heuristic proof. Probably, the results of a good mathematician, working heuristically, are not less reliable than the results of an average rigorous mathematician. (One can consider this statement as a definition of a good mathematician.)

Albert Schwarz

I do still believe that rigor is a relative notion, not an absolute one. It depends on the background readers have and are expected to use in their judgment.

René Thom

The only proof capable of being given that an object is visible is that people actually see it ... In like manner, I apprehend, the sole evidence it is possible to produce that anything is desirable is that people do actually desire it.

John Stuart Mill

Don't use manual procedures ... and ... don't rely on social processes for verification.

David Dill

Philosophers have frequently distinguished mathematics from the physical sciences. While the sciences were constrained to fit themselves via experimentation to the real world, mathematicians were allowed more or less free reign within the abstract world of the mind. This picture has served mathematicians well for the past few millenia, but the computer has begun to change this.

J. Borwein, P. Borwein, R. Girgensohn, and S. Parnes

A mathematician is a machine that turns coffee into theorems.

Paul Erdős

We will grieve not, rather find strength in what remains behind.

William Wordsworth

5.1 Humble Beginnings

In 1852 Francis W. Guthrie, a graduate of University College London, posed the following question to his brother Frederick:

> Imagine a geographic map on the earth (i.e., a sphere) consisting of countries only—no oceans, lakes, rivers, or other bodies of water. The only rule is that a country must be a single contiguous mass—in one piece, and with no holes—see Figure 5.1. As cartographers, we wish to *color* the map so that no two adjacent countries will be of the same color (Figure 5.2—note that R, G, B, Y stand for red, green, blue, and yellow). How many colors should the map-maker keep in stock so that he can be sure he can color any map?

Frederick Guthrie was a student of Augustus De Morgan (1806–1871), and ultimately communicated the problem to his mentor. The problem was passed around among academic mathematicians for a number of years (in fact De Morgan communicated the problem to William Rowan Hamilton (1805–1865)). The first allusion in print to the problem was by Arthur Cayley (1821–1895) in 1878.

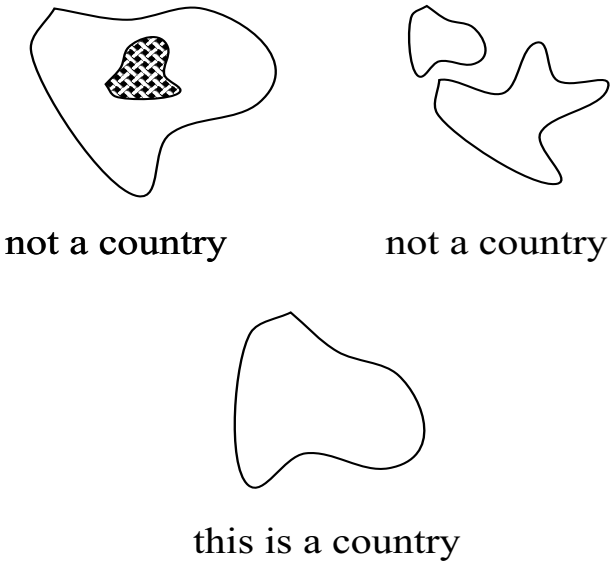


Figure 5.1

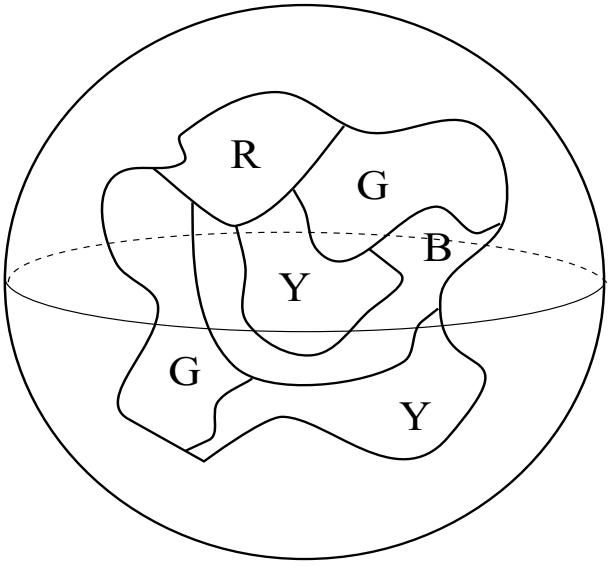


Figure 5.2

The eminent mathematician Felix Klein (1849–1925) in Göttingen heard of the problem and declared that the only reason the problem had never been solved is that no capable mathematician had ever worked on it. *He*, Felix Klein, would offer a class, the culmination of which would be a solution of the problem. He failed.

In 1879, A. Kempe (1845–1922) published a solution of the four-color problem. That is to say, he showed that any map whatever could be colored with four colors. Kempe's proof stood for eleven years. Then a mistake was discovered by P. Heawood (1861–1955). Heawood studied the problem further and came to a number of fascinating conclusions:

- Kempe's proof, particularly his device of "Kempe chains", *does* suffice to show that any map whatever can be colored with five colors.

- Heawood showed that if the number of edges around each region in the map is divisible by 3, then the map is 4-colorable.

- Heawood found a formula that gives an estimate for the "chromatic number" of any surface. Here the chromatic number $\chi(g)$ of a surface is the least number of colors it will take to color *any* map on that surface. We write the chromatic number as $\chi(g)$. In fact the formula is

$$\chi(g) \leq \left\lfloor \frac{1}{2}\left(7+\sqrt{48g+1}\right)\right\rfloor$$

  so long as $g \geq 1$.

  Here is how to read this formula. It is known, thanks to work of Camille Jordan (1838–1922) and August Möbius (1790–1868), that any surface in space is a sphere with handles attached. See Figure 5.3. The number of handles is called the *genus*, and we denote it by g. The Greek letter chi (χ) is the chromatic number of the surface—the least number of colors that it will take to color any map on the surface. Thus $\chi(g)$ is the number of colors that it will take to color any map on a surface that consists of the sphere with g handles. Next, the symbols $\lfloor\ \ \rfloor$ stand for the "greatest integer function". For example $\lfloor \frac{9}{2} \rfloor = 4$ just because the greatest integer in the number "four and a half" is 4. Also $\lfloor \pi \rfloor = 3$ because $\pi = 3.14159\ldots$ and the greatest integer in the number pi is 3.

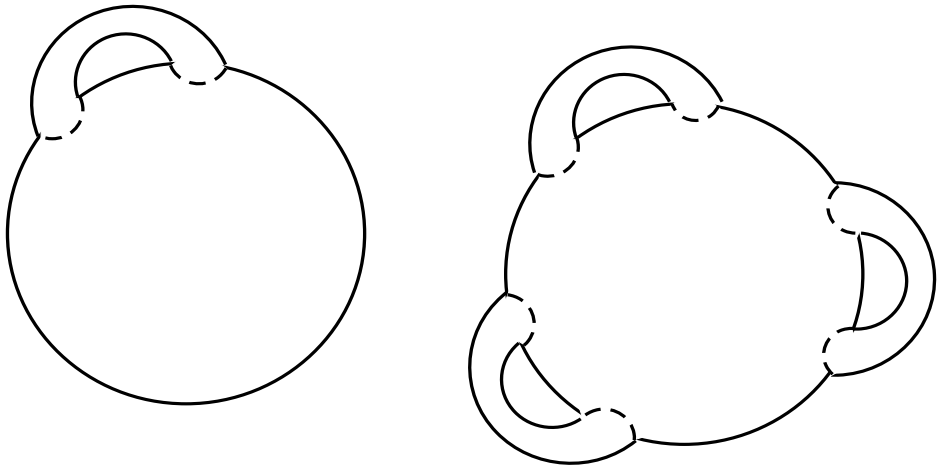


Figure 5.3

Now a sphere is a sphere with no handles, so $g = 0$. We may calculate that

$$\chi(g) \leq \left\lfloor \frac{1}{2}\left(7 + \sqrt{48 \cdot 0 + 1}\right) \right\rfloor = \left\lfloor \frac{1}{2}\,(8) \right\rfloor = 4\,.$$

This is the four-color theorem! Unfortunately, Heawood's proof was only valid when the genus is at least 1. It gives no information about the sphere.

The torus (see Figure 5.4) is topologically equivalent to a sphere with one handle. Thus the torus has genus $g = 1$. Then Heawood's formula gives the estimate 7 for the chromatic number. And in fact we can give an example—see Figure 5.6—of a map on the torus that requires 7 colors. Here is what Figure 5.5 shows. It is convenient to take a pair of scissors and cut the torus apart. With one cut, the torus becomes a cylinder; with the second cut it becomes a rectangle. The arrows on the edges indicate that the left and right edges are to be identified (with the same orientation), and the upper and lower edges are to be identified (with the same orientation). We call our colors "1", "2", "3", "4", "5", "6", "7". The reader may verify that there are seven countries shown in our Figure 5.6, and every country is adjacent to (i.e., touches) every other. Thus they all must have different colors! This is a map on the torus that requires 7 colors; it shows that Heawood's estimate is sharp for this surface.

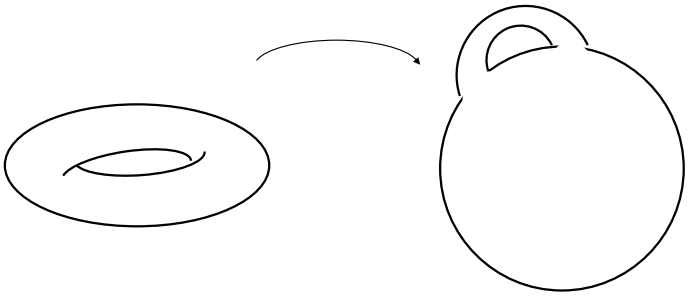


Figure 5.4

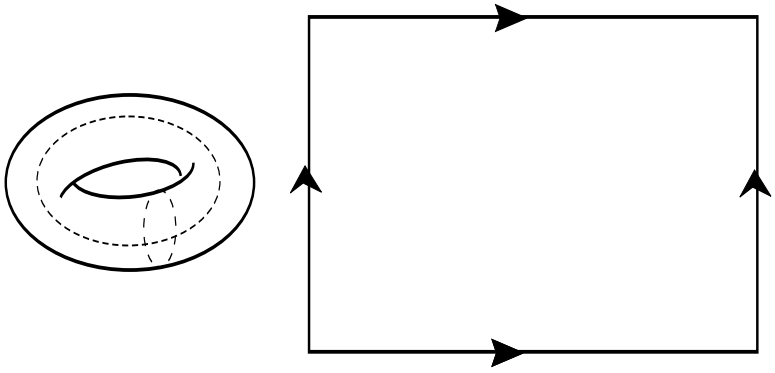


Figure 5.5

Heawood was unable to decide whether the chromatic number of the sphere is 4 or 5. He was also unable to determine whether any of his estimates for the chromatic numbers of various surfaces of genus $g \geq 1$ were sharp or accurate. That is to say, for the torus (the closed surface of genus 1), Heawood's formula says that the chromatic number does not exceed 7. Is that in fact the best number? Is there a map on the torus that really requires 7 colors? And for the torus with two handles (genus 2), Heawood's estimate gives an estimate of 8. Is that the best number? Is there a map on the double torus that actually *requires* 8 colors? And so forth: we can ask the same question for every surface of every genus. Heawood could not answer these questions.

The mathematician Tait produced another resolution of the four-color problem in 1880. Peterson pointed out a gap in 1891. Another instance of eleven years lapsing before the error was found!

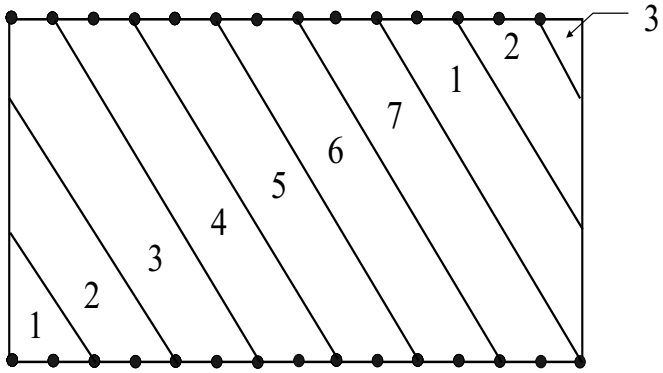


Figure 5.6

The four-color problem, as this conundrum came to be known, has a long and curious history. The great American mathematician G. D. Birkhoff did foundational work on the problem that allowed Philip Franklin (1898–1965) in 1922 to prove that the four-color conjecture is true for maps with at most 25 countries. Heesch made seminal contributions to the program, and in fact introduced the techniques of reducibility and discharging which were ultimately used by Appel and Haken in 1976 to solve the four-color problem. Walter Stromquist proved in his 1975 Harvard Ph.D. thesis [STR1] that, for any map with 100 or fewer countries, 4 colors will always suffice. See also [STR2]. What is particularly baffling is that Ringel and Youngs were able to prove in 1970 that all of Heawood's estimates, for the chromatic number of any surface, are sharp. So the chromatic number of a torus is indeed 7. The chromatic number of a "super torus" with two holes is 8. And so forth. But the Ringel/Youngs proof does not apply to the sphere. They could not improve on Heawood's result that 5 colors will always suffice.

Then in 1974 there was blockbuster news. Using 1200 hours of computer time on the University of Illinois supercomputer, Kenneth Appel and Wolfgang Haken showed that in fact 4 colors will always work to color any map on the sphere. Their technique is to identify 633 fundamental configurations of maps (to which all others can be reduced) and to prove that each of them is reducible in the sense of Heesch. But the number of "fundamental examples" was very large, and the number of reductions required was beyond the ability of any human to count. And the reasoning is extremely intricate and complicated. Enter the computer.

In those days computing time was expensive and not readily available, and Appel and Haken certainly could not get a 1200-hour contiguous time slice for their work. So the calculations were done late at night, "off the record", during various down times. In point of fact, Appel and Haken did not know for certain whether the calculation would ever cease. Their point of view was this:

- If the computer finally stopped then it will have checked all the cases and the 4-color problem was solved.
- If the computer never stopped then they could draw no conclusion.

Well, the computer stopped. But the level of discussion and gossip and disagreement in the mathematical community did not. Was this really a proof? The computer had performed tens of millions of calculations. Nobody could ever check them all. In 1974 our concept of a proof was etched in stone after 2500 years of development: a proof was a logical sequence of steps that one human being recorded on a piece of paper so that another human being could check them. Some proofs were quite long and difficult (for example, the proof of the celebrated Atiyah-Singer Index Theorem from the mid-1960s was four long papers in the *Annals of Mathematics* and used a great deal of mathematical machinery from other sources). But, nonetheless, they were always checkable by a person or persons. The new "proof" of Appel and Haken was something else again. It required one to place a certain faith in the computer, and in the algorithm that the computer was processing. The old IBM adage "Garbage In, Garbage Out" was in the forefront of everyone's mind.

But now the plot thickens. Because in 1975 a mistake was found in the proof. Specifically, there was something amiss with the algorithm that Appel and Haken fed into the computer. It was later repaired. The paper was published in 1976 [APH1]. The four-color problem was declared to be solved.

In fact Oscar Lanford has pointed out that, in order to justify a computer calculation as part of a proof, you must not only prove that the program is correct, but you must understand how the computer rounds numbers, how the operating system functions, and how the time-sharing system works. It would also help to know how the CPU (central processing unit, or "chip") stores data. There is no evidence thus far that those who use computers heavily in their proofs go beyond the first desideratum in this list.

In a 1986 article [APH2] in the *Mathematical Intelligencer*, Appel and Haken point out that the reader of their seminal 1976 article [APH1] must face

> 50 pages containing text and diagrams, 85 pages filled with almost 2500 additional diagrams, and 400 microfiche pages that contain further diagrams and thousands of individual verifications of claims made in the 24 lemmas in the main section of the text.

They go on to acknowledge that their proof required more than 1200 hours of computer time, and that there were certainly typographical and copying errors in the work. But they offer the reassurance that readers will understand "why the type of errors that crop up in the details do not affect the robustness of the proof." Several errors found subsequent to publication, they record, were "repaired within two weeks." By 1981 "about 40%" of 400 key pages had been independently checked, and 15 errors corrected, by U. Schmidt.

In fact, for many years after that, the University of Illinois Mathematics Department had a postmark that appeared on every outgoing letter from their department. It read:

Four Colors Suffice

Quite a triumph for Appel and Haken and their supercomputer.

But it seems as though there is always trouble in paradise. According to one authority, who prefers to remain nameless, errors continued to be discovered in the Appel/Haken proof. There was considerable confidence that any error that was found could be fixed. And invariably the errors were fixed. But the stream of errors never seemed to cease. So is the Appel/Haken work really a proof? Is a proof supposed to be some organic mass that is never quite right, that is constantly being fixed? Not according to the paradigm set down by Euclid 2300 years ago!

Well, there is hardly anything more reassuring than another, independent proof. Paul Seymour and his group at Princeton University found another way to attack the problem. In fact they found a new algorithm that seems to be more stable. They also needed to rely on computer assistance. But by the time they did their work computers were *much*, much faster. So they required much less computer time. In any event, this paper appeared in 1994

(see [SEY]). It has stood the test of 12 years, with no errors found. And in fact today Gonthier (2004) has used a computer-driven "mathematical assistant" to check the 1994 proof.[1]

But it is still the case that nobody can check the Seymour proof, in the traditional sense of "check". The computer is still performing many millions of calculations, and it is not humanly possible to do so much checking by hand—nor would anyone want to! The fact is, however, that over the course of twenty years, from the time of the original Appel/Haken proof to the advent of the Seymour proof, we as a community of scholars have become much more comfortable with computer-assisted proofs. There are still doubts and concerns, but this new methodology has become part of the furniture. There are enough computer-aided proofs around (some of them will be discussed in this book) that a broad cross-section of the community has come to accept them—or at least to tolerate them.

It is still the case that mathematicians are most familiar with, and most comfortable with, a traditional, self-contained proof that consists of a sequence of logical steps recorded on a piece of paper. We still hope that some day there will be such a proof of the four-color theorem. After all, it is only a traditional, Euclidean-style proof that offers the understanding, the insight, and the sense of completion that all scholars seek. For now we live with the computer-aided proof of the four-color theorem.

With the hindsight of thirty years, we can be philosophical about the Appel-Haken proof of the four-color theorem. What is disturbing about it, and about the Hales proof of the Kepler conjecture (discussed elsewhere in this book) is that these proofs lack the sense of *closure* that we ordinarily associate with mathematical proof. Traditionally, we invest several hours—or perhaps several days or weeks—absorbing and internalizing a new mathematical proof. Our goal in the process is to *learn something*.[2] The end result is new understanding, and a definitive feeling that something has been internalized and accomplished. These new computer proofs do not offer that reward. The Grisha Perelman proof of the Poincaré conjecture, and the classification of the finite, simple groups, also lack the sense of closure. In those cases computers do not play a role. Instead the difficulty is with the soci-

[1] This process has in fact become an entire industry. There is now a computer verification by Bergstra of Fermat's Last Theorem, and there is also a computer verification of the prime number theorem (about the distribution of the primes).

[2] And of course the admittedly selfish motivation is to learn some new techniques that will help the reader with his/her own problems.

ology of the profession and the subject. Those proofs are also discussed in these pages.

The real schism is, as Robert Strichartz [STR] has put it, between the quest for knowledge and the quest for certainty. Mathematics has traditionally prided itself on the unshakable absoluteness of its results. This is the value of our method of proof as established by Euclid. But there are so many new developments that have undercut the foundations of the traditional value system. And there are new societal needs: theoretical computer science and engineering and even modern applied mathematics require certain pieces of information and certain techniques. The need for a workable device often far exceeds the need to be *certain* that the technique can stand up to the rigorous rules of logic. The result may be that we shall re-evaluate the foundations of our subject. The way that mathematics is practiced in the year 2100 may be quite different from the way that it is practiced today.

Chapter 6

Computer-Generated Proofs

Automatic theorem proving remains a primitive art, able to generate only the most rudimentary arguments.

Arthur Jaffe

Arthur Jaffe represents the school of mathematical physicists who view their role as providing rigorous proofs for the doubtful practices of physicists. This is a commendable objective with a distinguished history. However, it rarely excites physicists who are exploring the front line of their subject. What mathematicians can rigorously prove is rarely a hot topic in physics.

Michael Atiyah

Thus it seems that the problem can be reduced to the determination of the minimum of a function of a finite number of variables, providing a programme realizable in principle. In view of the intricacy of this function we are far from attempting to determine the exact minimum. But, mindful of the rapid development of our computers, it is imaginable that the minimum may be approximated with great exactitude.

L. Fejes Tóth

As wider classes of identities, and perhaps even other kinds of classes of theorems, become routinely provable, we might witness many results for which we would know how to find a proof (or refutation); but we would be unable or unwilling to pay for finding such proofs, since "almost certainty" can be bought so much cheaper. I can envision an abstract of a paper, c. 2100, that reads, "We show in a certain precise sense that the Goldbach conjecture is true with probability larger than 0.99999 and that its complete truth could be determined with a budget of $10 billion."

As absolute truth becomes more and more expensive, we would sooner or later come to grips with the fact that few non-trivial results could be known with old-fashioned certainty. Most likely we will wind up abandoning the task of keeping track of price altogether and complete the metamorphosis to nonrigorous mathematics.

Doron Zeilberger

> Traditional papers are . . . a poor way to introduce a subject, but they are the whetstones which give the final edge to mastery. And they are at least a last resort in the learning process.
>
> Arthur Jaffe and Frank Quinn

> *Although there is an ongoing crisis in mathematics, it is not as severe as the crisis in physics. The untestability of parts of theoretical physics (e.g., string theory) has led to a greater reliance on mathematics for "experimental verification."*
>
> J. Borwein, P. Borwein, R. Girgensohn, and S. Parnes

> *The chief aim of all investigations of the external world should be to discover the rational order and harmony which has been imposed on it by God and which He revealed to us in the language of mathematics.*
>
> Johannes Kepler

6.1 A Brief History of Computing

The first counting *devices* were counting boards. The most primitive of these may be as old as 1200 B.C.E. and consisted of a board or stone tablet with mounds of sand in which impressions or marks could be made. Later counting boards had grooves or metal discs that could be used to mark a position. The oldest extant counting board is the Salamis tablet, dating to 300 B.C.E. See Figure 6.1.

The *abacus* was the first computing machine. Usually credited to the Chinese, it is first mentioned (as far as we know) in a book of the Eastern Han Dynasty, written by Xu Yue in 190 C.E. This is a frame, often made of wood, equipped with cylindrical dowels along which beads may slide. See Figure 6.2. The beads in the lower portion of the abacus each represent one unit, while the beads in the upper portion each represent five units.

Counting boards were prevalent in ancient Greece and Rome. The abacus developed slowly over a period of 1000 years. The modern abacus was popularized during the Song dynasty in China during the period 960–1127 C.E. In the succeeding centuries the use of the abacus spread to Japan and Korea. It is still in common use today. In fact in Taiwan and China one sees shopkeepers using an abacus to check the accuracy of the electronic cash register!

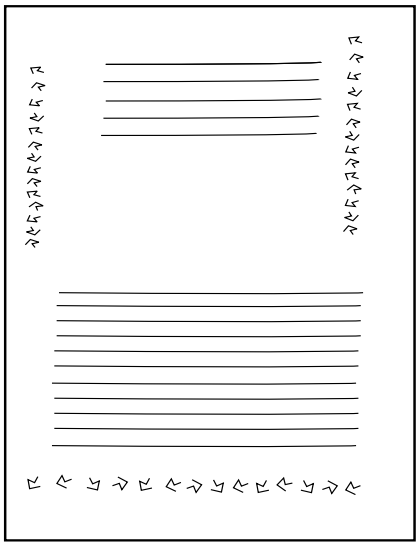


Figure 6.1. The Salamis tablet.

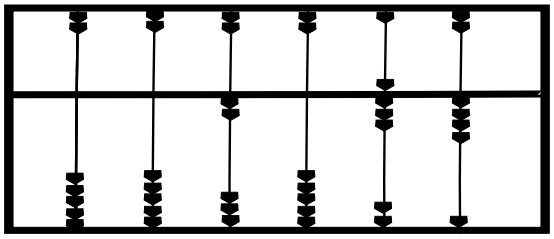


Figure 6.2. An abacus.

The first mechanical calculator was devised by Blaise Pascal (1623 C.E.–1662 C.E.), a mathematician who is remembered today for his work in probability theory and for his philosophical writings. Pascal built his machine in 1642, inspired by a design of Hero of Alexandria (c. 10 C.E.–70 C.E.). Pascal was interested in adding up the distance that a carriage traveled.

Pascal's design is still used today in water meters and automobile odometers. It is based on a single-tooth gear engaged with a multi-tooth gear (Figure 6.3). The basic idea was that the one-tooth gear was large enough that it only engaged the multi-tooth gear after one kilometer had been traversed by the carriage. None other than Gottfried Wilhelm von Leibniz (1646–1716) augmented Pascal's calculator so that it could multiply. In fact multiplication for this machine consisted of multiple additions. It is possible that techno-

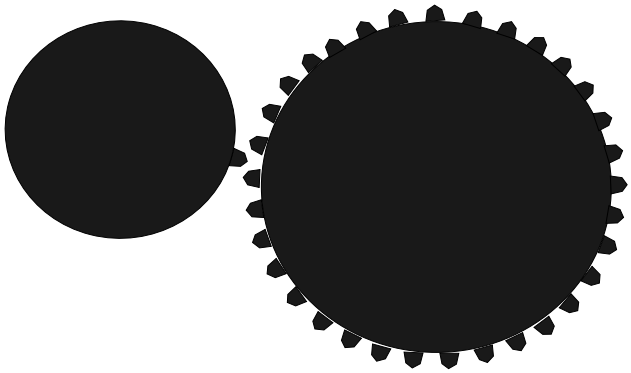


Figure 6.3

phobia was instigated by Pascal's invention, for even then mathematicians feared for their careers because of this new machine.

Tomas of Colmar (1785 C.E.–1870 C.E.) invented the first genuine mechanical calculator in 1820. It could add, subtract, multiply, and divide. Charles Babbage (1791 C.E.–1871 C.E.) capitalized on Colmar's idea by adding an insight of his own. Babbage realized as early as 1812 that many long calculations involved repetitive steps. He reasoned that one should be able to build a calculating machine that could handle those redundant steps automatically. He built a prototype of his "difference engine" in 1822. Soon thereafter he obtained a subsidy from the British government to develop his idea. In 1833, however, Babbage got an even better idea.

The better idea was to produce an "analytical engine" that would be a real parallel decimal computer. This machine would operate on "words" of 50 decimal places and was able to store 1000 such numbers. The analytical engine would have several built-in mathematical operations, and these could be executed in any order (as specified by the operator). The instructions for the machine were stored on punch cards (this idea was inspired by the Jacquard loom).

It is an interesting historical side-note that Lord Byron's daughter Augusta (1815 C.E.–1852 C.E.) played a role in Babbage's work. She was a skilled programmer (perhaps the first ever!), and created routines for his machine to calculate the Bernoulli numbers. Lord Byron is remembered as a poet, and he celebrated his daughter with these words:

> My daughter! with thy name this song begun—
> My daughter! with thy name thus much shall end—

I see thee not,—I hear thee not,—but none
Can be so wrapt in thee; thou art the friend
To whom the shadows of far years extend:
Albeit my brow thus never shouldst behold,
My voice shall with thy future visions blend,
And reach into thy heart,—when mine is cold,—
A token and a tone even from thy father's mould.

In 1890, Herman Hollerith (1860 C.E.–1929 C.E.) created a machine that would read punch cards that he produced. He was working for the U. S. Census Bureau, and he used the machine to tabulate census data. His "tabulating machine" was a terrific innovation, because it increased accuracy and reliability by a dramatic measure. Also the cards were a convenient and reliable way to store the data. In fact Hollerith's tabulator became so successful that he started his own firm to market the device; this company eventually became International Business Machines (or IBM).

Hollerith's machine was limited in its abilities; it could only do tabulations, not direct more complex computations. In 1936, Konrad Zuse (1910 C.E.–1995 C.E.) produced a machine (called the Z1) that could be termed the first freely programmable computer. The point here is that this was the first computing machine that was not dedicated to a particular task. Like a modern computer, it could be programmed to do a variety of things. The Z1 was succeeded in 1941 by a more advanced machine (called the Z3) that might be termed the first stored-program computer. It was designed to solve complex engineering problems. The machine was controlled by perforated strips of discarded movie film. It should be noted however that, unlike a modern computer, the Z3 was not all-electronic. It was mostly mechanical.

One remarkable feature of Zuse's Z3 was that it used the binary system for numbers. Thus all numbers were encoded using only 0s and 1s (as an instance, the number that we call 26 would be represented as 11010). Zuse may have been inspired by Alan Turing's "Turing Machine". In any event, the binary system is used almost universally on computers today.

The year 1942 saw the premiere of the world's first electronic-digital computer by Professor John Atanasoff (1903 C.E.–1995 C.E.) and his graduate student Clifford Berry (1918 C.E.–1963 C.E.) of Iowa State University. Their computer was called the ABC Computer, and it featured—among its many innovations—a binary system of arithmetic, parallel processing, regenerative memory, and a separation of memory and computing functions. Atanasoff

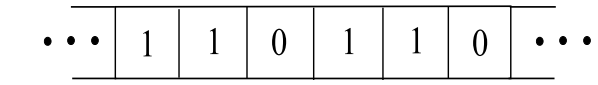


Figure 6.4

was awarded the National Medal of Science and Technology in 1990 in recognition of his work.

During the years 1939–1944, Howard Aiken (1900 C.E.–1973 C.E.) worked in collaboration with engineers at IBM to produce a large, automatic digital computer (based on standard IBM electromechanical parts). The machine, officially known as the "IBM automatic sequence controlled calculator" or ASCC, soon earned the nickname "Harvard Mark I". A remarkable piece of technology, the Harvard Mark I had 750,000 components, was 50 feet long, 8 feet tall, and weighed about 5 tons. It manipulated numbers that were 23 digits long; in fact the machine could add or subtract two such numbers in 0.3 seconds, multiply two numbers in 4 seconds, and divide them in 10 seconds.

The Harvard Mark I read data from punch cards and received instructions from a paper tape. It consisted of several different calculators which worked on different parts of the problem (much like a parallel processing machine today). Aiken went on to develop many computing machines, but the Harvard Mark I was perhaps the most significant. It was historically important, and was still in use at Harvard as late as 1959.

As an engineer, Howard Aiken was farseeing and innovative. As a social engineer, perhaps less so. For example, he predicted that six computers would be adequate in the future for all the computing needs in the United States.

Contemporary with the work we have been describing are the influential ideas masterminded by Alan Turing (1912–1954) in Great Britain. In 1936 Turing penned a paper called *On computable numbers* in which he described a theoretical device now known as a *Turing machine*.

A *Turing machine* is a device for performing effectively computable operations. It consists of a machine through which a bi-infinite paper tape is fed. The tape is divided into an infinite sequence of congruent boxes (Figure 6.4). Each box has either a numeral 0 or a numeral 1 in it. The Turing machine has finitely many "states" $S_1, S_2, \ldots, S_n$. In any given state of the Turing machine, one of the boxes is being scanned.

After scanning the designated box, the Turing machine does one of three

things:

(1) It either erases the numeral 1 that appears in the scanned box and replaces it with a 0, or it erases the numeral 0 that appears in the scanned box and replaces it with a 1, or it leaves the box unchanged.

(2) It moves the tape one box (or one unit) to the left or to the right.

(3) It goes from its current state S_j into a new state S_k.

Turing machines have become a model for anything that can be effectively computed. The theory of recursive functions and the famous "Church's thesis" from formal logic both have interesting and compelling interpretations in terms of Turing machines.

It is safe to say that, by 1945, Alan Turing had all the key ideas for what would become the modern stored program computer. He knew that there could be a single machine that could perform all possible computing tasks, and that the key to this functionality was the program stored in memory. Turing was virtually unique in that he understood the mathematical theory behind computing but he also had hands-on experience with large-scale electronics.

The next American breakthrough in the computer race came in 1946 from John W. Mauchly (1907 C.E.–1980 C.E.) and J. Presper Eckert (1919 C.E.–1995 C.E.) at the University of Pennsylvania. Known as the *ENIAC*, their computer used a record 18,000 vacuum tubes. Of course vacuum tubes (which have now been superseded by solid-state devices) generate a good deal of heat, and the cooling machinery for the ENIAC occupied 1800 square feet of floor space. The computer itself weighted 30 tons.

The ENIAC had punch card input and was able to execute complex instructions. The machine had to be reconfigured for each new operation. Nevertheless, it is considered by experts to have been the first successful high-speed electronic digital computer. It was used productively for scientific calculation from 1946 to 1955. Eckert and Mauchly went on to form a computer company in Philadelphia that produced, among other innovations, the UNIVAC computer. This was the first commercially available computer in the United States. It is curious that a patent infringement case was filed in 1973 (Sperry Rand vs. Honeywell) that voided the ENIAC patent as derivative of John Atanasoff's invention.

The year 1947 saw another computing innovation from Tom Kilburn (1921 C.E.–2001 C.E.) and Frederic Williams (1911 C.E.–1977 C.E.) of the University of Manchester in England. They succeeded in developing a technology for storing 2048 bits of information (a *bit* is a unit of computer information) on a CRT. They built a computer around this device and it became known as the Williams Tube. By 1948 the Tube had evolved into a new machine called The Baby. Its innovative feature was that it could read and reset at high speed random bits of information, while preserving a bit's value indefinitely between resettings. And this was the first computer ever to be able to hold any (small) program in electronic storage and process it at electronic speeds. Kilburn and Williams continued to develop their ideas, and this work led to the Manchester Mark 1 in 1949. In fact that computer contained the first random access memory (analogous to the sort of memory used in computers today). The point is that data was not stored in sequence—as on a tape. Instead it was stored in such a way that it could be electronically "grabbed" at high speed.

John von Neumann (1903 C.E.–1957 C.E.) had a chance encounter in 1944 with Herman Goldstine (1913 C.E.–) at a train station in Aberdeen, Maryland. There von Neumann learned of the EDVAC (Electronic Discrete Variable Automatic Computer) project at the University of Pennsylvania. He became involved in the project, and in 1945 produced a paper that would change the course of the development of computer science. Until then, computers (such as the ENIAC) were such that they had to be physically reconfigured for each new task. Von Neumann's vision was that the instructions to the computer could be stored *inside the machine*, and he was the first to conceive of a *stored program computer* as we think of it today. Ultimately von Neumann and Goldstine moved back to the Institute for Advanced Study (IAS) in Princeton, New Jersey where they developed the IAS Computer which implemented von Neumann's ideas. Many consider von Neumann (a mathematician by pedigree) to be the father of the modern computer.

An interesting byroad in the modern history of computing is the supercomputer. Seymour Cray (1925 C.E.–1996 C.E.) was arguably the godfather of supercomputing. Supercomputers, in Cray's vision, are founded on the notion of "parallel processing". The idea is that every computer contains a central processing unit (CPU). This is the chip in which all the calculations—the manipulations of 0s and 1s—take place. A supercomputer will have several CPUs—sometimes more than 100 of them—and the different computing tasks will be parceled out among them. Thus many different calculations can

be taking place at the same time. The consequence is that the end result is reached more quickly. In the early days of supercomputing, the user had to learn a special programming language that had routines for doing this division of labor among the different processors. In today's supercomputers, the parceling is done automatically, just by parsing the standard code that the user inputs. But one upshot of this fact is that modern computer languages, like `C++`, cannot be used with a supercomputer, because they are too structured. So many supercomputer users today use the otherwise antiquated scientific computing language `Fortran`.

For many years, the fastest parallel processing machines came out of Seymour Cray's plant outside of Minneapolis. And many of the hottest new high tech companies were spin-offs from Cray. For quite a time, the speed of a Cray was the benchmark for the computing world. As a simple example, when the first Pentium chips came out people crowed that "now we have a Cray I on a chip." That chip ran at 100 megaflops.[1]

Seymour Cray was a special and enigmatic man. When he had a deep and difficult problem to think about, he would go underground. Literally. For many years, Cray worked on digging—with a hand shovel—a $4' \times 8'$ tunnel that was to connect his house with a nearby lake. Cray said he always did his best thinking while he was shoveling away at his tunnel. He died in an automobile accident in Colorado before the tunnel was completed.

Today, with the advent of the *personal computer*, computing has become a major feature of life for all of us. Steve Jobs (1955–) and Steve Wozniak (1950–) invented the personal computer in 1977.[2] IBM revolutionized the

[1] A *flop* is, in computer jargon, a "floating point operation". This is one elementary arithmetic calculation, like addition. One *megaflop* is one million floating point operations. So a 100 megaflop machine can perform 100 million elementary arithmetic operations per second.

[2] It should be noted, however, that a version of the personal computer was produced earlier at the Xerox PARC research facility in Palo Alto. The Xerox PARC personal computer was never a commercial product. Xerox PARC also invented the mouse, the laser printer, and the graphics user interface that has today developed into the operating system `Windows`.

The operating system called Microsoft `Windows` is compiled from 100 million lines of computer code. This is one of the most remarkable engineering feats in history. When the merits of President Ronald Reagan's *Star Wars* program were being debated, it was pointed out that the control system for the project would involve tens of thousands of lines of computer code. At that time such complexity was considered to be infeasible. We have clearly evolved far beyond that state of sophistication. The verification of the reliability of something as complex as `Windows` is tantamount to proving a mathematical theorem.

personal computer industry in 1981 with the introduction of the first PC. This machine used an operating system that was developed by Bill Gates at Microsoft.

6.2 The Difference Between Mathematics and Computer Science

When the average person learns that someone is a mathematician, he or she often supposes that that person works on computers all day. This conclusion is both true and false.

Computers are a pervasive aspect of all parts of modern life. As we learned in the last section, the father of modern computer design was John von Neumann, a mathematician. He worked with Herman Goldstine, also a mathematician. Today most every mathematician uses a computer to do e-mail, to typeset his or her papers and books, and to post material on the WorldWide Web. A significant number (but well less than half) of mathematicians use the computer to conduct *experiments*. They calculate numerical solutions of differential equations, they calculate propagation of data for dynamical systems and differential equations, they perform operations research, they engage in the examination of questions from control theory, and many other activities as well. But the vast majority of (academic) mathematicians still, in the end, pick up a pen and write down a *proof*. And that is what they publish.

The design of the modern computer is based on mathematical ideas—the Turing machine, coding theory, queuing theory, binary numbers and operations, high-level languages, and so forth. Certainly operating systems, high-level computing languages (like `Fortran`, `C++`, `Java`, etc.), central processing unit (CPU) design, memory chip design, bus design, memory management, and many other components of the computer world are mathematics-driven. The computer world is an effective and important implementation of the mathematical *theory* that we have been developing for 2500 years. But the computer *is not mathematics*. It is a device for manipulating data.

Still and all, exciting new ideas have come about that have altered the

---

Indeed, T. Ball at Microsoft has developed theorem-proving and model-checking software that is used to verify portions of the `Windows` operating system. Just to check 30 functions of the Parallel Port device driver, the proving software was invoked 487,716 times.

way that mathematics is practiced. We have seen that the earliest computers could do little more than arithmetic. Slowly, over time, the idea developed that the computer could carry out *routines*. Ultimately, because of work of John von Neumann, the idea of the stored-program computer was developed. In the 1960s, a group at MIT developed the idea that a computer could perform high-level algebra and geometry and calculus computations. Their product was called `Macsyma`. It could only run on a very powerful computer, and its programming language was very complex and difficult.

Today, thanks to Stephen Wolfram (1959–)[3] and the Maple group at the University of Waterloo[4] and the MathWorks group in Natick, Massachusetts,[5], and many others, we have *computer algebra systems*. A computer algebra system is a high-level computer language that can do calculus, solve differential equations, perform elaborate algebraic manipulations, graph very complicated functions, and perform a vast array of sophisticated mathematical operations. And these software products will run on a personal computer! A great many mathematicians and engineers and other mathematical scientists conduct high-level research using these software products. Many Ph.D. theses present results that are based on explorations using `Mathematica` or `Maple` or `MatLab`. Important new discoveries have come about because of these new tools. Stephen Wolfram used `Mathematica` to perform most of the calculations that went into the development of his new theory of the universe that is presented in [WOL]—see Section 9.4.

There is no doubt that computers are now an important part of the mathematical life. They are not mathematics as it has traditionally been practiced. But they are a *part* of mathematics. And computers can be used to actually *search for new theorems* and *search for new proofs*. We shall explore some of these new developments, of the interaction of mathematics with computers, in the succeeding sections of this chapter.

[3]His famous product is `Mathematica`.
[4]Their famous product is `Maple`.
[5]Their famous product is `MatLab`.

6.3 How a Computer Can Search a Set of Axioms for the Statement and Proof of a New Theorem

Michael Atiyah [ATI2] once again gives us food for thought:

> Much of mathematics was either initiated in response to external problems or has subsequently found unexpected applications in the real world. This whole linkage between mathematics and science has an appeal of its own, where the criteria must include both the attractiveness of the mathematical theory and the importance of the applications. As the current story of the interaction between geometry and physics shows, the feedback from science to mathematics can be extremely profitable, and this is something I find doubly satisfying. Not only can we mathematicians be useful, but we can create works of art at the same time, partly inspired by the outside world.

With modern, high-level computing languages, it is possible to program into a computer the definitions and axioms of a logical system. And by this we do not simply mean the *words* with which the ideas are conveyed. In fact the machine is given information about how the ideas fit together, what implies what, what are the allowable rules of logic, and so forth. The programming language (such as `Otter`) has a special syntax for entering all this information. Equipped with this data, the computer can then search for valid chains of reasoning (following the hardwired rules of logic, and using only the axioms that have been programmed in) leading to new, valid statements—or *theorems*.

This theorem-proving software can run in two modes: **(i)** interactive mode, in which the machine halts periodically so that the user can input further instructions, and **(ii)** batch mode, in which the machine runs through the entire task and presents a result at the end. In either mode, the purpose is for the computer to find a new mathematical truth and create a logical chain of thought that leads to it.

Some branches of mathematics, such as real analysis, are rather synthetic. Real analysis involves estimates and subtle reasoning that does not derive directly from the twelve axioms in the subject. Thus this area does not lend

itself well to computer proofs, and computer proofs have pretty well passed this area by.[6]

Other parts of mathematics are more formalistic. There is still insight and deep thought, but many results can be obtained by fitting the ideas and definitions and axioms together in just the right way. The computer can try millions of combinations in just a few minutes, and its chance of finding something that no human being has ever looked at is pretty good. The Robbins conjecture, discussed below, is a vivid example of such a discovery.

There still remain aesthetic questions. After the computer has discovered a new "mathematical truth"—complete with a proof—then some human being or group of human beings will have to examine it and determine its significance. Is it interesting? Is it useful? How does it fit into the context of the subject? What new doors does it open?

One would also wish that the computer reveal its chain of reasoning so that it can be recorded and verified and analyzed by a human being. In mathematics, we are not simply after the result. Our ultimate goal is *understanding*. So we want to see and learn and understand the *proof*.

One of the triumphs of the art of computer proofs is in the subject area of Boolean algebra. Created by George Boole in the late nineteenth century, Boolean algebra is a mathematical theory of switching and circuits. In one standard formulation, the theory has just five definitions and ten axioms. They are these:

> The primitive elements of Boolean algebra are a binary operation $\cup$ (which we can think of as "union" or "combination") and a binary operation $\cap$ (which we can think of as "intersection"). There is also a unary function symbol $\overline{}$ which denotes complementation. The axioms of a Boolean algebra are these:

$(\boldsymbol{B_1})$	$x \cup (y \cup z) = (x \cup y) \cup z$	$(\widetilde{\boldsymbol{B}}_{\boldsymbol{1}})$	$x \cap (y \cap z) = (x \cap y) \cap z$
$(\boldsymbol{B_2})$	$x \cup y = y \cup x$	$(\widetilde{\boldsymbol{B}}_{\boldsymbol{2}})$	$x \cap y = y \cap x$
$(\boldsymbol{B_3})$	$x \cup (x \cap y) = x$	$(\widetilde{\boldsymbol{B}}_{\boldsymbol{3}})$	$x \cap (x \cup y) = x$
$(\boldsymbol{B_4})$	$x \cap (y \cup z) = (x \cap y) \cup (x \cap z)$	$(\widetilde{\boldsymbol{B}}_{\boldsymbol{4}})$	$x \cup (y \cap z) = (x \cup y) \cap (x \cup z)$

[6]Though it must be acknowledged that there are *areas* of analysis that are quite algebraic. The theory of $\mathcal{D}$-modules, the theory of von Neumann algebras, and the theory of Lie groups can be quite algebraic. And certainly computers have been put to good use in all three subjects.

$(\boldsymbol{B_5})$ $x \cup \overline{x} = 1$ $\qquad\qquad$ $(\widetilde{\boldsymbol{B}_5})$ $x \cap \overline{x} = 0$

In the 1930s, Herbert Robbins conjectured that these ten axioms were in fact implied by just three rather simple axioms:

$(\boldsymbol{R_1})$ $x \cup (y \cup z) = (x \cup y) \cup z$ (associativity)

$(\boldsymbol{R_2})$ $x \cup y = y \cup x$ (commutativity)

$(\boldsymbol{R_3})$ $\overline{\overline{x \cup y} \cup \overline{x \cup \overline{y}}} = x$ (Robbins equation)

It is not difficult to show that every Boolean algebra, according to the original ten axioms $(\boldsymbol{B_1})$–$(\boldsymbol{B_5})$ and $(\widetilde{\boldsymbol{B}_1})$–$(\widetilde{\boldsymbol{B}_5})$, is also a Robbins algebra (i.e., satisfies the three new axioms $(\boldsymbol{R_1})$–$(\boldsymbol{R_3})$). But it was an open question for sixty year whether every Robbins algebra is a Boolean algebra. The question was finally answered in the affirmative by William McCune of the Argonne National Laboratory using software (developed at Argonne) called EQP (for "Equational Theorem Prover". The computer was simply able to fit the Robbins axioms together in millions of ways, following the strict rules of logic of course, and find one that yields the ten axioms of Boolean algebra. It is important to note that, after the computer found a proof of the Robbins conjecture, Allen L. Mann [MAN1] produced a proof on paper that a human being can read.

Computers have been used effectively to find new theorems in projective geometry and other classical parts of mathematics. Even some new theorems in Euclidean geometry have been found (see [CHO]). Results in algebra have been obtained by Stickel [STI]. New theorems have also been found in set theory, lattice theory, and ring theory. One could argue that the reason these results were never found by a human being is that no human being would have been interested in them. Only time can judge that question. But certainly the positive resolution of the Robbins conjecture is of great interest for theoretical computer science and logic.

One of the real pioneers in the search for computer proofs is Larry Wos of Argonne National Labs. He is a master at finding new proofs of unproved theorems, and also of finding much shorter proofs of known theorems. He does all this on the computer, often using the software `Otter` by W. McCune. One of Wos's recent coups is the computer-generated solution of the SCB problem in equivalential calculus. Wos's work is well archived in various books, including [WOS1] and [WOS2].

6.4 How the Computer Generates the Proof of a New Result

There are now dedicated pieces of software that can search an axiom system for new results. For example, the program `Otter` is an automated reasoning program. You can tell it the problem you want it to think about and what type of reasoning to employ and then off it goes. It has been used successfully to analyze a number of different situations in many different fields of mathematics.

As indicated in the last section, the Robbins conjecture in Boolean algebra is an outstanding example of an important new fact that was discovered with a computer search. It should be understood that the axioms of Boolean algebra are crisp and neat and fit together like building blocks. This is a logical system that lends itself well to this new technology of computer-aided theorem proving. Subject areas like real analysis, or partial differential equations, or low-dimensional topology—which use a great deal of synthetic reasoning and *ad hoc* argument—would not (at least with our current level of knowledge in computer proofs) lend themselves well to computer searches for proofs. There are those, such as Doron Zeilberger [ZEI], who predict that in 100 years *all* proofs will be computer-generated. Given the current state of the art, and the evidence that we have at hand, this seems like wishful thinking at best.

It must be recorded that the fact that a computer can perform a high-level task like searching for a new truth and then searching for its proof is really quite impressive. Consider that computers were first constructed—about sixty years ago—to do calculations of the weather and of artillery data. All the calculations were numerical, and they were quite elementary. The point of computers in those days was *not* that the computer could do anything that a person could not do. Rather, it was that the computer was faster and more accurate.

There is an amusing story that may tend to gainsay the point made in the last paragraph. As noted elsewhere in this book, John von Neumann conceived and developed the first stored-program computer. He had a whole team of people, including Herman Goldstine, working with him on this project.

John von Neumann was an amazing mathematician and an amazing calculator. He also had a photographic memory: he could effortlessly recite

long passages from novels that he had read twenty years before. He of course played an instrumental role in developing one of the first stored-program computers at the Institute for Advanced Study in Princeton. In those days, von Neumann was extremely active as a consultant. He was constantly coming and going, all over the country, to government agencies and companies, disseminating the benefit of his erudition. It was said that his income from these activities (on top of his princely Institute salary) was quite substantial. And he was already rather wealthy with family money.

During one of von Neumann's consulting trips, Herman Goldstine and the others working on the new computer got it up and running for a test. They fed it a large amount of data from meteorological observations, ran it all night, and came up with very interesting conclusions in the morning. Later that day, von Neumann returned from his trip. Wanting to pull a prank on Johnny von Neumann, they decided not to tell him that they had the computer up and running, but instead to present their results as though they had obtained them by hand. At tea, they told von Neumann that they had been working on such and such a problem, with thus and so data, and in the first case had come up with .... "No, no," said von Neumann. He put his hand to his forehead, threw his head back, and in a few moments gave them the answer. It was the same answer that the machine had generated. Then they said, "Well, in the second case we got .... "No, no," said von Neumann. "Let me think." He threw his head back—it took longer this time—but after several moments he came up with the answer. Finally his collaborators said, "Now in the third case ...." Again, von Neumann insisted on doing the calculation himself. He threw his head back and thought and thought and thought. After several minutes he was still thinking and they blurted out the answer. John von Neumann came out of his trance and said, "Yes, that's it. How did you get there before I did?"

Another story of von Neumann's calculating prowess goes like this. He was at a party, and someone asked him the following chestnut:

> Two trains are fifty miles apart. They travel toward each other, on the same track, at a rate of 25 miles per hour. A fly begins on the nose of one train, travels at a rate of 40 miles per hour to the nose of the other train, and then turns around and flies back to the first train. The fly travels back and forth between the trains until it is crushed between the trains. How far does the fly travel in total?

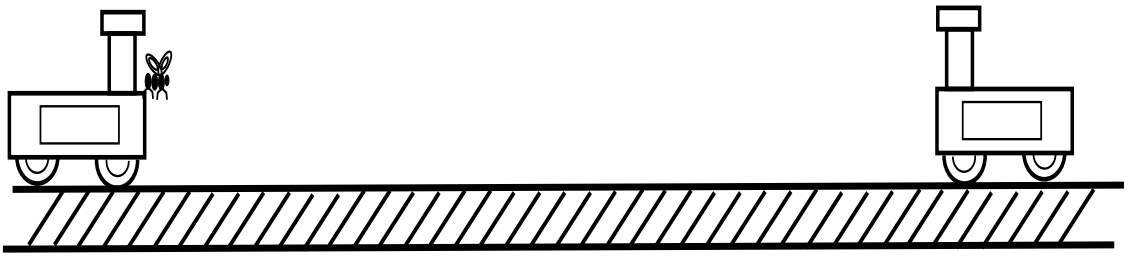


Figure 6.5

Johnny von Neumann instantly blurted out the correct answer of 40 miles. His friends said, "Oh, you saw the trick." [The trick is to notice that it takes the trains one hour to meet. Hence the fly travels for one hour, or 40 miles.] But von Neumann said, "No, I just added up the infinite series."

It is fun—if extremely tedious—to note what that series actually is. First observe that if the situation is as in Figure 6.5, with the fly beginning on the lefthand train (located at the origin), then we can calculate when the fly first meets the righthand train with the equation

$$40t = 50 - 25t\,.$$

Thus the first flight of the fly, from the lefthand train to the righthand train, has duration $t = 50/65$. The second flight of the fly is found by solving

$$\frac{50 \cdot 40}{65} - 40t = \frac{50}{65} \cdot 25 + 25t$$

hence that second flight has duration $t = (50 \cdot 15)/65^2$. Continuing in this manner, we end up with the infinite series of times

$$T = \frac{50}{65} + \frac{50}{65^2} \cdot 15 + \frac{50}{65^3} \cdot 15^2 + \frac{50}{65^4} \cdot 15^3 + \cdots = \frac{50}{65}\left(1 + \frac{15}{65} + \left(\frac{15}{65}\right)^2 + \left(\frac{15}{65}\right)^3 + \cdots\right).$$

Now one may use standard techniques to sum the infinite series in the large parentheses and find that the answer is 65/50. Thus the total elapsed time before the fly is crushed is 1 hour. In conclusion, the fly travels a total of 40 miles.

Today even a modest personal computer runs at a speed of one hundred megaflops. Here a "flop" is a floating point operation, or one basic manipulation of arithmetic. So the computer is performing 100 million basic arithmetic

operations per second. Clearly no man—not even John von Neumann—can match that speed.

In the next chapter we begin to explore new ways that computers are being used in communication, in teaching, and in proof. There is no denying that computers have changed many aspects of our lives, and they continue to do so.

Chapter 7

The Computer as an Aid to Teaching and a Substitute for Proof

Light; or, failing that, lightning: the world can take its choice.

Thomas Carlyle

Things and men have always a certain sense, a certain side by which they must be got hold of if one wants to obtain a solid grasp and a perfect command.

Joseph Conrad

In what we really understand, we reason but little.

William Hazlitt

It takes a long time to understand nothing.

Edward Dahlberg

Be sure of it; give me the ocular proof.

William Shakespeare

The computer has already started doing to mathematics what the telescope and microscope did to astronomy and biology. In the future not all mathematicians will care about absolute certainty, since there will be so many exciting new facts to discover: mathematical pulsars and quasars that will make the Mandelbrot set seem like a mere Galilean moon. We will have (both human and machine) professional theoretical mathematicians, who will develop conceptual paradigms to make sense out of the empirical data and who will reap Fields medals along with (human and machine) experimental mathematicians. Will there still be a place for mathematical mathematicians?

Doron Zeilberger

Experimental mathematics is that branch of mathematics that concerns itself ultimately with codification and transmission of insights within the mathematical community through the use of experimental exploration of conjec-

tures and more informal beliefs and a careful analysis of the data acquired in this pursuit.

J. Borwein, P. Borwein, R. Girgensohn, and S. Parnes

When a computer program applies logical reasoning so effectively that the program yields proofs that are published in mathematics and in logic journals, an important landmark has been reached. That landmark has been reached by various automated reasoning programs. Their use has led to answers to open questions from fields that include group theory, combinatory logic, finite semigroup theory, Robbins algebra, propositional calculus, and equivalential calculus.

Keith Devlin

7.1 Geometer's Sketchpad

`Geometer's Sketchpad` is a learning tool, marketed by Key Curriculum Press, designed for teaching Euclidean geometry to high school students. There has been a great trend in the past twenty-five years to reverse-engineer high school geometry so that proofs are de-emphasized and empiricism and speculation more highly developed. `Geometer's Sketchpad` fits in very nicely with this program.

`Geometer's Sketchpad` enables the user to draw squares and triangles and circles and other artifacts of classical geometry and to fit them together, dilate them, compare them, measure congruences, and so forth. It is a great way to experiment with geometrical ideas. And, perhaps most important, it is an effective device for generating student interest in learning mathematics. Students today are loth to read the dry text of a traditional mathematical proof. They are much more enthusiastic about jumping in and doing experiments with `Geometer's Sketchpad`. Thus this new software can, in the right hands, be a dynamic and effective teaching tool. And it has been a great success in the marketplace. Entire countries—most recently Thailand—have purchased site licenses for `Geometer's Sketchpad`.

7.2 Mathematica, Maple, and MatLab

The software `MACSYMA`, developed from 1967 to 1982 at MIT by William A. Martin, Carl Engelman, and Joel Moses, was revolutionary because it was the

first *symbol manipulation* package. What does that mean? Prior to `MACSYMA`, what a computer could do was manipulate numbers. The user inputted numerical data and the computer, after crunching for a while, outputted numerical data. `MACSYMA` changed all that by offering the capability to perform *algebraic operations*. `MACSYMA` could solve systems of equations, calculate integrals, invert matrices, solve differential equations, compute eigenvalues, and do a number of other calculations that involve *symbols* rather than numbers. The somewhat daunting feature of `MACSYMA` was that it only ran on a fairly powerful computer, and its programming language was in the nature of an artificial intelligence language. It was difficult to program in `MACSYMA`. But it was all we had for a number of years. It would be quite daunting, for example, to calculate the eigenvalues of a 20×20 matrix by hand; `MACSYMA` can do it in a trice. It would be nearly prohibitive to solve a system of ten ordinary differential equations in ten unknowns by hand; for `MACSYMA` the matter is straightforward. Even though there are a number of software products today that have superseded `MACSYMA`, there are legacy routines—developed for instance to study questions of general relativity—that were written in `MACSYMA`. So people continue to use the product.

In the early 1980s, the nature of symbolic manipulation changed dramatically. MacArthur Prize winner Stephen Wolfram developed a new package called `Mathematica`. This new tool had many advantages over `MACSYMA`:

- `Mathematica` will run on a microcomputer, for example on a PC or a Macintosh.

- `Mathematica` has a very sensible and transparent syntax. Anyone with some mathematical training will find programming in `Mathematica` to be straightforward.

- `Mathematica` is very fast. It can do calculations to any number of decimal places of accuracy very quickly.

- `Mathematica` is a wizard at graphing functions of one or two variables. The user just types in the function—no matter how complicated—and the graph is produced in seconds. Graphs can be viewed from any angle, and they can be rotated in space. [`MACSYMA` does not do graphing at all.]

- `Mathematica` is a whiz at displaying data.

Stephen Wolfram was quite aggressive about marketing his new product, but the fact is that it virtually sold itself. If ever there was a new tool that allowed us to see things that we couldn't see before, and calculate things that we couldn't dream of calculating before, then this was it. The product `Mathematica` was and continues to be rather expensive, and the licensing policies of Wolfram Research (the company that produces `Mathematica`) rather restrictive. But `Mathematica` has sold like hotcakes and Wolfram is quite a wealthy man.

Some competing products have come about that give `Mathematica` a good run for its money. These include Waterloo `Maple` and `MatLab` by MathWorks. Each offers something a little different from `Mathematica`. For example, many people prefer the syntax of `Maple` to that of `Mathematica`. Also `Maple` is deemed to be more reliable, and offers many functions that `Mathematica` does not have. Engineers prefer `MatLab`, just because it is more of a numerical engine than the other two products. And `MatLab` has a `Maple` kernel! So it shares many of the good features of `Maple`.

It happens that `MatLab` is particularly adept at handling complex numbers; `Mathematica` and `Maple` are rather clumsy in that context. It has become fashionable to refer to all these software packages as "computer algebra systems". But they are really much more than that.

There are some remarkable products that are built *atop* some of these computer algebra systems. For example, `Scientific Workplace` by Makichan Software is a very sophisticated word processing system. It is particularly designed for producing mathematics documents. Suppose that you are writing a textbook using `Scientific Workplace`. And you display an example. It turns out that `Scientific Workplace` has a `Maple` kernel, and it will work the example for you! This is a terrific accuracy-checking device. And a great convenience for authors and other working mathematicians.

A great many mathematicians now spend their day with one of these computer algebra systems at their side. Now they can very easily ask rather sophisticated "what if?" questions and get rapid and reliable answers. They can draw complicated graphs with grace and ease. They can create instructional *Notebooks* to use in class: students can benefit from the computing power of *Mathematica* without learning the technical computer language. The Notebook interfaces with the user just using ordinary English together with point-and-click. It is interesting to note that the most popular venue for *Mathematica*—the context in which the most copies have been sold—is business schools. For Professors of Business really like the sophisticated graphing

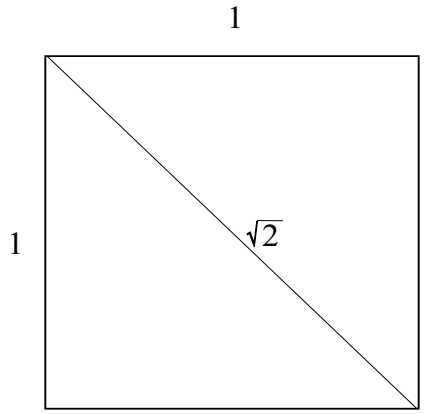


Figure 7.1

capabilities that *Mathematica* offers.

7.3 Numerical Analysis

The idea of numerical analysis has already been illustrated in the discussion of Newton's method in Section 4.5. The philosophy is that there are many problems that come from the real world which we cannot solve precisely—in closed form. Just as an instance, the differential equations that describe an airfoil are too complex for us to solve. Yet we still manage to design very effective airplane wings. How can this be? Mathematicians and aeronautical engineers use a combination of sophisticated guessing together with numerical analysis—which is a very particular way of doing scientific approximation.

Consider for example that we wish to calculate the square root of 2. This is in fact an irrational number (see Section 1.3). Thus its decimal expansion is infinite and nonrepeating. It is of interest to know the value of $\sqrt{2}$, because this is the length of the diagonal of a square of side 1—see Figure 7.1. But we cannot write down $\sqrt{2}$ exactly. We can only write down a decimal approximation. In fact

$$\sqrt{2} = 1.41421356237\ldots \qquad (*)$$

to eleven decimal places of accuracy. This degree of accuracy is adequate for most purpose. After all, it is off by less than 10^{-10}. But perhaps for microchip design or missile guidance or some other exacting task a higher

degree of accuracy would be required. Then we can expand the number further:

$$\sqrt{2} = 1.41421356237309504880168872\ldots \qquad (**)$$

Now we have 25 decimal places of accuracy. We might note that one way to perform this calculation is to apply Newton's method (again see Section 4.5) to the function $x^2 - 2$ with initial guess $x_1 = 1$. After about 6 iterations, one would arrive at the expansion $(**)$.

And, if need be, one could generate even more digits of the decimal expansion for $\sqrt{2}$. It is interesting and important, for theoretical purposes, to know that the decimal expansion of $\sqrt{2}$ is nonterminating and nonrepeating. But for practical applications a suitable approximation, with terminating decimal expansion (as in $(*)$ or $(**)$) is certainly sufficient.

Numerical analysis is the systematic study of this type of scientific approximation. The subject area finds its roots in Newton's method, but today it has developed and expanded into a sophisticated array of computer techniques. Numerical analysis is used in all aspects of engineering and many other parts of applied science.

7.4 Computer Imaging and the Visualization of Proofs

One of the marvelous features of high speed digital computers is that they enable us to see things that we could not before even begin to visualize. We can exhibit complex three-dimensional data sets, graph very sophisticated functions, and use graphics in powerful new ways to display data. As a simple example, recent work by this author and reconstructive surgeon Michael Cedars has created software that will decompose the face into pieces that reflect the subtle geometric structure of the facial surface; this enables the plastic surgeon to see the object of his/her procedures in a new light, and to more effectively plan his/her surgeries.

Today, automobile body design, ship hull design, and many other structural tasks are performed on a computer screen (by contrast, not too many years ago these tasks were performed with clay models). Tool design, fracture analysis, decay of materials, and many other important parts of applied science are studied and analyzed on a computer screen *because the computer will allow us to see things more clearly and in more detail and in new ways.*

As an instance of these last ideas, mathematician Bjorn Dahlberg of Washington University headed up a team of mathematicians, and engineers from Volvo Corporation, to develop software called `SLIP` to design automobile bodies. This was quite revolutionary. The traditional method for designing an automobile body had been this: The artists and designers would discuss the design of a new auto body, and produce various sketches and color renditions—these were paper hard copies. Then a clay model of the auto body would be produced. At that stage the engineers examined the new design and critiqued it: Where would the engine go? Do the passengers have sufficient room? Is it safe? Is it aerodynamic? And so forth. Their comments and criticisms would be conveyed to the artists and designers, and then new sketches and a new clay model would be produced. Then the engineers would have another go at it. This loop would be repeated a great many times until a workable model was created. The entire process took thousands of person hours and many months. In the early 1980s, Volvo decided that it wanted to modernize and automate the process. The company gave the problem to its in-house technical staff for development. Unfortunately they found it too difficult; they could make no progress with the task.

Enter Bjorn Dahlberg. He was a brilliant theoretical mathematician—holder of the prestigious Salem Prize in harmonic analysis. And he took a real shine to this problem. He applied sophisticated techniques from differential equations, differential geometry, linear programming, convex surface theory, harmonic analysis, real variable theory, and other parts of mathematics to help develop the elaborate software package `SLIP` that Volvo now uses every day to design new autos. The key insight of `SLIP` is that an auto body is the union of convex surfaces. The user inputs data points through which the surface will pass, specifies criteria for air resistance, refraction of light, and other desiderata, and then `SLIP` does an elaborate calculation and produces the surface—both graphically and analytically. The entire subject of surface design was revolutionized by `SLIP`.

Similarly, computer visualization can be an important tool in proofs. Hoffman, Hoffman, and Meeks [HHM] studied embedded minimal surfaces in three-dimensional space by first generating pictorial examples using methods of numerical analysis and then, after analyzing these examples and determining various patterns and relationships, writing down a traditional theorem and proving it in a traditional manner. The point here is that the computer enabled these scientists to *see things* that they otherwise could not. It enabled them to envision certain important examples that pointed to new

directions in the field. Then, equipped with new understanding and insight, they were able to write down a theorem—in the traditional manner—and prove it—also in the traditional manner.

The notion of using the computer to develop visualization techniques is a relatively new one. For computer graphic calculations are "computationally expensive"—meaning that they require considerable processing power, considerable memory, and considerable hard disk space. Not to mention a good chunk of computing time. Even twenty-five years ago, one was at the mercy of the computer system at the university. One had to compete with computer scientists and engineers and other high-profile users for a time slice on the mainframe (or perhaps the mini-computer in later years) so that the necessary calculations could be performed. Some mathematicians developed the habit of going into work at 2:00am on Saturday night in order to be able to do their work. Now we have desktop microcomputers that perform at the speed of some of the early supercomputers. My present notebook computers is faster than the Cray I, one of the benchmark machines produced by Seymour Cray in the late 1960s and early 1970s. It operates at greater than 100 megaflops—meaning that it can perform more than 100 million arithmetic operations per second. It is only recently that such powerful tools have become readily available and relatively inexpensive.

7.5 Mathematical Communication

In the ancient world the great mathematicians had schools. Euclid, for example, ran a large and powerful school in Alexandria. Archimedes had a school. In this way the powerful scholar could disseminate his ideas, and also curry young people to help him in his work.

But there was not a great deal of communication among mathematicians. Travel was difficult, and there was no postal system. In many ways mathematicians worked in isolation. An extreme illustration of the lack of communication is that the ancient Chinese anticipated the Pythagorean theorem and many other results that we have commonly attributed to western mathematicians. Yet this fact was not learned until modern times. In fact the Chinese were rather sophisticated mathematicians in a number of ways that are still not known in the West. In the third century C.E., Liu Hui managed to inscribe a 3072-sided polygon in a circle, and thereby to calculate the number π to five decimal places. In the fifth century C.E., the father-son

team of Tsu Ch'ung-Chih and Tsu Keng-Chih was able to calculate π to *ten* decimal places; the details of their methodology have been lost. But we have their result, and it is correct.

As we have discussed elsewhere in this book, mathematicians of the Renaissance tended to be rather secretive about their work. Many of these scholars worked in isolation, and not at universities. There was a considerable amount of professional jealousy, and little motivation to publish or to share ideas. It was not at all uncommon for a scholar to announce his results, but not reveal his methods. Some such investigators, such as Pierre de Fermat, did so out of a sense of puckishness. Fermat would prove his theorems and then pose them as challenges to his fellow mathematicians. Some scholars did it because they did not trust other scientists. Henry Oldenburg pioneered the idea of the refereed scientific journal in 1665, and that really changed the tenor of scientific communication.

In the nineteenth century, at least among European mathematicians, communication blossomed. Weierstrass, Cauchy, Fermat, and many others conducted copious correspondence with scholars at other universities. Many of these letters survive to this day, and they make fascinating reading. In fact Gösta Mittag-Leffler's rather palatial home in Djurshom, Sweden is now a mathematics institute. Conferences and meetings and long-term workshops are held there regularly. What is remarkable about this particular math institute is that it still appears very much as it did when Mittag-Leffler lived in it. Much of his furniture survives, his library is intact, his photo albums and scrapbooks still sit on the shelf, and the papers in his study are all intact. Residing on a shelf in the study are a box with letters from Cauchy, a box with letters from Weierstrass (Mittag-Leffler's teacher), and many others as well. This is a real treasure trove, and fodder for future historians of mathematics.

In the twentieth century the habit of regular correspondence among mathematicians continued to develop. But something new was added. Certainly in the early part of the twentieth century there were few formal mathematics journals and few mathematical book publishers. So it became quite common for people to circulate sets of notes. This came about because a professor would give a course at the university developing some of his new ideas. The consensus often was that these ideas were valuable and should be promulgated. So a secretary was enlisted to type up the notes, copies were made, and one could request a copy simply by sending in a modest amount of money to cover copying costs and postage.

Certainly sets of notes were one of the primary tools of mathematical communication in the 1920s, 1930s, 1940s, and 1950s. One thing that is special about the Princeton University mathematics library is that it has a special room with a magnificent collection of sets of notes from all over the world. Many of these are priceless, and contain extremely valuable information that is not reproduced elsewhere.

On October 4, 1957 the Soviets launched their unmanned satellite Sputnik. This event caught the Americans by total surprise; they suddenly realized that they were behind in the "space race", and behind in the development of technology in general. The government launched a huge push to accelerate our high-tech growth, and this included a major development in American education. As a result, American universities in the 1960s enjoyed an enormous mushrooming and development. Many new campuses were built, and existing campuses were expanded enormously. And there were ripple effects. Mathematical publishing, in particular, really took off. A great many mathematical publishing houses sprang forth, and produced a plethora of mathematics books. Several publishers began special series called "Lecture Notes", and they then published informal manuscripts that formerly would have been circulated privately as mimeographed sets of notes. Mathematical communication truly grew and prospered during this period.

Today of course we have electronic communication and this has changed everything—probably for the better. In our modern world people routinely post sets of notes—created in TEX (see Section 8.5) so that they are typeset to a very high standard—on Web sites. People post their papers, just as soon as they are written, on preprint servers. These are Web sites equipped with a mechanism that makes it easy for the user to `ftp` or upload a TEX or `*.dvi` or `*.pdf` file and then have it posted—for all the world to see—automatically and without human intervention. Furthermore, the end user can download the paper, print it out, edit it, and manipulate it in other ways. He can send it on to others, share it with his/her collaborators, or cut it up and combine it with other documents (although this latter is not encouraged).

Of course mathematicians can also send their papers directly to selected individuals using e-mail; the device of the e-mail *attachment* is particularly useful for this purpose.

A special feature of e-mail is that it can affect the *quality* of communication. An e-mail exchange can escalate in intensity rather rapidly. We have all had the experience of having a polite e-mail exchange become ever more heated until the participants are virtually shouting at each other over the

wires. The trouble with electronic communication is that you don't have the opportunity to look your interlocutor in the eye; you cannot use body language; you do not have room to interpret and to allow for nuance. It is just that e-mail is too hard-edged.

And this feature has consequences for scientific communication as well. An interesting story, coming to us from Harvard mathematician Arthur Jaffe [JAF], illustrates the point:

> In 1982, Simon Donaldson, then finishing his doctoral thesis at Oxford, advocated the study of the space of solutions to the Yang-Mills [field equations] on 4-manifolds as a way to define new invariants of these manifolds. Literally hundreds of papers had been written after Donaldson's initial work in 1983, and an industry of techniques and results developed to study related problems ...
>
> This mathematical focus changed overnight in 1994, following a suggestion by Seiberg and Witten that a simpler approach might be possible. I heard this as a concluding remark in a physics seminar on October 6, 1994, held at MIT ...
>
> So after that physics seminar on October 6, some Harvard and MIT mathematicians who attended the lecture communicated the remark by electronic mail to their friends in Oxford, in California, and in other places. Answers soon began to emerge at breakneck speed. Mathematicians in many different centers gained knowledge and lost their sleep. They reproved Donaldson's major theorems and established new results almost every day and every night. As the work progressed, stories circulated about how young mathematicians, fearful of the collapse of their careers, would stay up night after night in order to announce their latest achievement electronically, perhaps an hour—or even a few minutes—before some competing mathematician elsewhere. This was a race for priority, where sleep and sanity were sacrificed in order to try to keep on top of the deluge of results pouring in. Basically ten years of Donaldson theory were re-established, revised, and extended during the last three weeks of October 1994.

Electronic communication—e-mail and the Worldwide Web—have transformed the development of mathematics in marvelous and profound ways. In fact it has changed the nature of our discourse. For papers that are sent

around by e-mail or posted on a preprint server are not published in the usual fashion.[1] They are not vetted or refereed or reviewed in any manner. There is no filter for quality or appositeness or correctness. There is nobody to check for plagiarism or calumny or libel or slander.[2] There is just an undifferentiated flood of information and pseudoinformation.

And of course the end user must then somehow figure out what is worth reading and what is not. A traditional journal of high quality conducts a serious filtering operation. It is in their best interest to maintain a high standard and to showcase only the best work. This keeps their clientele coming back for more, and makes them willing to pay the rather high tariff for a modern scientific journal.[3] Of course the scientist looking for literature is likely to gravitate to names that he knows and trusts. It is quite easy to do a `Google` search for a particular author and to find all his works. Likely as not they will all be posted on his Web site, or on a preprint server, and easily downloaded and printed out. But such a system does not serve the tyros well. A mathematician or scientist who is just starting out will not be well known and it is rather unlikely that some bigwig at Harvard is going to do a Web search looking for his work.

For many mathematicians, posting work on the Web has replaced dealing with traditional publishers who purvey traditional journals. When you post on the Web, you don't need to deal with pompous editors and surly referees. You just post and then go on about your business. You get feedback from interested readers, but that will mostly be benign and welcome. The point here is that the WorldWide Web is magnificent for disseminating work broadly, and virtually for free. It is not good for archiving, as nobody knows how to archive electronic media. Each copy of an electronic product is unstable, and sunspots or cosmic rays or an electromagnetic storm could wipe out every copy on earth in an instant. A new, insidious virus could destroy 95% of the world's computers simultaneously. Certainly mirror sites, disaster backups, the modified-Tower-of-Hanoi protocol, and other devices make our electronic media somewhat secure. But there is a long ways to go. Each copy of a hard copy journal is stable: Print a thousand hard copies of the new issue of your journal and distribute them to libraries all over the world and you can be reasonably confident that at least some of these will survive for several

[1] They could eventually be published in a refereed journal, but at first they are not.

[2] But `Google` *can* be used to check for plagiarism, and teachers commonly do so.

[3] This could be as high as $500 or $1000 per annum or much higher. One brain science journal these days costs $23,000 per year.

hundred years.

And while there are electronic journals that conduct reviewing and refereeing in the traditional manner, and to a very high standard, they are in the minority. Many electronic journals are a free-for-all, with the expected consequences and side effects. And there are also side effects on paper journals. Editors and referees, taken as a whole, are overworked. There are far too many papers.[4]

Electronic dissemination of scientific work raises a number of important side issues. We again borrow from Arthur Jaffe [JAF] as we lay out what some of these are (see also the book [KRA3] for useful information):

- How can we maintain standards of quality when the volume of publication is theoretically unlimited? For example, how can we continue to ensure serious refereeing, in spite of an enormous increase of volume brought about by the ease of word processing? This phenomenon will overload a system already near the breaking point. In fact there are already so many publications that careful refereeing is hard to find.

- How can we organize publication so we can find what we need? Casual browsing of the web illustrates this problem.

- How can we establish the priority of ideas and assign credit? Volume makes this question extremely difficult, as do multiple versions of the same work. The possibility of interactive comments makes it even more difficult. Will some people resort to secrecy to ensure priority, while others will make exaggerated claims? Will much work get lost because it was not fashionable at the time?

- How can we prevent fads from overwhelming ground-breaking long-term investigation? To what extent will frequency of citation be taken as a gauge of success?

- How can we archive our work? In other words, how can we assure access to today's publications in 5, 10, 30, or in 100 or 300 years? Libraries have learned to live with paper, and the written word is generally readable. But technology develops rapidly. Formats, languages, and operating systems change. We know that the American Mathematical

[4]The vehicle *Math Reviews* of the American Mathematical Society reviews 75,000 papers per year.

Society has a roomful of 10 year old computer tape, unusable because the computer operating system has changed. How can we be assured that our mathematical culture will not disappear because a change in programs, a change in media, or another change in technology makes it too expensive to access a small number of older works?

Chapter 8

Beyond Computers: The Sociology of Mathematical Proof

It's one of those problems [the Kepler sphere-packing problem] that tells us that we are not as smart as we think we are.

D. J. Muder

If a whole chain of boring identities would turn out to imply an interesting one, we might be tempted to redeem all these intermediate identities; but we would not be able to buy out the whole store, and most identities would have to stay unclaimed.

Doron Zeilberger

The packing will be the tightest possible, so that in no other arrangement could more pellets be stuffed into the same container.

J. Kepler

It became dramatically clear how much proofs depend on the audience. We prove things in a social context and address them to a certain audience.

William P. Thurston

Results discovered experimentally will, in general, lack some of the rigor associated with mathematics, but will provide general insights into mathematical problems to guide further exploration, either experimental or traditional. Conjectures experimentally verified will give us more confidence in our direction, even when strongly held beliefs elude proof. One can hope to produce an intuitive view of mathematics that can be transferred in concrete examples and analysis, as opposed to the current system where intuitions can be transmitted only from person to person.

J. Borwein, P. Borwein, R. Girgensohn, and S. Parnes

Whether the turn of the century will find many open questions under attack by a team consisting of a researcher and an automated reasoning program is left to time to settle. Clearly, the preceding decade has witnessed a sharp increase in this regard, sometimes culminating in the answer to a question that had remained open for decades. Thus we have evidence that an automated reasoning program can and occasionally does contribute to mathematics and logic.

Keith Devlin

This suggests to me that sometimes to prove more one must assume more, in other words, that sometimes one must put more in to get more out. . . . Euclid declared that an axiom is a self-evident truth, but physicists are willing to assume new principles like the Schrödinger equation that are not self-evident because they are extremely useful.

G. J. Chaitin

The paper distorts the relation of experiment to theoretical physics. To paraphrase Fermi (perhaps badly): an experiment which finds the unexpected is a discovery; an experiment which finds the expected is a measurement.

Daniel Friedan

Logistic is not sterile; it engenders antinomies.

Henri Poincaré

8.1 The Classification of the Finite, Simple groups

The idea of "group" came about in the early nineteenth century. A product of the work of Evariste Galois (1812–1832) and Auguste Cauchy (1789–1857), this was one of the first cornerstones of what we now think of as *abstract algebra*. So what is a group?

The idea is simplicity itself. A *group* is a set (or a collection of objects) G equipped with a binary operation that satisfies three axioms. Now we are seeing mathematics in action. In particular, we are encountering definitions and axioms.

What is a "binary operation"? This is a way of combining two elements of G to produce a new element. For instance, if our set is the whole numbers (or integers), then ordinary addition is a binary operation. For addition gives us a way to combine two whole numbers in order to produce another whole

number. As an example,

$$2 + 3 = 5\,.$$

We combine 2 and 3 to produce 5. Multiplication is another binary operation. An example of multiplication is

$$2 \times 3 = 6\,.$$

We combine 2 and 3 to produce 6.

These last two examples are rather pedestrian. What is a more exotic example of a binary operation? Consider 2×2 matrices. We multiply two such matrices according to this rule:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} = \begin{pmatrix} a\alpha + b\gamma & a\beta + b\delta \\ c\alpha + d\gamma & c\beta + d\delta \end{pmatrix}\,.$$

We see that this gives a way of combining two 2×2 matrices to give another 2×2 matrix.

So the group G has a binary operation, as described above, which we denote by $\cdot$ (we do this as a formal convention, whether the operation in any specific instance may turn out to be addition or multiplication or some other mode of combining elements). And the axioms that we put in place which govern the behavior of this binary operation are these:

(1) The binary operation is associative:

$$a \cdot (b \cdot c) = (a \cdot b) \cdot c\,.$$

(2) There is a special element $e \in G$, called the *identity element of the group*, such that $e \cdot g = g \cdot e = g$ for every $g \in G$.

(3) Each element $g \in G$ has an *inverse element*, called g^{-1}, such that

$$g \cdot g^{-1} = g^{-1} \cdot g = e\,.$$

It turns out that groups arise in all aspects of mathematics, physics, and even engineering. Some examples are these:

- The integers (i.e., the positive and negative whole numbers), equipped with the binary operation of addition, form a group. Of course integer addition is associative: We know that $a + (b + c) = (a + b) + c$. The identity element e is just 0, for $0 + a = a + 0 = a$ for all elements a in the integers. And finally the additive inverse for any integer a is $-a$.

- The positive real numbers (i.e., all positive whole numbers, all positive rational numbers, and all positive irrational numbers), equipped with the binary operation of multiplication, form a group. In this context associativity is a standard arithmetical fact. The identity element is just 1. And the multiplicative inverse of a is $1/a$.

- The 2×2 matrices having nonvanishing determinant, so that

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc \neq 0\,,$$

  and equipped with the binary operation of matrix multiplication (defined above), form a group. Associativity of matrix multiplication is a standard fact from linear algebra. The identity element is the matrix $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. And the inverse of a given matrix is its matrix multiplicative inverse.

- Group theory is used to describe the algebra of bounded operators on a Hilbert space. These in turn, according to a profound idea of Werner Heisenberg (1901–1976), Schrödinger (1887–1961), and John von Neumann, can be used to explain the structure of quantum mechanics.

- Every cell phone is encoded with a copy of the Cayley numbers. This is a special group that finds applications in mathematical physics and that is used to encrypt the information being transmitted by a cell phone.

Since groups are such universal mathematical objects, and since they can be used to describe or control or analyze so many different types of physical phenomena, there is great interest in classifying all the groups which may arise. Great progress has been made in this regard with respect to the *finite, simple groups*. A group is finite if it has just finitely many elements. An elementary example of a finite group is this. Consider the polygon (i.e., the *square*) exhibited in Figure 8.1. It has certain symmetries: **(i)** We can flip it left-to right, **(ii)** We can flip it top to bottom, **(iii)** We can flip it along the main diagonal, **iv)** We can flip it along the minor diagonal, **(v)** We can rotate it through 90° or 180° or 270°. The collection of all these symmetries (see Figure 8.2) forms a group—-where the binary operation is composition

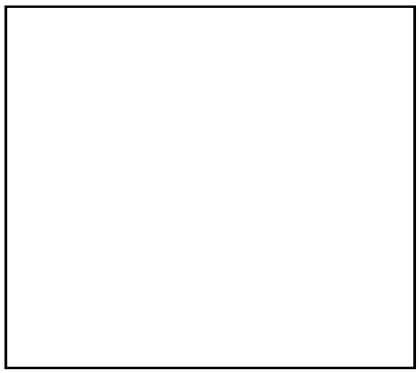


Figure 8.1

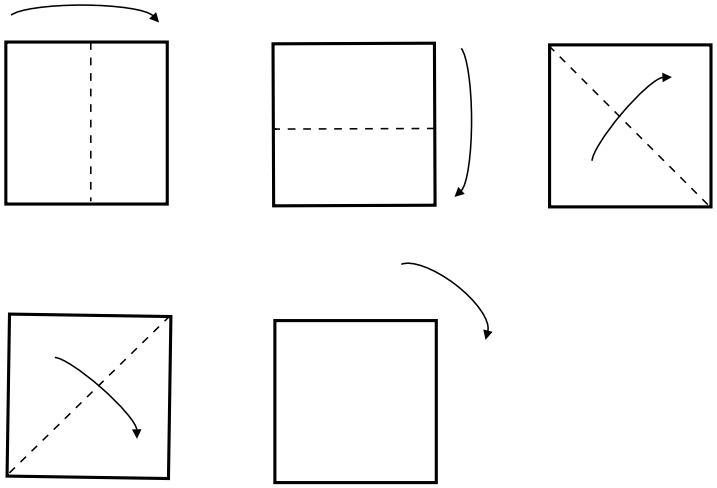


Figure 8.2

(i.e., superposition) of operations. And it is clear that this group has just finitely many elements.

The concept of "simple group" is somewhat more technical. A group is *simple* if it cannot be decomposed into more elementary pieces, each of which is a group. It is a fundamental structure theorem that *any* finite group is made up of simple groups. So the problem of classifying all finite groups reduces, in some sense, to classifying all the finite, simple groups.

The classification of the finite, simple groups is one of the great triumphs of twentieth century mathematics. What is interesting about this achievement—from our point of view—is that this mathematical result is

not the work of any one scientist. It is also not the work of two collaborators, or of a small team working together at a mathematics institute. In point of fact, the classification of the finite, simple groups follows from the aggregate of the work of hundreds of mathematicians in dozens of countries stretching back to the middle of the nineteenth century.

The classification of finite, simple groups comprises more than 10,000 pages of dense mathematical scholarly writing. There exists no published outline of the proof.[1] The proof has existed in some form since 1981. Gaps have been found along the way.[2] The ones that are known have now been fixed.

The nearest thing that we have to a "record" of the classification of the finite, simple groups is the four-volume work *The Classification of the Finite Simple Groups* by Gorenstein, Lyons, and Solomon. It totals 2139 pages of dense mathematical reasoning, and gives a substantial idea of what the program is all about. But Michael Aschbacher, a leading authority in the field, has estimated recently that the *full proof* will be more than 10,000 pages. And bear in mind that that is 10,000 pages of published mathematics in the standard modern argot. Which means that it is pithy and condensed and brief, and leaves a fair amount of work to the reader.

It was Daniel Gorenstein (1923–1992) of Rutgers University who, in the early 1970s, convinced people that the holy grail was in sight. He organized a conference of over 100 experts in the field and took it upon himself to organize them into an army that would attack the considerable task of nailing down this proof. He summarized the state of the art and pointed out what was accomplished and what needed to be done, and to actually convince people to fill in the gaps and prove the results that were still outstanding. He assigned specific tasks to individuals and groups from all over the world. To quote Michael Aschbacher [ASC]:

[1] The book [GLS] contains something that resembles an outline of the classification of the finite, simple groups as it stood at that time (1994). The proof has certainly evolved since then.

[2] There are gaps in mathematics and there are *gaps*. The biggest gap in the finite, simple group program was to classify the "quasithin" groups. This task had been assigned to a young Assistant Professor at a West Coast university. This is what the fellow had studied in his thesis. But it turned out that there were significant *lacunae* in his arguments. These have recently been filled in and rectified by experts Michael Aschbacher and Stephen Smith (see [ASM]). Their two volumes, written to address the questions about the quasithin groups, comprise 1200 pages!

> ... Danny Gorenstein began to speculate on a global strategy for a proof. In effect he called attention to certain subproblems, which appeared to be approachable, or almost approachable, and he put forward a somewhat vague vision of how to attack some of the subproblems, and how his various modules might be assembled into a proof. While his program was sometimes a bit far from what eventually emerged, in other instances he was fairly prescient. In any event, Gorenstein focused on the problem of classifying the finite simple groups, in the process making the effort more visible. He also gave it some structure and served as a clearing house for what had been done, and was being done. In short, Gorenstein managed the community of finite simple groups theorists, and to a lesser extent, managed part of the development of the proof itself.

It finally got to the point, in the late 1970s, that there was just one remaining hole in the program. People suspected that there existed one very large group, with

$$808017424794512875886459904961710757005754368000000000$$

elements, but they could not in fact identify this group. The group held a special fascination, because the prime factorization

$$\begin{aligned} &808017424794512875886459904961710757005754368000000000 \\ &= \ 2^{46} \cdot 3^{20} \cdot 5^9 \cdot 7^6 \cdot 11^2 \cdot 13^3 \cdot 17 \cdot 19 \cdot 23 \cdot 29 \cdot 31 \cdot 41 \cdot 47 \cdot 59 \cdot 71 \end{aligned}$$

also exhibits the atomic weights of the factors in an important molecule. It became dubbed *the monster group*. In 1981, Robert Griess of the University of Michigan was able to construct the monster group (in fact the group can be generated by two 196882×196882 matrices over the group with two elements). And that put the cherry on the top of the program. The classification of the finite, simple groups was (in principle) complete!

The next step, naturally, was for Daniel Gorenstein to arrange for some books to be published that would archive the entire proof, going back to the very beginning in the early nineteenth century and running up to the present day when the final profound ideas were put into place. It was estimated that this work would comprise several volumes of total length in the (many) thousands of pages. To date, this comprehensive work on the classification of the finite, simple groups has not been completed.

But now the experts in finite, simple group theory are working on a "second generation" proof, one that will be more accessible, more coherent, and (one hopes) rather shorter. This mathematical achievement is important to document no matter how many pages it takes, but it is more likely that people can read it, understand it, verify it, and internalize it if it is briefer and more attractively written.

This really is a grand saga, and a tribute to the spirit of cooperation that is prevalent among modern mathematicians. But there have been some stumbles and hiccups along the way. We have noted that there was one mathematician at a West Coast university who was assigned a certain part of the program. He labored away at this for years, produced a manuscript of several hundred pages, but ultimately became discouraged and never finished his task. Unfortunately Gorenstein died and the rest of the world lost track of this little piece of the puzzle. On those rare occasions when, at a conference, somebody said, "But what about that part—the quasithin stuff—that the man at Santa Cruz is supposed to be doing?", the typical answer was, "Don't worry. The program is so robust that it's bound to work out. We have more important concerns at this time." It was finally Fields Medalist Jean-Pierre Serre who wrote a paper pointing out this nontrivial gap in the literature—a gap that must be filled. Serre's article caused quite a stir, and in the end Aschbacher and Smith [ASM] had to work out the ideas and write them up. This took two volumes and 1200 pages. And that was to fill just one gap.

Of course, as we have indicated, there are experts all over the world who spend virtually all their time studying the classification of the finite, simple groups and searching for ways to simplify and clarify the arguments. And it does happen that, every now and then, someone discovers a notable glitch in the big picture. But the program does appear to be stable, the glitch always gets fixed, and there is every reason to believe that eventually there will exist a multi-volume work documenting this theorem that took two centuries to prove.

There are still many more volumes to be written in the saga of the finite simple groups. Part of the problem is that this is a human endeavor, and a great many of the key participants are aging or retiring or dying or some combination of these. It is not a sure bet that Gorenstein's vision will be carried out and validated as he originally anticipated.

8.2 Louis de Branges's Proof of the Bieberbach Conjecture

The Bieberbach Conjecture is one of the grand old questions in complex variable theory. It concerns the nature of the power series coefficients of certain types of analytic functions on the unit disc in the complex plane. More precisely, we have an analytic function

$$f(z) = z + \sum_{j=2}^{\infty} a_j z^j$$

defined on the unit disc $D = \{z \in \mathbb{C} : |z| < 1\}$. We assume that the function takes distinct domain points z_1, z_2 to distinct targets. The conjecture is that it must be that the j^{th} coefficient a_j cannot be any larger (in modulus) than j.

The full details need not concern us here. Suffice it to say that this is a rather technical mathematical question that only a specialist would care about, and Louis de Branges of Purdue University cared about it passionately.

Louis had the reputation of a talented mathematician with a lot of good ideas. But he also had the reputation of something of a crank, because he had cried "wolf" once too often. It should be stressed that all mathematicians make mistakes. Any mathematician worth his/her salt is going to take some risks, work on some hard problems, shoot for the moon. In doing so, he/she may become convinced that he/she has solved a major problem. And thereby mistakes get made. Almost any good mathematician has published a paper with mistakes in it. Many of us have published papers that are just plain wrong. And the referee did not catch the mistake either, so it must have been a pretty good mistake.

There is another famous problem in mathematical analysis called the "invariant subspace problem." A great many people would love to solve this problem—about the nature of bounded operators on a Hilbert space (this is the language in which the modern theory of quantum mechanics is formulated). But nobody has succeeded. Unfortunately, Louis had announced that he could do it, and he fell on his face. Yet another example is de Branges's claimed proof of the Ramanujan conjecture—later in fact proved by Fields Medalist Pierre Deligne.

So de Branges's credibility was somewhat in doubt in 1984 when he announced that he had a solution of the classic Bieberbach conjecture—a problem that went back for its original formulation to 1916. To further confound matters, de Branges claimed not that his proof was in a paper of 30 or 40 or 50 pages that one could just sit down and read. Rather, Louis de Branges declared that the proof was part and parcel of the new edition of his book *Hilbert Spaces of Entire Functions* [DEB1]. Well, there was hardly anyone who was going to sit down and read a 326-page book in order to determine whether Louis had really done it or not. All other things being equal, Louis de Branges's "proof" of the Bieberbach conjecture could easily have moldered away in manuscript form for many years—with nobody having the time or interest to check it.

But Providence intervened on Louis's behalf! In the harsh winter of 1984, de Branges took a sabbatical leave which he spent at the Steklov Mathematics Institute in St. Petersburg, Russia. Now the Russians have a strict and powerful mathematical tradition, and a great determination and work ethic. They spent the entire semester with de Branges distilling his proof of the Bieberbach conjecture from the 300+ page book manuscript. The result is a 16-page paper [DEB2] that appears in *Acta Mathematica* and that anyone can check. Now many hundreds of mathematicians have read de Branges's proof, and confirmed it. There is no doubt that the Bieberbach Conjecture is proved, and that Louis de Branges proved it (with the help of a terrific team of St. Petersburg mathematicians).

But there are even further developments. Louis de Branges engaged in the common mathematical practice of circulating copies of his paper long before its formal publication. Such a script is called a *preprint*. This is a manuscript that is not handwritten—it is a formal typescript. It is an official enunciation of what this mathematician thinks he/she has proved. And Louis sent hundreds of these all over the world. Christian Pommerenke and Carl Fitzgerald were among the recipients of this largesse. And then an unfortunate accident occurred. Namely, Fitzgerald and Pommerenke found a number of interesting and decisive ways to simplify and clarify de Branges's arguments. Their ideas were sufficiently important to justify the publication of a separate paper. They wrote up their ideas and submitted them to the *Transactions of the American Mathematical Society* (see [FIP]). Through a variety of clerical SNAFUs, it came about that the very brief (only 8 pages) Fitzgerald/Pommerenke proof of the Bieberbach conjecture appeared in print *before* the original de Branges proof appeared.

Well, Louis was not happy. Even though Fitzgerald and Pommerenke had been eminently respectful of Louis (they even used his name in their title, and not Bieberbach's!), there is no question that a major screw-up had occurred. At a celebratory conference the purpose of which was to hail Louis as the hero of the hour, Carl Fitzgerald introduced de Branges to speak. Professor de Branges stood up and announced to one and all that Fitzgerald and his collaborator were gangsters who were set to steal his ideas. None of it was very pretty.

And there are further developments. Lenard Weinstein wrote a thesis at Stanford University (his ideas are published in [WEI]) in which he dramatically simplified de Branges's proof. In fact his proof, which begins with the classical Loewner equation, uses nothing more than calculus. It is just four pages. Although not well known, this was a real milestone. But Doron Zeilberger [ZEI] took things a decisive step further. In his paper *A high-school algebra, 'formal calculus', proof of the Bieberbach conjecture [after L. Weinstein]* [EKZ], Zeilberger and Shalosh B. Ekhad provide a proof of the Bieberbach Conjecture that is less than half a page. It consists primarily of verifying a large combinatorial identity, and that verification can in fact be assigned to a computer. It is noteworthy that the original title of the Zeilberger/Ekhad paper was *A wallet-sized, high-school-level proof of the Bieberbach conjecture*. Journal editors, stodgy as they are, will not tolerate a paper with a whimsical title such as this. It could not stand. Instead the paper appears under the indicated title. The reader should find it particularly interesting to note that Shalosh B. Ekhad is *not* a person. In fact this is Zeilberger's nickname for his computer!

8.3 Wu-Yi Hsiang's Solution of the Kepler Sphere-Packing Problem

Together with his brother Wu-Chung Hsiang of Princeton University, Wu-Yi Hsiang of U. C. Berkeley established a pre-eminent reputation in the 1960s as an expert in the theory of group actions on topological spaces. This was a hot field at the time, and earned each of them jobs at the best mathematics departments in the world. At some point late in his career, Wu-Yi's interests began to wander, and they settled on a fascinating problem that found its genesis with the inestimable Sir Walter Raleigh. As we all know,

Raleigh played a key role in the history of the United States. Among other things, he is responsible for our addiction to tobacco. But he was also an adventurer and something of a pirate. One day he questioned one of his gunners about the most efficient way to stack cannonballs. This question turned out to have mathematical interest, and was eventually communicated to the distinguished mathematician and astronomer Johannes Kepler. The classical formulation of Kepler's problem is then, "What is the most efficient way to pack balls of the same size into space?" Kepler published this question in a booklet called *Strena sue de nive sexangula* in 1611. Interesting.

This is in fact a question that grocers face every day. If the produce person in a supermarket wants to display oranges (assuming that all the oranges are about the same size and shape), what is the most efficient means to stack them up and display them? That is to say, how can we fit the most oranges into the least space? This may seem like a frivolous question, but it is not. And its solution is by no means obvious. The analogous question in two dimensions has had a ready solution for a long time, and it is well worth examining. Look first at Figure 8.3. It shows a number of two-dimensional balls, or discs—all having the same radius—packed into the plane. These are displayed in the "rectilinear packing", which is one of the obvious ways to fit discs efficiently into a small space. It turns out that this is not the best way to do it. A sophisticated exercise in matrix theory shows that in fact the *best* way to pack two-dimensional discs into the plane is the *hexagonal packing*, displayed in Figure 8.4. One can calculate that this is a distinctly better method, and in fact show that it is the optimal method. The Kepler sphere-packing problem is to find an analogous answer in three (or higher) dimensions.

It is generally believed that the three-dimensional analogue of the planar hexagonal packing is the optimal packing for spheres in three dimensions (refer to Figure 8.5). This is the "default" manner in which a grocer would stack oranges in a store display, or a gunner would stack his cannon balls. It has density $\alpha = 0.74048$. That is to say, with the described packing, 74.048 percent of space will be occupied by the cannon balls and 25.952 percent of space will be occupied by air. Certainly nature, with the way that she packs atoms into molecules, provides evidence for this belief. But there was, until recently, no proof.

Kepler first formulated his problem in 1611. So this is one of the very oldest unsolved mathematics problems. Carl Friedrich Gauss contributed the first result on Kepler's problem. He showed that the Kepler conjecture

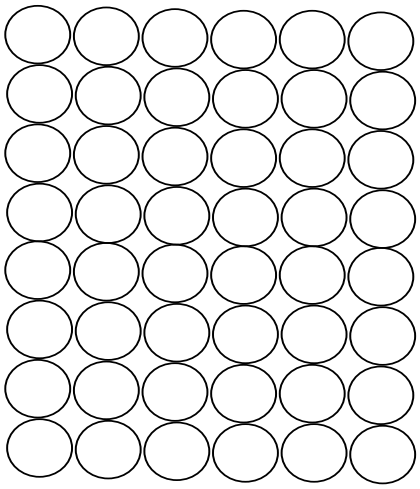


Figure 8.3

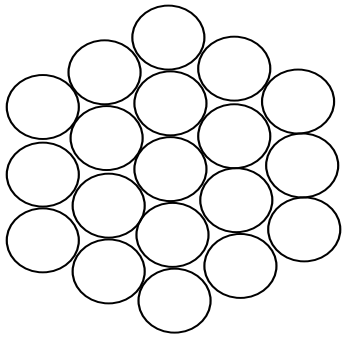


Figure 8.4

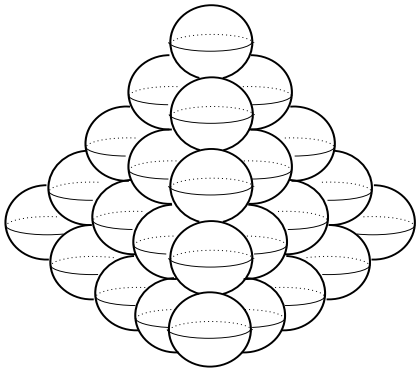


Figure 8.5

is true if the spheres are assumed in advance to be arranged in a regular lattice. Thus, if there were to be a counterexample to the Kepler conjecture, then it would exhibit balls arranged in an irregular fashion. One of the most significant modern attacks on the problem was levied by Laszlo Fejes-Tóth (see [FTO]). In 1953 he demonstrated that the problem could be reduced to a finite (but very large) number of calculations. Architect, geometer, and entrepreneur Buckminster Fuller claimed in 1975 to have a proof of Kepler's conjecture, but this proved to be incorrect.

Interesting ideas about the problem in higher dimensions were offered by John Horton Conway in [CON]. Wu-Yi Hsiang proposed to use a new implementation of spherical trigonometry—in fact he virtually reinvented the subject—to give a resolution of Kepler's problem in dimension three. In 1993 he wrote a long paper [HSI1] (92 pages) together with a secondary paper [HSI2] explaining what was going on in the first paper, laying out his solution of the Kepler sphere-packing problem. This caused quite a sensation, as there had been no serious attack on the problem for a great many years.

But it was not long before the experts began to mount aspersions on Hsiang's efforts. It seems that Hsiang had declared that "Thus and such is the worst possible configuration of balls and we shall content ourselves with examining this configuration." Unfortunately, Hsiang did not provide a convincing reduction to the indicated special case. In fact it is generally believed that problems like this *do not have* a worst-case scenario. One has to come up with arguments that address all cases at once. One of the leaders of the anti-Hsiang movement was Thomas Hales of the University of Michigan (now of the University of Pittsburgh). He published a polite but detailed article [HAL1] in the *Mathematical Intelligencer* taking Hsiang's efforts to task. It should be noted that others, including John Horton Conway of Princeton, had endeavored to get Hsiang to fess up to his errors. Conway published an important book in the subject area and made no reference to Hsiang's work. But it was Hales who bit the bullet and made public statements to the effect that Hsiang was wrong.

Thomas Hales was not just blowing smoke. He had *his own* proof of the Kepler sphere-packing problem which he had intended to publish as a long sequence of papers. He studied the work of L. Fejes-Tóth and concluded that the problem could be solved by minimizing a function of 150 variables. This in turn would entail solving about 100,000 linear programming problems.[3]

[3]Linear programming is a subject developed by George Dantzig in the late 1940s. It is

In 1992, assisted by his graduate student Thomas Ferguson, Hales embarked on a program to prove the celebrated conjecture. They produced a total of six papers along the way. The final piece of the puzzle appears in the paper *A proof of the Kepler conjecture* [HAL1], which has a colorful history.

Hales's capstone paper, in manuscript form, was more than 200 pages long. He submitted it to the *Annals of Mathematics*, which is the "journal of record" for modern mathematics. It is arguably the most prestigious mathematics journal in the world. Being a journal of such high rank, the *Annals* has rather strict rules. Generally speaking, it does not consider exceedingly long papers. And it only wants to consider modern problems of current interest. Finally, it usually will only consider papers that have a traditional mathematical form. That is to say, the *Annals of Mathematics* has been a bastion for the classical form of mathematical proof: tightly knit sequences of statements connected together by the strict rules of logic. Thomas Hales's paper, whatever its merits may have been, did not conform to that standard. It relied heavily on computer calculation (much like the proof of the four-color theorem from 25+ years before). But Robert MacPherson, Managing Editor of the *Annals*, was somewhat enamored of computer proofs. He liked the idea of a computer proof of this crusty old problem. And he was willing to consider it.

But who could referee such a work? MacPherson was able to use the stature and clout of the *Annals* to recruit a team of twenty Hungarian mathematicians, led by Gábor Fejes-Tóth (son of Laszlo Fejes-Tóth, the previously mentioned pioneer in the study of the Kepler problem), to study the paper and render a judgment. They spent several years poring over the details of this long and tedious mathematical work. In fact the refereeing process was so protracted and so arduous that there was attrition among the referees: some quit and others retired or died. At the end they said that they were able to confirm the mathematical parts of the paper but it was impossible to check the computer work. Thus no conclusive report could be rendered. Well, there is a bit more to it than that.

The twenty Hungarians ran a seminar for three semesters on the paper. They studied it for four years altogether. At the end they said they were 99% sure it was right.

But MacPherson was not daunted. He accepted Hales's paper for the

used to find extrema for systems of linear equations. The methodology has far-reaching applications in scheduling airlines, routing Internet messages, and doing `Google` searches.

Annals of Mathematics—it is [HAL2]. Furthermore, he planned to publish the paper with a displayed disclaimer that said, in effect,

> The *Annals* cannot be certain that this paper is correct. Nevertheless we feel that the paper is worthwhile.

This was a bellwether for modern mathematics! Think about the long tradition of proofs in mathematics. Think about what proofs represent. This is what distinguishes mathematics from biology and physics and engineering. We do not perform experiments and come to plausible conclusions (which could be, and often are, refuted later). We instead prove theorems once and for all. Once a theorem is proved, and its proof checked and validated, then the theorem stands forever. It is just as true, and just as useful, today as when it was proved. Thus we use the Pythagorean theorem with confidence today, because Pythagoras proved it 2500 years ago. We use the Prime Number Theorem with never a doubt today because Hadamard and de la Valleé Poussin proved it 100 years ago. The *Annals* had always stood a hard line in defense of the traditional notion of mathematical theorems and proofs. It had always had its papers refereed assiduously. The *Annals* wanted to maintain its position as the journal of record for major advances in mathematics. It wanted to be *certain* that the papers it published were correct. There is rarely a published correction or retraction of a paper published in the *Annals of Mathematics*. But now it was deviating from that line. The venerable *Annals of Mathematics* was going to publish a result that nobody could confirm!! And they were going to acknowledge this shortfall with a *disclaimer*.

As this book has amply demonstrated, mathematics is an intensely human activity—not the cut and dried calculations that some may suspect is all that it is. Certainly human beings can have an effect on the directions that the subject takes, or the problems that people choose to work on, or the value that people attach to certain pieces of work. It is human beings who decide who gets the awards and who gets the plum appointments and who is elected to the National Academy. But human beings can also play a role in what is accepted as a proof, and what stands as valid mathematics.

When John Horton Conway of Princeton found out the plans for Hales's paper in the *Annals*, he was quite perturbed. He phoned up MacPherson and gave him a piece of his mind. Conway's view was that the *Annals* should not publish a paper with a disclaimer. The appellation *Annals* means just what it says. This journal should be a showcase for mathematics that has been

certified to be correct. Nothing else. He convinced MacPherson to remove the disclaimer. It has now been replaced by a ringing endorsement.[4] Note that, according to the Web page of the *Annals of Mathematics*, the paper was received on September 4, 1998 and was finally accepted on August 16, 2005. It is well known that some journals can be rather slow. One and a half to two years may be a typical time for a journal to get a new paper refereed, typeset, and in print. The slower journals may take three years or more. But seven years is quite extraordinary. Of course it must be borne in mind that over four years of this time was for the refereeing!

The current version of the truth is that the *Annals* will publish an *outline* of Hales's proof.[5] Entitled *A Proof of the Kepler Conjecture*, this outline is 121 pages long (see [HAL2]). It is in the November, 2005 issue. The full details will appear elsewhere, in the July, 2006 issue of *Discrete and Computational Geometry*. Hales's work will take up the entire issue, and will be divided into six papers. These will comprise 265 pages.

This is also new territory for the estimable *Annals*. The *Annals of Mathematics* is, and has been, a stodgy old girl. She publishes complete, self-contained papers with complete proofs of new and important results. Certainly its publication of Thomas Hales's work, whatever form it may ultimately take, is charting new ground. And it is setting an example for other journals. It is possible that the entire nature of the publication of mathematical research will be affected by these actions.[6]

A lovely and detailed history of the Kepler sphere-packing problem, and related mathematical questions, appears in [ASW].

[4] At least that was the original plan. The paper [HAL2] has now appeared, and it contains *neither* a disclaimer *nor* a ringing endorsement.

[5] This is also new territory for the venerable *Annals of Mathematics*. The *Annals* does *not* publish outlines.

[6] At a recent meeting, held in England, to discuss the changing nature of mathematical proof, MacPherson commented on the handling of the Hales paper by the *Annals*. He claimed that the *Annals* has a new policy of accepting computer-generated and computer-assisted proofs. There is no public record of such a decision. He further asserted that the usual refereeing process of the *Annals* had "broken down" in the case of the solution of the Kepler sphere-packing problem. The evidence, as described here, suggests that there was more at play than such a simple explanation would suggest.

8.4 Thurston's Geometrization Program

In 1866, Alfred Nobel (1833–1896) invented dynamite. This was in fact a dramatic event. It was just the technology that was needed for the Panama Canal and other big engineering projects of the day. Thus there was great demand for tri-nitro toluene (TNT), and Nobel became a rich and successful man. He amassed quite a fortune. As is the case with many such wealthy and influential people, Nobel began, at the end of his life, to ponder his legacy. And he decided to create a prize to recognize the pinnacle of human achievement. It would entail considerable fanfare and honor for the recognized individual, and a substantial financial award. Thus was born the Nobel Prize.

Now Alfred Nobel was a practical man of the world. There was never any possibility that he would endow a Nobel Prize in metaphysical epistemology. It is likewise that he gave never a thought to mathematics. To this day, there is no Nobel Prize in mathematics.

But history is a funny thing. For as long as anyone can remember, mathematicians have been telling each other that the reason that Nobel did not endow a prize in mathematics is that a mathematician ran off with Nobel's wife. This story may require some explanation.

A notable and celebrated contemporary of Alfred Nobel was Gösta Mittag-Leffler (1846–1927). Mittag-Leffler, a student of the celebrated mathematician Karl Weierstrass, was a prominent mathematician in his own right. He married well, and as a result lived in a grand mansion in Djursholm, Sweden—just outside of Stockholm.[7] Now Mittag-Leffler was really a celebrity; his name was in the newspapers all the time. He dressed like a dandy, and was really a man about town. Nobel was a dowdy, stodgy, solitary *bachelor*. He never married, and as far as we know he never had a lady friend in his entire adult life.[8] He was extremely jealous of Mittag-Leffler and the lifestyle that he led. It might also be noted that the extremely beautiful and brilliant

[7] Djursholm is still, even to this day, a very ritzy place. Ordinary Swedes cannot afford to live there. In point of fact the town is packed with the mansions of the Ambassadors to Sweden from foreign countries.

[8] Some fairly recent revisionist biographies of Nobel depart from the version of his life that appeared in the "official" biography of himself that Nobel arranged and sanctioned. These newer books claim that, after Nobel's death, a woman came forward and filed a claim against Nobel's estate. She said that she had been, in effect, Nobel's common-law wife.

mathematician Sonja Kowalevski was an associate of Mittag-Leffler and lived in his house for a period of time. It seems that their relationship was more than Platonic. In any event, Mittag-Leffler represented a way of life that was anathema to Nobel. Mittag-Leffler was the most prominent and celebrated scientist in all of Sweden. Some thought it likely that, were there a Nobel Prize in mathematics, Mittag-Leffler would have received it. This may have influenced Alfred Nobel's decision *not* to found a prize in mathematics.

In any event, there is and was no Nobel Prize in mathematics. Period.[9] This eventuality had an interesting upshot. Mittag-Leffler was quite irritated at the fact of no Nobel Prize in mathematics, as he too figured that he would be the obvious recipient of any such prize. So he used his considerable resources to establish the *Mittag-Leffler Prize* in Mathematics. And he specified explicitly that the medal for the prize would be twice as large as the medal for the Nobel Prize. Mittag-Leffler was the founder of the quite prestigious mathematical journal called *Acta Mathematica*. The winner of the Mittag-Leffler Prize would receive a complete leather-bound edition of the journal. Another aspect of the prize would be a lavish banquet prepared by a famous French chef. Thus Mittag-Leffler strove to outclass the Nobel Prize in many different respects.

Unfortunately, whereas the Nobel Prize has survived and prospered for more than a century, the Mittag-Leffler Prize folded after being awarded only once or twice. The first time it was awarded to Charles de la Valleé Poussin, a brilliant French Fourier analyst. Professor de la Valleé Poussin was vacationing in the Alps at the time of the award, so it was arranged that the leather-bound copies of *Acta Mathematica* and the fancy French dinner be delivered to him up in the mountains.

[9]It should be noted, however, that the Nobel Prize is an organic entity. It grows and changes. As an example, the Nobel Prize in Economics was added as recently as 1966. And, even more recently than that, the Crafoord and Schock Prizes were added to the Nobel stable of awards. The Crafoord fund comes from a large pharmaceutical family, and the Schock money from a similar source. What is notable is that mathematicians are eligible for these two new prizes. Professor Louis Nirenberg of the Courant Institute of Mathematical Sciences was the recipient of the first Crafoord Prize. Professor Elias M. Stein of Princeton University has been awarded the Schock Prize. The money for these prizes is comparable in magnitude to that for the Nobel. Another new prize, which was first awarded in 2004, is the very prestigious Abel Prize. Sponsored by the Norwegian government, and named after the remarkable nineteenth century Norwegian mathematician Niels Henrick Abel, this prize has emerged as the nearest thing to the Nobel Prize for mathematics.

As anybody who has thought about the matter would know, the way that a prize like this survives is that its founder invests the money. The awards are made each year from the income of the investment. That is where the Mittag-Leffler Prize met its sorry end. For Mittag-Leffler chose to invest in the Italian Railroad System and German World War I Bonds. End of the Mittag-Leffler Prize.[10]

One corollary of all this colorful history is that the mathematicians have created their own prize. Known as the Fields Medal, this prize was established by John Charles Fields of the Canadian Mathematical Society. Fields was in fact the President of the International Congress of Mathematicians that was held in Toronto in 1924, and was editor of the Proceedings. He suggested that the considerable funds raised through the sale of those Proceedings be used to establish a research prize for young mathematicians. In the 1932 meeting of the International Congress in Zurich, it was voted to approve this new award. It was dubbed the *Fields Medal*.

The purpose of the prize is to recognize talent in promising young mathematicians. First awarded in 1936—to Lars Ahlfors and Jesse Douglas—the prize began as a modest encomium to recognize developing talent. Over time the Fields Medal has become the greatest honor that can be bestowed on a mathematician. To win the Fields Medal is to be virtually beatified in mathematical circles. The set of Fields Medalists is a very select and distinguished group.

The Canadian sculptor R. Tait McKenzie was enlisted to design the actual Fields *Medal*. The medal is 2.5 inches in diameter and the obverse side depicts the head of Archimedes facing to the right.

William P. Thurston was awarded the Fields Medal in 1982 in Warsaw, Poland. He had done brilliant work on foliation theory (a branch of topology, which is part of the modern theory of geometry)—*in his Ph.D. thesis!!* These results completely revolutionized the field—solving many outstanding problems and opening up new doors. So the recognition was richly deserved. Thurston went on to do groundbreaking work in all aspects of low-dimensional topology. He became a cultural icon for mathematicians young and old. He also had a great many brilliant students, and was able to spread his intellectual influence in that manner as well.

Around 1980, Thurston made a dramatic announcement. He had found

[10]Well, not quite. A search of the Web reveals that there is still some vestige of the Mittag-Leffler Prize that is awarded these days. But it is nothing like the original prize.

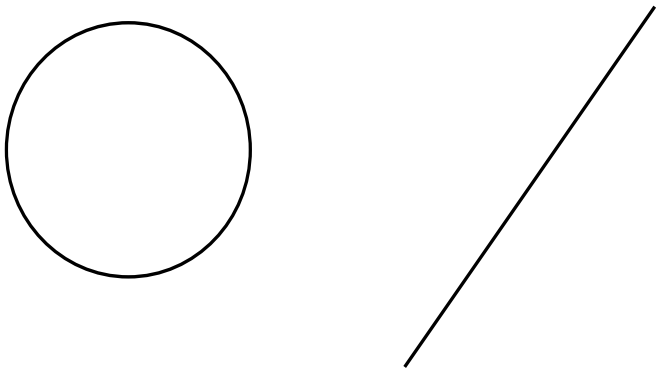


Figure 8.6

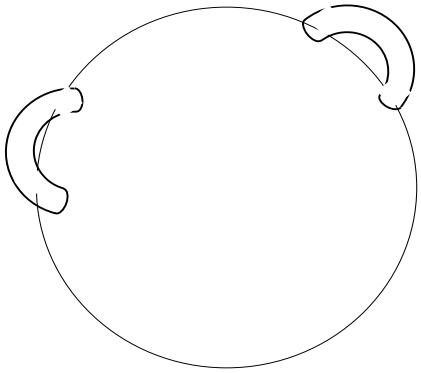


Figure 8.7

a way to classify all 3-dimensional manifolds. Thurston's set of ideas was dubbed the "geometrization program." For a mathematician, a manifold is a surface with a specified dimension. This surface may or may not live in space; it could instead be an abstract construct. It was a classical fact from nineteenth century mathematics that all 1-manifolds and 2-manifolds were completely understood and classified. The 1-manifolds are the circle and the line—see Figure 8.6. Any 1-dimensional "surface" is, after some stretching and bending, equivalent to a circle or a line. Any 2-dimensional surface (in space) is, after some stretching and bending, equivalent to a sphere with a certain number of handles attached; see Figure 8.7.

Thurston's daring idea was to break up any 3-dimensional manifold into pieces, each of which supports one of eight special, classical geometries.[11]

[11]In fact it was the nineteenth-century mathematician Bianchi (1856–1928) who, in

Thurston worked out in considerable detail what these eight fundamental geometries must be. His theorem would give a structure theory for three-dimensional manifolds.

Of course 3-dimensional manifolds are much more difficult to envision, much less to classify, than are 1- and 2-dimensional manifolds. Prior to Thurston, almost nothing was known about this problem. Certainly 3-dimensional manifolds are of interest from the point of view of cosmology and general relativity—just because we live in a three-dimensional space. For pure mathematicians, the interest of the question, and perhaps the driving force behind the question, was the celebrated Poincaré conjecture. Formulated in 1904 by Henri Poincaré of Paris, it posited that any 3-dimensional surface with the geometry/topology of the sphere actually is equivalent to the sphere. Poincaré formulated this question following upon earlier investigations of homology spheres. He had some ideas about proving it, but they turned out to be inadequate to the task. The problem has fascinated mathematicians for 100 years. It has important implications for the geometry of our universe, and is of central interest to mathematicians and cosmologists alike. Every few years there is an announcement—that even makes it into the popular press—of a new proof of the Poincaré conjecture. In 1986, Colin Rourke of Warwick announced a proof. He is a man of considerable reputation, and his proof survived until it was dissected in a seminar in Berkeley. In 2002, M. J. Dunwoody of the University of Southampton announced a proof. He even wrote a 5-page preprint. The effort quickly died.

Many mathematicians have tried and failed to prove the Poincaré conjecture. But, if Thurston's geometrization program were correct, then the Poincaré conjecture would follow as an easy corollary. Suffice it to say that there was considerable excitement in the air pursuant to Thurston's announcement. He had already enjoyed considerable success with his earlier work—he was arguably the greatest geometer who had ever lived—and he was rarely wrong. People were confident that a new chapter of mathematics had been opened for all of us.

But the problem was with the proof. The geometrization program is not something that one proves in a page or two. It is an enormous enterprise that reinvents an entire subject. This is what historian of science Thomas Kuhn [KUH] would have called a "paradigm shift". Although Thurston was

1898, first identified these eight fundamental geometries. But Thurston saw further than Bianchi, or anyone else, insofar as to how they could be used.

absolutely convinced that he could prove his new way of looking at low-dimensional geometry and topology, he was having trouble communicating his proof to anyone else. There were so many new ideas, so many new constructs, so many unfamiliar artifacts, that it was nearly impossible to write down the argument. After a period of time, Thurston produced a set of "notes" [THU3] explaining the geometrization program.

It is important to understand what the word "notes" means to a mathematician. As this book explains and attests, mathematics has a long legacy—going back to Euclid and even earlier—of deriving ideas rigorously, using rigid rules of logic, and recording them according to the strict axiomatic method. Correctly recorded mathematics is crisp, precise, clear, and written according to a very standard and time-honored model. As a counterpoint to this Platonic role model, modern mathematics is a fast-moving subject, with many new ideas surfacing every week. There are exciting new concepts and techniques springing forth at an alarming rate. Frequently a mathematician finds that he/she simply cannot take the time to write out his/her ideas in a linear, logical fashion. If the idea is a big and important one, it could take a couple or several years to get the recorded version just right. Frequently one feels that he/she just doesn't have time for that. So a commonly used alternative is "notes". What the mathematician does is give a set of lectures, or perhaps a course (at the advanced graduate level) and get some of the students to take careful notes. The professor then edits the notes (rather quickly) and then disseminates them. One of the treasures of the Princeton University Mathematics Library is a rather comprehensive collection of sets of notes—going back 75 years or more. Many a mathematician has cut his/her teeth, and laid the basis for a strong mathematics education, by studying those notes.

And this is where William Thurston found himself in 1980.[12] He had

[12]The Poincaré conjecture is one of those problems that frequently finds its way into the popular press. It is one of the really big problems in mathematics, and when someone claims to have solved it that is news. In the mid-1990s, Valentin Poenaru professed that he had a proof of the Poincaré conjecture. Poenaru is a Professor at the University of Paris, and a man of considerable reputation. He produced a 1200-page manuscript containing his thoughts. Unfortunately none of the experts were able to battle their way through this weighty tract, and no definitive decision was ever reached on Poenaru's work. In 1999 he published an expository tract with a summation of his efforts. Certainly Poenaru has contributed a number of important ideas to the subject of low-dimensional topology. But the jury is still out on whether he has proved the Poincaré conjecture.

In 2002, M. J. Dunwoody announced *his* proof of the Poincaré conjecture. The good

one of the most profound and exciting new ideas to come along in decades. It would take him a very long time to whip all these new ideas into shape and shoehorn them into the usual mathematical formalism. So he gave some lectures and produced a set of notes. The Princeton University Mathematics Department, ever-supportive of its faculty, reproduced these notes and sold them (with postage) to anyone—worldwide—who would send in a modest fee.

It is safe to say that these notes were a blockbuster. Many, many copies were sold and distributed all over the planet. There were so many beautiful new ideas here, and many a mathematician's research program was permanently affected or changed because of the new directions that these notes charted. But the rub was that nobody believed that these notes constituted a proof of the geometrization program. Thurston found this very frustrating. He continued to travel all over the world and to give lectures on his ideas. And to produce Ph.D. students who would carry the torch forth into the mathematical firmament. But he felt that this was all he could do—given the time constraints and limitations of traditional mathematical language—to get his ideas recorded and disseminated. The catch was that the mathematical community—which in the end is *always* the arbiter of what is correct and accepted—was not ready to validate this work.

In fact Thurston's pique with this matter was *not* transitory. In 1994, he published the paper *On proof and progress in mathematics* [THU1] which is a remarkable polemic about the nature of mathematical proof. It also, *sotto voce*, castigates the mathematical community for being a bit slow on the uptake in embracing his ideas. This article was met with a broad spectrum of emotions ranging from astonishment to anger to frustration.

Many years later Thurston published a more formal book [THU2]—in the prestigious Princeton University Press Mathematics Series—in which he began to systematically lay out the details of his geometrization program. In this tract he began at square one, and left no details to the imagination. He in effect invented a new way to look at geometry. His former Ph.D. student Silvio Levy played a decisive role in developing that book. And it

news this time was that the paper that he wrote was only five pages long. Anyone could read it. But that was also the bad news. The fact that anyone could read it meant that it was not traditional, rigorous, hardcore mathematics, written in the usual argot. In fact the paper was rather informal and fanciful. It was difficult to tell whether the work should be taken seriously (even though Dunwoody is certainly a solid mathematician of good repute). In the end, Dunwoody's contribution was deemed flawed.

is a remarkable and seminal contribution to the mathematical literature. In fact the book recently won the important AMS Book Prize. But it should be stressed that this book is the first step in a long journey. If the full saga of Thurston's proof of the geometrization program is to be revealed in this form, then a great many more volumes must appear. And they have not appeared yet.

8.5 Grisha Perelman's Attack on the Poincaré Conjecture and the Geometrization Program

In 1982, mathematician Richard Hamilton introduced a new technique into the subject of geometric analysis. Called the method of "Ricci flows", this is a means of studying a flow generated from a source—like a heat flow. See Figure 8.8. What Hamilton does is to write down a differential equation on a given manifold that, in effect, mandates a notion of distance (what mathematicians call a "metric") and a motion of the manifold so that the velocity at any given point can be expressed in terms of the curvature. This idea is inspired perhaps by what happens to a closed, one-dimensional curve in the plane when you do the same thing to it: as the process evolves, the curve smooths out, all its dents and twists and kinks disappear, and it becomes a circle. See Figure 8.9.

So this is what is also supposed to happen to a higher-dimensional manifold when it is subjected to the Ricci flow. The trouble is that things are much more complicated in higher dimensions. It is difficult to prove existence of a solution to the partial differential equation. And also some nasty singularities can arise during the deformation process (that we somewhat poetically depicted in Figure 8.9). Particularly tricky are the so-called "cigars" that are like long, thin tubes—see Figure 8.10. There is a sophisticated technique, developed at Princeton University by Bill Browder and John Milnor, called "surgery theory" that allows one to cut out these nasty singularities—as with a knife—and then to plug up the holes. The trouble was that, in the situation for the Poincaré conjecture, the singularities could run out of control. It was possible that one singularity could be removed and several others could spring up to take its place. Perelman's deep insight was to be able to show that the singularities evolved in finite time. This gave the means for controlling them and, ultimately, eliminating them. The successive process of eliminating the singularities results in a nicer manifold—closer to the goal.

Hamilton could see that this was potentially a method for solidifying the method initiated by Thurston in his "geometrization program". For one could start Ricci flows at different points of a surface and thereby generate the "geometric pieces" that Thurston predicted—refer to Figure 8.11. The idea of each of these special pieces is that it contains an "ideal geometry"—one in which a minute creature, equipped with a tape measure for the special

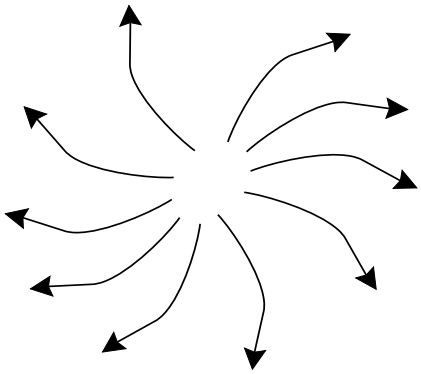


Figure 8.8

metric on that piece, could not tell one point from another.

Hamilton was only able to harness the analytic techniques to completely carry out this idea for surfaces of dimension 2. [He made significant inroads in the case of dimension 3—enough to convince people that this was a potential program for proving the Poincaré conjecture—but some of the difficult estimates eluded him.] And, as we have noted, 2-dimensional surfaces had already been classified by Jordan and Möbius in the mid-nineteenth century. Nothing new resulted from Hamilton's application of the Ricci flow to the study of 2-dimensional surfaces. He was able to obtain some partial results about existence of solutions for Ricci flows in dimension 3, but only over very small time intervals. He could make some interesting assertions about Ricci curvature, but these were insufficient to resolve the Poincaré conjecture in 3 dimensions. And this is where Ricci flows have stood for more than twenty years.[13]

Enter Grigori (Grisha) Perelman. Born in 1966, Perelman exhibited his mathematical genius early on. But he never had any particular designs to be a mathematician. His father, an enineer, gave him problems and things to read, but Perelman entered the profession crablike.

He won the International Mathematics Olympiad with a perfect score in 1982. Soon after his Ph.D. at St. Petersburg State he landed a position at

[13]Hamilton has certainly emerged as a hero in this story. He had the early idea that broke the Poincaré conjecture wide open. Although he could not himself bring the program to fruition, he was greatly admired both by Perelman and by Yau. He gave the keynote address at the International Congress of Mathematicians in August, 2006 about the status of the Poincaré conjecture.

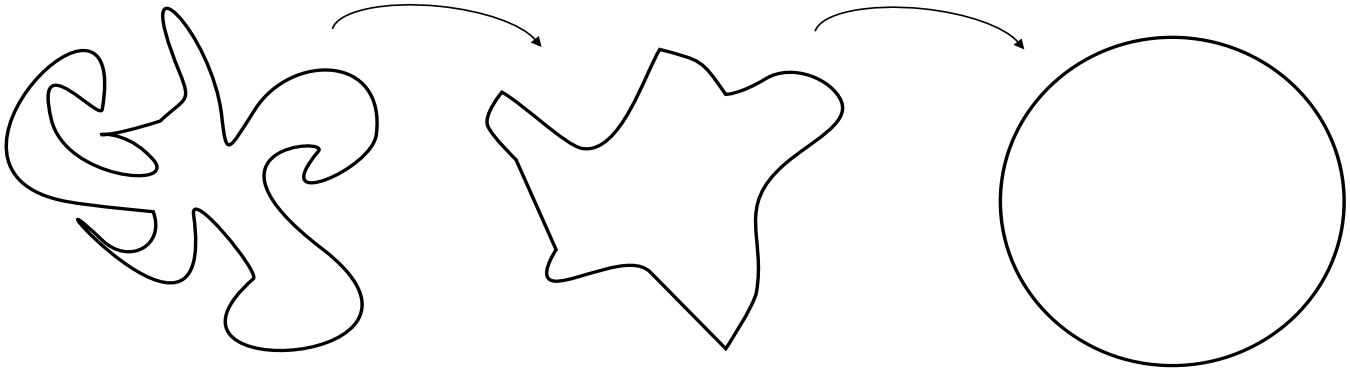


Figure 8.9

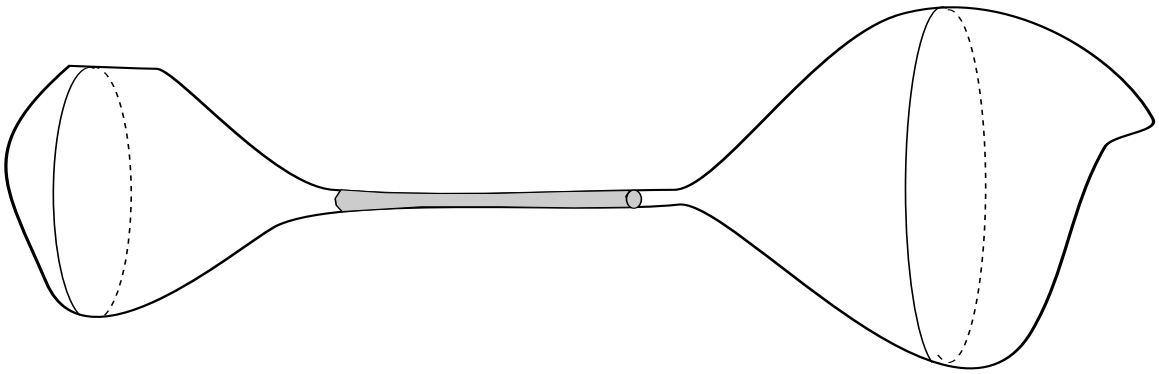


Figure 8.10

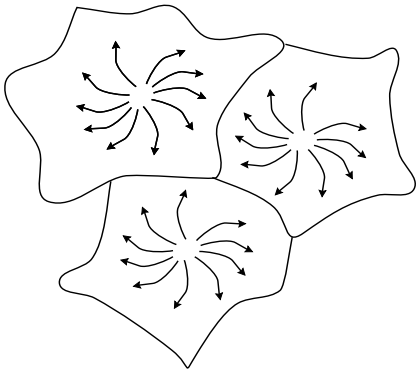


Figure 8.11

the estimable Steklov Institute in St. Petersburg (the Steklov Institutes are the most prestigious mathematics institutes in all of Russia). He accepted some fellowships at New York University and SUNY Stony Brook in 1992, and made such an impression that he garnered several significant job offers. He turned them all down. He certainly impressed people early on that he was an unusual person. He kept to himself, let his fingernails grow to 6 inches in length ("If they grow, why wouldn't I let them grow?" said Perelman), kept a spartan diet of bread and cheese and milk, and maintained an eccentric profile.

In 1993 Perelman began a two-year fellowship at U. C. Berkeley. At this time he was invited to be a speaker at the 1994 International Congress in Zurich. He was also invited to apply for jobs at Stanford, Princeton, the Institute for Advanced Study, and the University of Tel Aviv. He would have none of it. When asked to submit his *curriculum vitae* for the job application, Perelman said, "If they know my work, they don't need to see my CV. If they need my CV, then they don't know my work."

In 1996 Perelman declined a presigious prize for young mathematicians from the European Mathematical Society. It is said that Perelman claimed that the judging committee was not qualified to appreciate his work. But at this point everyone knew that Grisha Perelman was a man who was going places.

In the year 2002, on November 11 to be precise, Perelman wrote the groundbreaking paper *The entropy formula for the Ricci flow and its geometric applications* [PER1]. The fourth page of the introduction to this important paper contains the statement that, in Section 13, he will provide a brief sketch of the proof of the (Thurston) geometrization conjecture. Well, the "proof" was pretty sketchy indeed; and we are still unsure just what the paper [PER1] proves and does not prove. But the paper certainly set the world aflame.

On November 19, 2002, geometer Vitali Kapovitch sent Perelman an e-mail that read

> Hi Grisha, Sorry to bother you but a lot of people are asking me about your preprint "The entropy formula for the Ricci . . . " Do I understand it correctly that while you cannot yet do all the steps in the Hamilton program you can do enough so that using some collapsing results you can prove geometrization? Vitali.

Perelman's reply the very next day was, "That's correct. Grisha." Coming

from a mathematician of the established ability and credentials of Perelman, this was a bombshell.

On March 10, 2003, Perelman released a second paper entitled *Ricci flow with surgery on three-manifolds* [PER2]. Among other things, this new paper filled in many of the details for the sketch provided in the first paper.

In April, 2003 Professor Perelman gave a series of lectures at several high-level universities in the United States, including MIT (home of some of the most distinguished experts), SUNY at Stony Brook, New York University, and Columbia. These talks made a sufficiently strong impression that people began to take Perelman's program very seriously. In July of that same year, Perelman released a third paper entitled *Finite extinction time for the solutions to the Ricci flow on certain three-manifolds*. This paper provides a simplified version of a proof of a part of the geometrization program; that result is sufficient to imply the Poincaré conjecture.

One might cautiously compare Perelman's nine-month period—from November, 2002 to July, 2003—with Albert Einstein's "miracle year" (1905), in which he published five papers that completely changed the face of modern physics. In one of these papers Einstein introduced special relativity almost as an afterthought[14]—and it took several years for the idea really to catch on. But these were the five papers that really started everything. Just so for Perelman. His three papers profess to be able to harness the Ricci flow in dimension three. He claims to be able to prove the assertions of Thurston's geometrization program, and therefore to also be able to prove the Poincaré conjecture.

It must be stressed that Perelman's three papers are full of original and exciting ideas. But they are written in a rather informal style. And Perelman has no plans to publish them. They are posted on the preprint server `arXiv`, and that is where they will remain. They will never be vetted or refereed, at least not in any formal sense. So it is up to the mathematical community to assess what these papers have to offer, and how they fit into the firmament. A lovely discussion, on an informal level, of the Poincaré conjecture and Perelman's contributions, appears in [STRZ].

One rarely sees in mathematics the level of excitement and intense activity that these three papers by Perelman have generated. There are conferences being organized all over the world. Such an august institution as the Clay Mathematics Institute in Cambridge, Massachusetts has funded two math-

[14]This in the sense that relativity isn't even mentioned in the title!

ematicians (Bruce Kleiner and John Lott) at the University of Michigan to develop Perelman's ideas and produce a detailed and verifiable proof of the Poincaré conjecture. Two other very prominent mathematicians (John Morgan of Columbia and Gang Tian of Princeton) are giving a large segment of their time and effort to attempt to write up all the details of Perelman's program.

There are a number of unusual threads to this story—some of which we have introduced in the last paragraph—that require explication. Let us begin with the Clay Mathematics Institute.

It is just a fact of life that most mathematics institutes are funded with government money. The mathematics institute at Luminy in France, the Isaac Newton Institute in Cambridge, England, the Banff International Research Station in Banff, Canada, the famous institute IMPA in Brazil, the Mathematisches Forschungsinstitut in Oberwolfach, Germany are all funded with national government monies. There are seven federally funded mathematics institutes in the United States. But it is a remarkable fact that the United States enjoys a number of privately funded mathematics institutes.

The first private mathematics institute in this country was founded in the 1930s by Louis Bamberger of Bamberger's Department Stores. Working with the prominent academician Abraham Flexner, Bamberger was ultimately convinced to found an "Institute of Useless Knowledge". And thus was born the Institute for Advanced Study in Princeton, New Jersey. Centered at Fuld Hall on a lovely grounds that is also a game preserve, the Institute is a haven for very advanced research activity in mathematics, astrophysics, history, social science, and other areas. When Flexner and his collaborators were setting up the Institute for Advanced Study, one of their first jobs was to select the founding faculty. These faculty would be permanent, would be paid princely salaries, and would be charting an important course for intellectual development in this country. So the selection had to be made with the utmost care. What they found was that the only subject area in which they could get a real consensus on who were the best people was mathematics (and mathematical physics). So that is where they started. And one of their very first appointments was Albert Einstein. Certainly it was this appointment that really put the fledgling Institute for Advanced Study on the map. Over the years, the Institute for Advanced Study has had an astonishing panorama of powerful and influential faculty. It has hosted any number of important conferences and intellectual events. On the whole, it has been a great success as a scholarly institution.

John Fry was a college student at the University of Santa Clara. There he fell under the spell of mathematics—particularly under the tutelage of Professor Gerald B. Alexanderson. Fry was a football star in college, and also was a good friend of math major Brian Conrey. Well, Fry went on to become a successful retailer. Specifically, his father owned a chain of supermarkets. At some stage he gave each of his children $1 million to see whether they could make something of themselves. John Fry built the very successful *Fry's Electronics* chain of emporium-style electronics stores. There are now 33 of them in four states, there are plans to expand across the country and around the world, and John Fry is worth over $1 billion. In spite of his success and wealth (or perhaps because of it), Fry has maintained his interest in mathematics. He has been collecting rare mathematics books for some years. Included in his collection are Napier's book on logarithms, some of Euler's early books, and a first edition of Isaac Newton's *Principia* signed by Stephen Hawking (who presently holds Newton's Lucasian Chair at Cambridge University).

In 1994 John Fry decided to start a mathematics institute. Called The American Institute of Mathematics (AIM), it is presently located in Palo Alto, California in the former corporate headquarters of Fry's Electronics. AIM enjoys generous funding each year from John Fry, and it also has funding from the federal National Science Foundation (the Princeton Institute for Advanced Study has similar support from NSF). The Director of AIM is Brian Conrey, Fry's old college friend.

AIM has purchased a 1900 acre plot in Morgan Hill, California (just Southeast of Silicon Valley), and is building a world-class home for its mathematics institute. It is modeled, in considerable detail, after the medieval Spanish fortress known as the Alhambra; but it will be about twice the size. The grounds are also home to a beautiful 18-hole custom golf course where John Fry enjoys an occasional game of golf with his friends. Fry has also hired two famous chefs—who formerly worked at fancy resorts in Northern California. They are currently available to cook the occasional lunch for Fry and his golfing buddies. But, when the new institute is up and running in Morgan Hill, these two will prepare the meals for the hard-working mathematicians.

A new addition to the "privately funded mathematics institute" scene is the Clay Mathematics Institute in Cambridge, Massachusetts. Founded in 1998 with money from Landon T. Clay (a venture capitalist who is a direct descendent of Cassius Clay, the Civil War general), this institute has really

made a splash. Well, the things you can do with $50 million. And one of the remarkable things that the Clay Institute has done is to establish the "Millenium Problems". This is a collection of seven well-known mathematical problems that have been dubbed the most prominent and important and challenging unsolved problems in all of mathematics.[15] These are *the* mathematics problems for the twenty-first century. What is remarkable is that the Clay Institute has offered a $1 million prize for the solution of any of these seven problems.

Sketchily described, the problems are these:

- The Birch and Swinnerton-Dyer Conjecture [algebraic geometry]
- The Hodge Conjecture [complex geometry]
- The Navier-Stokes Equations [mathematical physics, partial differential equations]
- The **P vs. NP** Problem [logic, theoretical computer science]
- The Poincaré Conjecture [topology]
- The Riemann Hypothesis [analytic number theory, complex analysis]
- The Yang-Mills Theory [mathematical physics, partial differential equations]

The Clay Institute has detailed and strict rules for how purported solutions to these problems will be judged. A portion of the strictures is thus:

> Before consideration, a proposed solution must be published in a refereed mathematics journal of world-wide repute, and it must also have general acceptance in the mathematics community *two years* [after that publication]. Following this two-year waiting period, the SAB [Scientific Advisory Board of the Clay Institute] will decide whether a solution merits detailed consideration

[15] A lovely and authoritative discussion of these problems appears in [CJW]. See also [DEV1]. This situation is of course quite similar to what Hilbert did in 1900. But there are two important differences. One is that this time a *committee* of distinguished scientists put together the list of problems. The other is that each of these new problems has a $1 million bounty attached to it. Only the Riemann hypothesis was both on Hilbert's list and is on the Clay list.

> . . . The SAB will pay special attention to the question of whether a prize solution depends crucially on insights published prior to the solution under consideration. The SAB may (but need not) recommend recognition of such prior work in the prize citation, and it may (but need not) recommend the inclusion of the author of prior work in the award.

So far nobody has staked a claim for any of the Clay Prizes.[16] The person who is closest to being able to do so is Grisha Perelman with his "solution" of the Poincaré conjecture.

But there is the rub. For Perelman is not playing by any of the rules of the Clay Math Prizes. First of all, he has not published any of his three papers. They are posted on what is called a "preprint server". What is that?

Another intellectual detour is called for at this point. The traditional way of publishing mathematical research, and of thereby planting your flag and attaching your name to an important new result, is cut and dried. It goes like this. You write out your ideas in careful detail, with complete proofs, in longhand on $8\frac{1}{2}'' \times 11''$ paper. You proofread it several times, editing as you go, to make sure that every detail is absolutely correct. You take special care to see that the list of references is accurate and complete, so that due credit is given to all the people who have worked in this subject area prior to you. You double check all your citations, and you make sure that all the results you use are stated correctly. You check to be certain that each of your own new proofs is rock solid.

When you are confident that you have a winner, then you get the paper typed up. In the *really old days*, this was achieved in a rather primitive fashion. That is to say, the department would have a manuscript typist who typed all the English words in the paper, leaving spaces (of suitable sizes) for the mathematical expressions. Then the mathematical expressions were written in by hand—usually by the mathematician himself.

In the semi-old days (from 1961 to 1985), the creation of a mathematical paper became a new type of ordeal. Because what technology was available in the typical math department to render mathematics on paper? The one

[16] There is a longstanding tradition of people offering money for the solutions of various mathematical problems. Andrew Wiles won the Schock Prize, the Prix Fermat, the Cole Prize, and other honors for proving Fermat's Last Theorem. Mathematician Paul Erdős frequently offered monetary bounties for the solutions of his pet problems. Even though his resources were meager, he always paid up.

and only thing that we had at that time was the IBM Selectric Typewriter. If you have ever seen one of these machines, then you know that it did not have separate imprint slugs, each on the end of an arm, that struck the page in sequence. That is the way that the original typewriters, from the early twentieth century, worked. But the IBM Selectric was a revolutionary design. What it had was a metal ball, of diameter about $1.5''$, with different symbols embossed all over it. When you struck a key on the keyboard, the ball would rotate rapidly so that the designated embossed metal letter faced the page, and then it hammered the letter (through an inked ribbon) onto the page.

So the standard, default ball in the IBM Selectric had the standard roman alphabet characters a b c d ... A B C D ... and the numerals 1 2 3 ... and the standard typographical special symbols like & % # ... and so forth. But there was another IBM Selectric ball that had the Greek alphabet. And another that had the integral signs and differentiation signs and other artifacts of calculus. And this is how mathematics manuscripts were created for many years—say from 1961 to 1985.

It took some real talent and training, and a great deal of practice, to use the IBM Selectric effectively and well. A mathematical paper typed up on an IBM Selectric looked professionally prepared, and was fairly easy to read. But it certainly did *not* look typeset. Mathematics involves elaborate displays of symbols, using characters of different sizes, such as

$$\frac{\displaystyle\int_a^b \frac{ax^2+b}{\sin\frac{x}{3}-\tan\frac{x^3}{2}}\,dx}{\ln(e^{2x-x^2}-x^{3^b})} = \frac{\det\begin{pmatrix} x^3 & y-3 & \frac{z}{w} & y^z \\ \frac{\sin x}{\cos y} & \tan x\cdot x^x & x^3\cdot y^2 & \frac{x}{\cos x} \\ \frac{y-x}{x-z} & (x-w)\cdot e^{\cot x} & \cos\left(\frac{x}{y}\right) & x^{z^y} \\ x & 0 & y & e^x \end{pmatrix}}{\dfrac{x^3-x}{y^3+y}}$$

$$= \frac{\dfrac{\dfrac{a}{b}-\dfrac{b^a}{c^d}}{x^3+y^3}}{\dfrac{y}{x}}.$$

A typewriter, by nature, is a *monospaced* environment. This means that all the characters are the same size and width, with the same amount of

space between them. Vertical spacing is quite tricky. The result of the IBM Selectric rendition of a paper was readable mathematics, but not the way it is typeset in the mind of God.

But in the early 1980s everything changed dramatically. Donald Knuth of Stanford University invented the computer typesetting system TEX. This book is typeset in TEX. In fact today most mathematics is typeset in TEX. TEX is a high-level computing language like `Fortran` or `JAVA`. It is *not* a word processor. When you create a document in TEX, you are in fact issuing commands for how you want each character and each word to appear and to be positioned on the page. Some have called the invention of TEX the most important event in the history of typesetting since Gutenberg's printing press.

Just to give the reader an idea of how TEX works, we now provide a sample. The user opens up a new file and, using a text editor, inputs TEX code that looks something like this:

```
**********************************************************

\documentclass{article}

\newfam\msbfam
\font\tenmsb=msbm10 scaled \magstep1 \textfont\msbfam=\tenmsb
\font\sevenmsb=msbm7 scaled \magstep1 \scriptfont\msbfam=\sevenmsb
\font\fivemsb=msbm5 scaled \magstep1 \scriptscriptfont\msbfam=\fivemsb
\def\Bbb{\fam\msbfam \tenmsb}

\def\RR{{\Bbb R}}
\def\CC{{\Bbb C}}
\def\QQ{{\Bbb Q}}
\def\NN{{\Bbb N}}
\def\ZZ{{\Bbb Z}}

\begin{document}

Let $f$ be a function that is continuous from complex numbers
$\CC$ to itself.  Consider the
auxiliary function
$$
g(z) = \frac{\displaystyle \int_a^b \frac{\alpha z + \beta}{\gamma z
```

```
                  + \delta} \, dz}{\hbox{det} \
     \left (
   \begin{array}{cc}
         2z - 3 & z^2 + z \\
    z^z  & \sin z
   \end{array}
      \right )} \, .
$$
Then
$$
g \circ f (z)
$$
operates in a natural manner on the Banach space of continuous functions on
the unit interval $I$.

\end{document}
```

****

The user compiles this TEX code using a standard TEX compiler that is available commercially, or by download from the Web. The output looks like this:

****

Let f be a function that is continuous from the complex numbers $\mathbb{C}$ to itself. Consider the auxiliary function

$$g(z) = \frac{\displaystyle\int_a^b \frac{\alpha z + \beta}{\gamma z + \delta}\, dz}{\det \begin{pmatrix} 2z - 3 & z^2 + z \\ z^z & \sin z \end{pmatrix}}.$$

Then

$$g \circ f(z)$$

operates in a natural manner on the Banach space of continuous functions on the unit interval I.

****

The user may send this output to a screen, to a printer, to a FAX machine, or to a number of other devices. TEX is a powerful and flexible tool that has become part of the working life of every mathematician.

Today most mathematicians know how to use TEX, and they type their own papers in TEX. The result is that most mathematics papers now are actually typeset, just as they would be in a finished book. Mathematical characters are sized and positioned and displayed just as they would be in a typeset script for a first-class mathematical monograph. This opens up many possibilities.

A paper that is prepared in TEX is easily exported to `PostScript` or an Acrobat-readable format (such as `*.pdf`). Such a file in turn may easily be posted on the World Wide Web. A paper in `*.pdf` that is posted on the Web will have all mathematical formulas, all text, and all graphics (i.e., figures) displayed just as they would be in a published paper or book. And many mathematicians choose to disseminate their work in just that fashion as soon as it is completed. This is what electronic preprint servers are for. A preprint server is a computer hooked into the WorldWide Web that serves as a forum for mathematical work. It archives thousands of papers in an attractive and accessible fashion, offers the end user a choice of formats, and displays each paper just as it would appear in hard copy—with all the mathematical notation, all the text, all the graphics, all the formulas.

A preprint server is a clearing house for new mathematical work. It makes the fruits of our labors available for free to everyone in the world who has access to a computer. For rapid and free dissemination of new scientific ideas, there is nothing to beat a preprint server.

But it should be stressed that an electronic preprint server engages in *no vetting* of the work that it displays. There is no refereeing and no reviewing. And this is as it should be. For the people running the preprint server can assume no responsibility for the work that it displays (if they conducted reviewing, then they *would* bear some responsibility). Moreover, it is most desirable that the preprint server should run without human intervention. Anyone at all should be able to post his/her work on the server with no assistance or manipulation from the people who run the server.

The most prominent and influential and important mathematics preprint server in the world today is `arXiv`. Begun at Los Alamos by Paul Ginsparg—over the objections and distinct lack of support of his supervisors—`arXiv`

began as a rather small-scale preprint server in high energy physics and is now a very large enterprise that encompasses physics, mathematics, and other subjects as well. There are tens of thousands of preprints now available on `arXiv` for free and easy download to anyone who wishes to read the work. And the number is growing by leaps and bounds with each passing day. Ginsparg has won a MacArthur Prize for his work in developing `arXiv`. He has also moved to Cornell University, where he is a Professor of Physics. And now `arXiv` makes its home at Cornell. A happy ending for all.

Well, to make a long story short, Grisha Perelman put his seminal work on the Poincaré Conjecture and the Thurston geometrization program on `arXiv` and nowhere else. He has no intention of publishing his papers in the traditional fashion. Thus there has been no refereeing or reviewing. And the papers, sad to say, are not written in the usual tight, rigorous, take-no-prisoners fashion that is the custom in mathematics. They are rather loose and informal, with occasional great leaps of faith.

Perelman himself has—in spite of receiving a number of attractive offers from major university math departments in the United States—returned to St. Petersburg so that he can care for his mother. He is more or less *incommunicado*, not answering most letters or *e*-mails. His view seems to be that he has made his contribution, he has displayed and disseminated his work, and that is all that he has to say. Perelman's position at Steklov pays less than $100 per month. But he lives an ascetic life. The jobs in the West that he declined pay quite prestigious (six-figure) salaries. Perelman claimed that he made enough money at his brief jobs in the West to support himself for the rest of his life.

But Perelman has left the rest of us holding the bag. The most recent information is that he has resigned his prestigious position at the Steklov Institute in St. Petersburg so that he can enjoy the solitude. He now says that he is no longer a part of the mathematics profession. Because of a "'competing" paper by Cao and Zhu (see below), and because of vigorous campaigning by various highly placed mathematicians, Perelman has concluded that the mathematics profession is deficient in its ethical code. He says that he has now quit mathematics to avoid making a fuss:

> As long as I was not conspicuous, I had a choice. Either to make some ugly thing or, if I didn't do this kind of thing, to be treated as a pet. Now, when I become a very conspicuous person, I cannot stay a pet and say nothing. That is why I had to quit.

All rather sad, and reminiscent of Fields Medalist Alexander Grothendieck (1928–). Grothendieck won the Fields early on for his work (in functional analysis) on nuclear spaces. He later shifted interests, and worked intensely with his teacher Jean Dieudonné to develop the foundations of algebraic geometry. There is hardly any twentieth century mathematician who has received more honors, or more attention, than Grothendieck. He occupied a chair at the prestigious Institute des Hautes Études Scientifique for many years—indeed he was a founding member (along with Dieudonné). But, at the age of forty, Grothendieck decided to quit mathematics. Part of his concern was over government funding, but an equally large concern was lack of ethics in the profession. Even in 1988 Grothendieck turned down the prestigious Wolf Prize; his remarks at that time indicated that he was still disgusted with the lack of ethical standards among mathematicians. Today Grothendieck lives in the Pyrenees and is intensely introspective, to say the least. He believes that many humans are possessed by the devil.

Perelman was recently awarded the Fields Medal at the International Congress of Mathematicians in Madrid (the other recipients were Andrei Okounkov, Terence Tao, and Wendelin Werner). This is without question the highest encomium in the profession.[17] Perelman did not show up to accept his honor. In fact the Fields Committee determined at the end of May, 2006 to award Perelman (and three other mathematicians) the Fields Medal. President of the International Mathematical Union Sir John M. Ball traveled to St. Petersburg to endeaver to convince Perelman to accept the medal. Perelman was quite gracious, and spent a lot of time with Ball, but was adamant that he would not accept the award. He made it plain that what was important was solving the problem, not winning the prize.

One of the staunch rules of the Clay Institute Prizes is that any candidate for one of the $1 million prizes must submit the paper to a prominent journal. It must be published, and it must stand for a period of at least two years while it is reviewed and discussed by the world body of mathematicians. Perelman has not played by this game, so he will most likely not receive a Clay Prize—whether he deserves one or not.

There is recent information that Perelman does not want to publish his

[17]Indeed, being selected for a Fields Medal is analogous to being canonized in the Catholic Church. And the selection procedure is just about as complicated. Technically, Perelman was over forty years old (the traditional cutoff for the Fields)—his fortieth birthday was in June 13, 2006. But the Fields Committee did not want to put too fine a point on it. They gave him the prize.

proof because then he would be a legitimate candidate for one of the seven Clay Millenium Prizes. And then, once he has $1 millon, Perelman fears that some Russian gangster may murder him.

A footnote to the fascinating Perelman story is that, just as this book was being written, Swiss mathematician Peter Mani-Levitska announced his own proof of the Poincaré conjecture. He has written a self-contained, 20-page paper. And then he retired from his academic position and has not been heard from since. His proof uses combinatorial techniques that harken back to some of the earliest ideas about the Poincaré conjecture. And Mani-Levitska, although not a topologist by trade, is one of the ranking experts in these combinatorial techniques. One version of the story is that select experts are now reviewing Mani-Levitska's proof. Another version is that the journal *Commentarii Mathematicae Helvetici* has accepted the paper (suggesting that the paper has been refereed and verified by *someone*). The one sure fact is that most of the cognoscenti have been sworn to secrecy in the matter. It is difficult to get any hard information.

The very latest development—just breaking as this book is being put in final form—is that Huai-Dong Cao and Xi-Ping Zhu have a 334-page paper appearing in *The Asian Journal of Mathematics*—see [CAZ]—which purports to prove *both* the geometrization program *and* the Poincaré conjecture. That is S. T. Yau's journal, so this event carries some weight. It may be noted that the Cao/Zhu paper was published without the benefit of any refereeing or review. Yau obtained the approval of his board of editors, but *without* showing them the paper. On the other hand, Cao was a student of Yau. Yau is perhaps the premiere expert in the application of methods of nonlinear partial differential equations in geometry. He presumably read the paper carefully, and that counts for a lot.

Yau has aggressively promoted the Cao/Zhu work. Perelman himself has expressed some skepticism over what contribution this paper actually makes to the problem. He uses the paper as a touchstone to cast aspersions on the general ethical tenor in the mathematical profession. Unfortunately, it has recently come to pass that the Cao/Zhu paper has been cast in a negative light. First, a portion of the paper was cribbed from the work of Kleiner and Lott. Some apologies have been tendered for that gaffe. The final resolution of that difficulty has yet to transpire.

Perhaps the most compelling component to date in this complicated story is the book of John Morgan and Gang Tian. These are two widely respected authorities who have written a 473-page book giving all the details, and filling

in all the gaps, of Perelman's proof of the Poincaré conjecture (they make no attempt to deal with the geometrization program). Funded in part by the Clay Mathematics Institute, Gang and Tian have submitted their book to the Clay Institute. The America Mathematical Society will consider the book for publication. It is now being refereed. S. T. Yau and Gang Tian are now rivals (even though Tian was Yau's student), so Yau has not had much to say about the book. But it has a life of its own.

Because of these last developments, it can now be said that Perelman's work has been carefully refereed and vetted. Two world-renowned experts have pronounced it (after some considerable ministrations of their own) to be correct. The entire Morgan/Tian book is posted on the Web, and therefore available for checking to the entire world. So it seems likely that the saga of the Poincaré conjecture has been brought to closure. One wonders what Henri Poincaré himself might have thought of this development that his problem inspired. Or what he would say about the final solution.

The American Mathematical Society holds a large annual meeting, joint with the Mathematical Association of America, each January. In 2007 that meeting was held in New Orleans, Louisiana. James Arthur, President of the AMS, had planned to have a celebration of the Poincaré conjecture at the gathering. A whole day of talks and discussions was planned. Fields Medalists John Milnor and William Thurston were to give background in the morning on the Poincaré conjecture and the geometrization problem respectively. In the afternoon, Richard Hamilton and John Morgan and John Lott were to speak about their contributions to the program. Unfortunately, Hamilton backed out; he cited other commitments and general fatigue. After considerable negotiation and cogitation, it was decided to invite Zhu to replace him. At that point Lott said he would not share the stage with Zhu. Efforts were made to rescue the situation, but to no avail. The event was ultimately canceled. There are hopes that a new event may be organized in the near future.

Chapter 9

A Legacy of Elusive Proofs

Modern mathematics is nearly characterized by the use of rigorous proofs. This practice, the result of literally thousands of years of refinement, has brought to mathematics a clarity and reliability unmatched by any other science. But it also makes mathematics slow and difficult: it is arguably the most disciplined of human intellectual activities. Groups and individuals within the mathematics community have from time to time tried being less compulsive about details of arguments. The results have been mixed, and they have occasionally been disastrous.

Arthur Jaffe and Frank Quinn

I have the impression that applying rigor to a theoretical idea is given substantial credit when it disconfirms the theoretical idea or when the proof is especially difficult or when the ideas of the proof are original, interesting and fruitful.

Daniel Friedan

There are no theorems in analysis—only proofs.

John B. Garnett

One must make a start in any line of research, and this beginning almost always has to be a very imperfect attempt, often unsuccessful. There are truths that are unknown in the way that there are countries the best road to which can only be learned after having tried them all. Some persons have to take the risk of getting off the track in order to show the right road to others. . . . We are almost always condemned to experience errors in order to arrive at truth.

Humble thyself, impotent reason!

Blaise Pascal

Intuition is glorious, but the heaven of mathematics requires much more. . . . In theological terms, we are not saved by faith alone, but by faith and works.

Saunders Mac Lane

Philip Anderson describes mathematical rigor as "irrelevant and impossible". I would soften the blow by calling it "beside the point and usually distracting, even where possible".

Benoit Mandelbrot

Thrice happy souls! to whom 'twas given to rise
To truths like these, and scale the spangled skies!
Far distant stars to clearest view they brought,
And girdled ether with their chains of thought.
So heaven is reached—not as of old they tried
By mountain piled on mountains in their pride.

Ovid

And this prayer I make,
Knowing that Nature never did betray
The heart that loved her ...

William Wordsworth

9.1 The Riemann Hypothesis

Bernhard Riemann (1826–1866) was one of the true geniuses of nineteenth century mathematics. He lived only until the age of 40, ultimately defeated by poverty and ill health. He struggled all his life, and only finally landed a regular professorship when he was on his deathbed. But the legacy of profound mathematics that Riemann left us continues to have a major influence in modern mathematics.

When Riemann was taking his oral exams at Göttingen, the estimable Carl Friedrich Gauss assigned him to speak about geometry. And even the impatient and haughty Gauss was impressed; for Riemann completely reinvented the subject. Taking into account the work of Bolyai (1802–1860) and Lobachevsky (1793–1856) on non-Euclidean geometry, Riemann offered a farseeing program for equipping surfaces and manifolds and even more general spaces with a geometry that retained the key features of the Euclidean geometry that we already know but also adapted itself to the particular features of the space on which it lived. Riemann's ideas about geometry are still studied intensively today.

Riemann's mathematical interests were broad. He gave the definition of the integral in calculus that is most commonly used today. He contributed

fundamental ideas to the theory of trigonometric series and Fourier series. And he was definitely interested in number theory. In his seminal paper *On the number of primes less than a given magnitude* [RIE], Riemann laid out some key ideas about the distribution of the prime numbers. Recall that a prime number is a positive, whole number whose only whole number divisors are 1 and itself. By custom we do not count 1 as a prime. The first several prime numbers are:

$$2\ ,\ 3\ ,\ 5\ ,\ 7\ ,\ 11\ ,\ 13\ ,\ 17\ ,\ 19\ ,\ 23\ ,\ 29\ ,\ 31\ \ldots\ .$$

The Fundamental Theorem of Arithmetic tells us that every positive integer can be written in a unique manner as the product of primes. As an instance,

$$17540 = 2^3 \cdot 3^2 \cdot 5 \cdot 7^2\,.$$

Obviously the primes are the building blocks for everything that we might want to know about the positive integers. The key ideas in modern cryptography are built on the primes. Many of the ideas in image compression and signal processing are founded in the primes. One of *the* most fundamental issues is how the primes are distributed.

When Gauss was a young man he studied tables of the prime numbers. These tables went on for pages and pages and contained thousands of prime numbers. From his studies, Gauss concluded that the following must be true: For n a positive integer, let $\pi(n)$ denote the number of primes less than or equal to n. For example, $\pi(50) = 15$, because the primes up to 50 are

$$2\ ,\ 3\ ,\ 5\ ,\ 7\ ,\ 11\ ,\ 13\ ,\ 17\ ,\ 19\ ,\ 23\ ,\ 29\ ,\ 31\ ,\ 37\ ,\ 41\ ,\ 43\ ,\ 47\,.$$

Also $\pi(100) = 26$ because, after 50, the primes are

$$53\ ,\ 59\ ,\ 61\ ,\ 67\ ,\ 71\ ,\ 73\ ,\ 79\ ,\ 83\ ,\ 89\ ,\ 91\ ,\ 97\,.$$

Then Gauss conjectured that, for large n, the value of $\pi(n)$ is about $n/\log n$. More precisely, a refined version of Gauss's conjecture would be that the limit of the quotient

$$\frac{\pi(n)}{n/\log n}$$

as n tends to infinity is 1.

Gauss was unable to prove this result, but he definitely believed it to be true. In point of fact, this so-called "Prime Number Theorem" was not proved

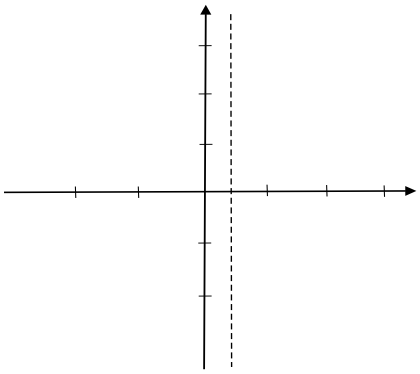


Figure 9.1. The critical line.

until 1896 (independently) by Jacques Hadamard (1865–1963) and Charles de la Vallée Poussin (1866–1962). Their proof was remarkable because it used complex analysis (a subject that is rather distant from number theory both in form and in style) in a profound way. Central to their study of the Prime Number Theorem was the notable zeta function of Bernhard Riemann.

In the aforementioned paper *On the number of primes less than a given magnitude*, Riemann introduced the zeta function. This is an analytic function of a complex variable that is defined by a particular infinite series. One of the first results that he proved about the zeta function was that it was intimately connected to the prime numbers. And Riemann *conjectured* that he knew the location of all the zeros (i.e., all the points of vanishing) of this zeta function. This was critical information for the study of the distribution of prime numbers. get this history straight. Look in Sabbagh and other sources The celebrated Riemann hypothesis, perhaps the most important unsolved problem in modern mathematics, concerns the location of the zeros of the Riemann zeta function. The conjecture is that, except for a collection of explicit and uninteresting zeros that were located at the negative integers by Riemann, all the other zeros (and there are infinitely many of these) are located on the critical line—a vertical line in the Cartesian plane located at $x = 1/2$. See Figure 9.1. G. H. Hardy (1877–1947) was able to show that infinitely many of these zeros are indeed on the critical line.

It is known that all the zeros of the zeta function—except for the trivial ones on the negative real axis that we noted above—lie in the critical strip, which is the set of complex numbers having real part between 0 and 1—refer to Figure 9.2. The issue is whether those zeros in the critical strip actually lie

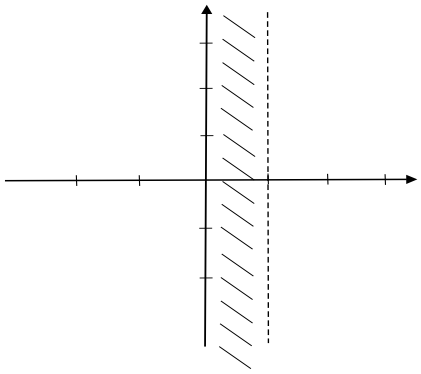


Figure 9.2. The critical strip.

on the critical line. Brian Conrey of the American Institute of Mathematics has proved[1] that at least 2/5 of the zeros of the Riemann zeta function lie on the critical line.

The Riemann hypothesis has a colorful history (the book [SAB], for example, is a great source for the detailed lore of the problem). Every major mathematician since the time of Riemann has spent time thinking about the Riemann hypothesis. In fact the *illuminati* refer to the problem simply as RH. It is possible that—some day—a mathematician will imitate Andrew Wiles (see Section 9.5) and give a lecture which he concludes by drawing a double arrow followed by "RH". That will be a dramatic day indeed.

One mathematician who had a real passion for RH was Godfrey Hardy. Every year he took a vacation to visit his friend Harald Bohr (1887–1951) in Denmark. The first thing they would do when they got together is sit down and write an agenda. And, always, the first thing to go on the agenda was "Prove the Riemann hypothesis." This is one agenda item that they were never able to bring to fruition.

G. H. Hardy, a lifelong bachelor, had many eccentricities. He was convinced that he and God had an ongoing feud.[2] One corollary of this belief is that Hardy believed that God had an agenda against Hardy. When Hardy

[1]Conrey built on earlier work that MIT mathematician Norman Levinson performed on his deathbed. Levinson at first thought that he could prove that 100% of zeros of the zeta function in the critical strip lie on the critical line. Then he had to modify his reasoning, and his claim, and make it 99%. Ultimately Levinson could prove that 33% of the zeros are on the critical line. Conrey's 2/5 or 40% is a distinct improvement, and seems to take Levinson's method to its logical limit.

[2]The truth is that Hardy was an atheist. But he frequently joked about God.

attended one of his beloved cricket matches he always brought an umbrella because he felt that this would guarantee that it would not rain: God would not give Hardy the satisfaction of being prepared for foul weather. One day Hardy had to make the passage across the English Channel—by boat of course—at a time of particularly rough seas. There was a genuine danger that he would not make it. As insurance, Hardy sent a postcard to Harald Bohr telling him that he had a proof of the Riemann hypothesis. This because Hardy was quite sure that God would not give him the satisfaction of going down with the ship and leaving the world to believe that he died with a proof of RH.

In modern times Louis de Branges (1932–)—of Bieberbach Conjecture fame (see Section 8.2)—has set out to prove the Riemann hypothesis. Worse, Louis has been up to his old tricks: He has announced more than once that he had a proof. He even, more than once, produced a manuscript. But more than that is true. On one of these occasions he submitted an elaborate proposal to the National Science Foundation to sponsor a conference celebrating his proof of the Riemann hypothesis. Of course all of de Branges's proofs have been found to be erroneous, and the conference was never funded.[3]

Unlike Fermat's last problem, which Andrew Wiles worked on in secret because he feared that people would think him on a fool's errand, the Riemann hypothesis is definitely mainstream mathematics. It is at the pinnacle of mathematical importance and influence, and even a partial result (such as Brian Conrey's dramatic theorem that we described above) is of considerable interest. In the year 1996, the American Institute of Mathematics sponsored a conference to discuss the current state of the problem. Many luminaries attended, including Fields Medalists Alain Connes (1947–) and Atle Selberg (1917–) and Paul Cohen (1924–). Some very frank views were exchanged, and in fact Connes claims that he came away with some important ideas. Some very nice papers by Connes resulted.

It must be stressed that, as of this writing, the Riemann hypothesis is still an open problem. And it cannot be said that we are within "shouting distance" of the proof. There are many encouraging partial results; but the end is nowhere in sight.

[3]To be fair, there is some evidence that the troublesome parts of de Branges's argument are amenable to the knowledge and talents of certain mathematicians in Israel. The unfortunate political situation in that part of the world makes it infeasible for de Branges to take advantage of their talents (as he did in 1984 with the Soviet mathematicians in St. Petersburg).

9.2 The Goldbach Conjecture

The Goldbach Conjecture is a tantalizing problem from eighteenth century mathematics. Christian Goldbach (1690–1764), a notable number theorist who conducted considerable correspondence with Leonhard Euler, asked in 1742 the following question:

> Can every even integer greater than 4 be written as the sum of two odd primes?

This is a genuine question, of real interest to serious mathematicians, that anyone can understand. And the evidence is all around us:

$$6 = 3 + 3\ ,\ 8 = 5 + 3\ ,\ 10 = 5 + 5\ ,\ 12 = 7 + 5\ ,\ 14 = 7 + 7\ ,$$

and so forth. Of course the assertion has been checked into the hundreds of millions using a computer. But, as we have stressed throughout this book, such immense evidence is definitely not the same thing as a mathematical proof. A mathematical proof is exhaustive, and covers *all* even integers greater than 4. A computer can only check finitely many of these.

This is a problem that frequently finds its way into the popular press. For it is readily accessible, and the entire world would appreciate the answer. Of course the fact of its immediacy and comprehensibility tempts all sorts of people to work on the problem—not just professional mathematicians. So all sorts of accounts of "solutions" have appeared in the newspapers, and they have all turned out to be incorrect.

The method that is usually used to attack the Goldbach conjecture is what we call the *sieve technique*. This tool finds its historic roots in the work of Eratosthenes (276 B.C.E.–194 B.C.E.). Eratosthenes was interested in finding all the prime numbers. They of course do not occur in any particular pattern, and they appear to be (and this would be one of the consequences of the Riemann hypothesis) randomly distributed in the integers.

The prime numbers are the units of arithmetic. A *prime number* is defined to be a positive whole number (i.e., an integer) that has no divisors except 1 and itself. By convention, we do not consider 1 to be a prime. Thus

- 2 is prime because the only divisors of 2 are 1 and 2.
- 3 is prime because the only divisors of 3 are 1 and 3.
- 4 is *not* prime because 2 divides 4.

- 5 is prime because the only divisors of 5 are 1 and 5.
- 6 is *not* prime because 3 divides 6.
- 7 is prime because the only divisors of 7 are 1 and 7.
- 8 is *not* prime because 2 divides 8.
- 9 is *not* prime because 3 divides 9.

and so forth. The *Fundamental Theorem of Arithmetic* states that every positive integer can be factored into primes in one and only one way. For example,

$$98 = 2 \cdot 7^2$$

and

$$12745656 = 2^3 \cdot 3^2 \cdot 7 \cdot 11^3 \cdot 19\,.$$

Except for rearranging the order of the factors, there is no other way to factor either of these two numbers.

The ancient Greeks had a particular fascination with primes. One such was Eratosthenes. Eratosthenes was born in Cyrene, Libya, North Africa. His teachers included Lysanias of Cyrene and Ariston of Chios. The latter made Eratosthenes part of the stoic school of philosophy. Around 240 B.C.E., Eratosthenes became the third librarian of the great library of Alexandria (this library was later destroyed by invading hordes). One of Eratosthenes's most important works was the *Platonicus*, a tract that dealt with the mathematics underlying Plato's *Republic*.

Eratosthenes devised a *sieve* for creating a list of the primes. In fact sieve methods are still used today to attack such celebrated problems as the Goldbach conjecture. Here is how Eratosthenes's method works.

We begin with an array of the positive integers:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64
65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96

...

First we cross out 1. Then we cross out all the multiples of 2 (but not 2 itself):

~~1~~	2	3	~~4~~	5	~~6~~	7	~~8~~	9	~~10~~	11	~~12~~	13	~~14~~	15	~~16~~
17	~~18~~	19	~~20~~	21	~~22~~	23	~~24~~	25	~~26~~	27	~~28~~	29	~~30~~	31	~~32~~
33	~~34~~	35	~~36~~	37	~~38~~	39	~~40~~	41	~~42~~	43	~~44~~	45	~~46~~	47	~~48~~
49	~~50~~	51	~~52~~	53	~~54~~	55	~~56~~	57	~~58~~	59	~~60~~	61	~~62~~	63	~~64~~
65	~~66~~	67	~~68~~	69	~~70~~	71	~~72~~	73	~~74~~	75	~~76~~	77	~~78~~	79	~~80~~
81	~~82~~	83	~~84~~	85	~~86~~	87	~~88~~	89	~~90~~	91	~~92~~	93	~~94~~	95	~~96~~

...

Now we proceed by crossing out all the multiples of 3 (but not 3 itself):

~~1~~	2	3	~~4~~	5	~~6~~	7	~~8~~	~~9~~	~~10~~	11	~~12~~	13	~~14~~	~~15~~	~~16~~
17	~~18~~	19	~~20~~	~~21~~	~~22~~	23	~~24~~	25	~~26~~	~~27~~	~~28~~	29	~~30~~	31	~~32~~
~~33~~	~~34~~	35	~~36~~	37	~~38~~	~~39~~	~~40~~	41	~~42~~	43	~~44~~	~~45~~	~~46~~	47	~~48~~
49	~~50~~	~~51~~	~~52~~	53	~~54~~	55	~~56~~	~~57~~	~~58~~	59	~~60~~	61	~~62~~	~~63~~	~~64~~
65	~~66~~	67	~~68~~	69	~~70~~	71	~~72~~	73	~~74~~	~~75~~	~~76~~	77	~~78~~	79	~~80~~
~~81~~	~~82~~	83	~~84~~	85	~~86~~	~~87~~	~~88~~	89	~~90~~	91	~~92~~	~~93~~	~~94~~	95	~~96~~

...

You can see that the numbers we are crossing out *cannot* be prime since, in the first instance, they are divisible by 2, and in the second instance, they are divisible by 3. Now we will cross out all the numbers that are divisible by 5 (why did we skip 4?) but not 5 itself. The result is:

~~1~~	2	3	~~4~~	5	~~6~~	7	~~8~~	~~9~~	~~10~~	11	~~12~~	13	~~14~~	~~15~~	~~16~~
17	~~18~~	19	~~20~~	~~21~~	~~22~~	23	~~24~~	~~25~~	~~26~~	~~27~~	~~28~~	29	~~30~~	31	~~32~~
~~33~~	~~34~~	~~35~~	~~36~~	37	~~38~~	~~39~~	~~40~~	41	~~42~~	43	~~44~~	~~45~~	~~46~~	47	~~48~~
49	~~50~~	~~51~~	~~52~~	53	~~54~~	~~55~~	~~56~~	~~57~~	~~58~~	59	~~60~~	61	~~62~~	~~63~~	~~64~~
~~65~~	~~66~~	67	~~68~~	~~69~~	~~70~~	71	~~72~~	73	~~74~~	~~75~~	~~76~~	77	~~78~~	79	~~80~~
~~81~~	~~82~~	83	~~84~~	~~85~~	~~86~~	~~87~~	~~88~~	89	~~90~~	91	~~92~~	~~93~~	~~94~~	~~95~~	~~96~~

...

Let us perform this procedure just one more time, by crossing out all multiples of 7 (why can we safely skip 6?), but not 7 itself:

~~1~~	2	3	~~4~~	5	~~6~~	7	~~8~~	~~9~~	~~10~~	11	~~12~~	13	~~14~~	~~15~~	~~16~~
17	~~18~~	19	~~20~~	~~21~~	~~22~~	23	~~24~~	~~25~~	~~26~~	~~27~~	~~28~~	29	~~30~~	31	~~32~~
~~33~~	~~34~~	~~35~~	~~36~~	37	~~38~~	~~39~~	~~40~~	41	~~42~~	43	~~44~~	~~45~~	~~46~~	47	~~48~~
~~49~~	~~50~~	~~51~~	~~52~~	53	~~54~~	~~55~~	~~56~~	~~57~~	~~58~~	59	~~60~~	61	~~62~~	~~63~~	~~64~~
~~65~~	~~66~~	67	~~68~~	~~69~~	~~70~~	71	~~72~~	73	~~74~~	~~75~~	~~76~~	~~77~~	~~78~~	79	~~80~~
~~81~~	~~82~~	83	~~84~~	~~85~~	~~86~~	~~87~~	~~88~~	89	~~90~~	~~91~~	~~92~~	~~93~~	~~94~~	~~95~~	~~96~~

...

And now here is the punchline: The numbers that remain (i.e., that are *not* crossed out) are those that are *not* multiples of 2, nor multiples of 3, nor multiples of 5, nor multiples of 7. In fact those that remain are not multiples of anything. They are the primes:

$$2\ ,\ 3\ ,\ 5\ ,\ 7\ ,\ 11\ ,\ 13\ ,\ 17\ ,\ 19\ ,\ 23\ ,\ 29\ ,\ 31\ ,\ 37\ ,\ 41\ ,\ 43\ ,$$

$$47\ ,\ 53\ ,\ 59\ ,\ 61\ ,\ 67\ ,\ 71\ ,\ 73\ ,\ 79\ ,\ 83\ ,\ 89\ldots$$

And on it goes. No prime was missed. The sieve of Eratosthenes will find them all.

But a number of interesting questions arise. We notice that our list contains a number of *prime pairs*: $\{3,5\}$, $\{5,7\}$, $\{11,13\}$, $\{17,19\}$, $\{29,31\}$, $\{41,43\}$, $\{71,73\}$. These are primes in sequence that differ by just 2. How many such pairs are there? Could there be infinitely many prime pairs? To date, nobody knows the answer to this question. Another old problem is whether the list of primes contains arbitrarily long arithmetic sequences, that is, sequences that are evenly spaced. For example, $3,5,7$ is a list of primes that is evenly spaced (by units of 2). Also $41,47,53,59$ is evenly spaced (by units of 6). It was only just proved in 2004 by Green and Tao that the primes *do contain* arbitrarily long arithmetic sequences.

An even more fundamental question is this: How many prime numbers are there altogether? Perhaps 100? Or 1000? Or 1,000,000? In fact it was Euclid (330–275 B.C.E.) who determined that there are infinitely many prime numbers. We have discussed his argument elsewhere in this book (Subsection 1.2.1).

So this is what a sieve technique is: A systematic method for crossing off elements of the positive integers in order to leave behind the numbers

that we seek. Many of the known attempts at the Goldbach conjecture have involved some sort of sieve technique. Although this problem is not as central or important as the Riemann hypothesis, it has attracted considerable attention. No less a vanguard than Brun, Rademacher, Vinogradov, Chen, Hua, Bombieri, and Iwaniec have worked on Goldbach and also on the twin prime conjecture (discussed in the next section). Perhaps what is most important about these problems is not the problems themselves, but rather the techniques that have been introduced to study them.

9.3 The Twin-Prime Conjecture

The twin-prime conjecture asks another tantalizing question that is accessible to anyone. We call two prime numbers "twin primes" if they occur in a row; that is to say, there is a difference of just 2 between them. Examples of twin primes are

$$\{3,5\}\ ,\ \{5,7\}\ ,\ \{11,13\}\ ,\ \{17,19\}\ ,\ \{29,31\}\ ,\ \{41,43\}\ \ldots.$$

Of course the twin primes become sparser and sparser as we move out into the larger integers. The question is: are there infinitely many twin primes?

Again, computers can be used to find twin primes that are quite far out in the number system. Thousands of twin primes have been found. And certainly the problem can be attacked using sieve methods; it has been subjected to such assaults on many occasions. It may be noted that Vinogradov has pioneered the method of trigonometric sums (based on ideas of Hardy and Littlewood) in the study of the twin primes problem.

The trouble with both the Goldbach conjecture and the twin-prime conjecture is that neither is very central to modern mathematics. They are more like historical footnotes. Still, it must be noted—once again—that a number of eminent mathematicians have worked on these problems. It is difficult to judge their true worth until we know that they are true and can spend some time with them.

Fermat's last theorem loomed large—in an obvious sort of way—because it spawned so much important mathematics. Ring theory, ideal theory, and many other critical ideas of modern abstract algebra grew out of attempts to prove Fermat's last theorem. The twin-prime conjecture has not been nearly so fecund, although it certainly has been the wellspring of some interesting research and some new ideas.

Certainly the mathematician who solves either of these old chestnuts (Goldbach or twin-prime) will achieve a certain amount of celebrity, and will reap certain rewards. And it is fitting that this will be so. But it will not be like proving the Riemann hypothesis or the Poincaré conjecture.

9.4 Stephen Wolfram and *A New Kind of Science*

Stephen Wolfram is one of the *Wunderkinder* of modern science. He earned his Ph.D. at Cal Tech at the tender age of 20. Just one year later he won a MacArthur Prize—the youngest recipient ever! Most people who win the MacArthur Prize—which is quite substantial (on the order of $500,000 or more)—just stick the money in the bank and stare at it. But Wolfram is quite the enterprising fellow, and he started a company (now called Wolfram Research) and developed one of the first computer algebra systems for the personal computer. Known as `Mathematica`, this product has really changed the landscape for mathematical scientists. Whereas formerly one used a computer to do strictly *numerical* calculations, now the computer could solve algebra problems, solve differential equations, calculate integrals, calculate derivatives, manipulate matrices, and perform many other basic mathematical operations. `Mathematica` can draw marvelous graphs and diagrams. Today there are many thousands of mathematicians who depend on `Mathematica` as a basic tool in their research. And, concomitantly, Stephen Wolfram is a wealthy man.

As already noted, Wolfram is an estimable scientist. And a man with diverse talents. When he was on the faculty at the University of Illinois, he was simultaneously a Professor of Computer Science, Mathematics, and Physics. A feat hardly ever equalled in the annals of academe. One of his major contributions to mathematical science has been to give a prominent berth to the theory of cellular automata. Just what is this animal? A cellular automaton is a computer system that begins with an array of boxes on a sheet of graph paper and a set of rules by which these boxes might evolve. For example, one might begin with a single interval as shown in Figure 9.3. Note that the key box is shaded, and the other boxes shown are *un*shaded. Then the rule might be as depicted in Figure 9.4.[4] The way to read this rule is as

[4]This is in fact Wolfram's Rule 90—see [WOL].

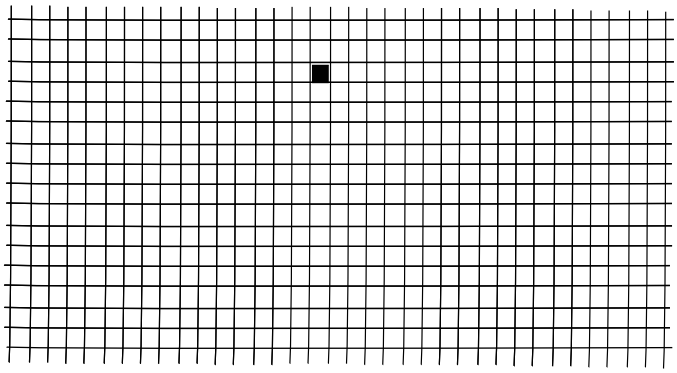


Figure 9.3. Beginning of a cellular automaton.

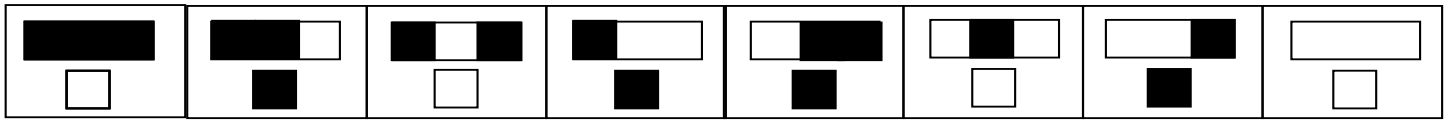


Figure 9.4. Wolfram's cellular automaton Rule 90.

follows. The first box shows a particular triple of boxes: they are all shaded. Whenever we encounter such a triple then the middle box is to be made blank. The second box shows a new triple in which the first two boxes are shaded and the third is not. Whenever we encounter such a triple then the middle box is to be shaded. And so forth. Note that, when we are applying this analysis to the N^{th} row on the graph paper, we record the result in the $(N+1)^{\text{th}}$ row.

Applying that rule to our initial configuration in Figure 9.3 (parts 7, 6, and 4 of the rule), we obtain Figure 9.5. Now apply the rule again to obtain Figure 9.6. And yet once more to obtain 9.7. And so forth.

The remarkable thing about a cellular automaton is that one can begin with a very simple configuration and a very simple rule of progeneration and end up, after a good many iterations of the rule, with something fantastically complicated. Figures 9.8, 9.9, 9.10 exhibits some configurations that were generated by fairly simple cellular phenomena.

Over time, Wolfram has become obsessed with cellular automata. He is convinced that they are a model for how nature works. From the spots on a leopard to the design of a snowflake to the structure of the human brain, Wolfram believes that there is a cellular automaton that encodes the design

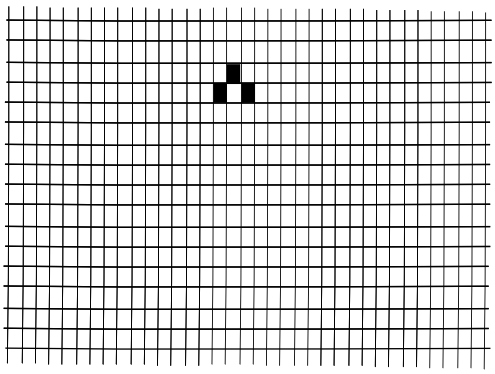


Figure 9.5. First iteration of Rule 90.

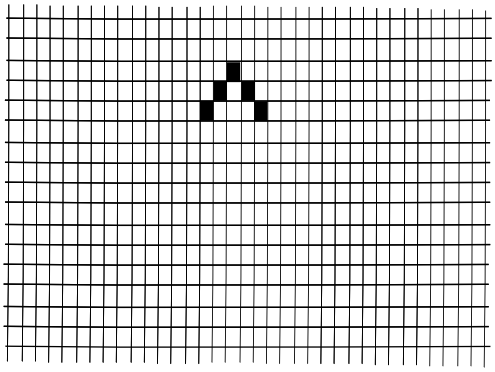


Figure 9.6. Second iteration of Rule 90.

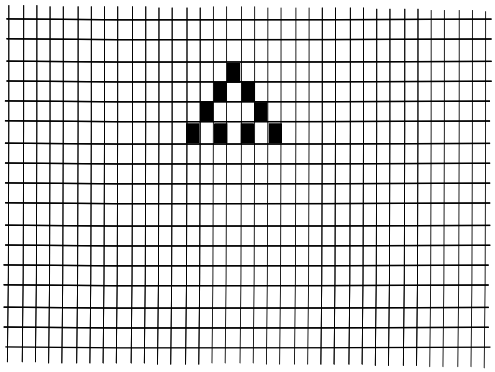


Figure 9.7. Third iteration of Rule 90.

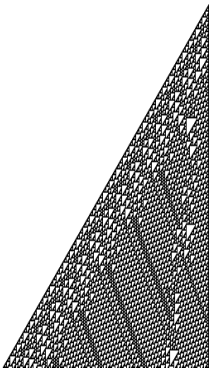


Figure 9.8. First cellular automaton.

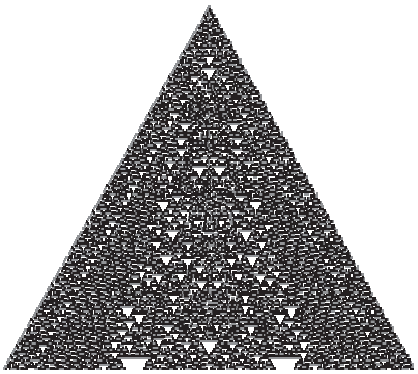


Figure 9.9. Second cellular automaton.

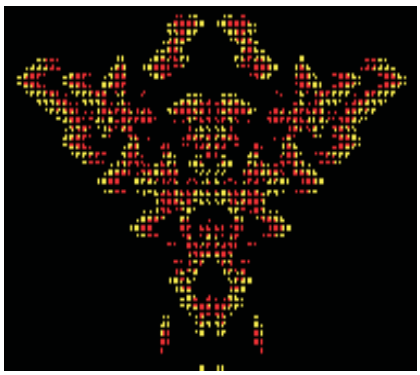


Figure 9.10. Third cellular automaton.

of each. He has written a 1286-page book [WOL] that explains his theory in some detail. The book has garnered quite a lot of attention—see the review [KRA6] as well as the Web site

`http://www.math.usf.edu/~eclark/ANKOS_reviews.html`

But *A New Kind of Science* is *not* Newton's *Principia* and it is not Darwin's *Origin of the Species*. Whatever its merits may be, this book has not changed anyone's view of the world. It has not changed any school curriculum (although Wolfram has written that it certainly will). It has not convinced anyone of anything.

Given the message of the present book, and given that the theory of cellular automata is certainly mathematical science, the subject merits some discussion in these pages. Wolfram is a physicist by training. He is not bound to the strict rigor of proofs as we have discussed it here. But he knows a heck of a lot of mathematics. He certainly knows what our standards are; and he is well acquainted with the standards in the world of theoretical physics. Unfortunately, the book [WOL] does not meet any of these standards. How could this be, and why would Wolfram commit such a gaffe?

What Wolfram professes in the pages of [WOL] is that he is writing this book for a popular audience. He thinks that his ideas are so important that he must leapfrog over the usual jury of his scientific peers and go directly to the populace at large. Unfortunately, one result of this decision is that he cannot write with any rigor or sophistication. He cannot indulge in any precise reasoning. He cannot access and utilize the scientific literature. The bottom line is then a panorama of rather vague discussion that is, to be generous, quite inconclusive. The considerations in [WOL] are largely phenomenological, and they are *not* convincing.

Wolfram's book contains a great many calculations and a great many pictures and a great many *descriptions*. But it has very little *scientific discourse*. Because Wolfram wants to bypass the usual jury of scientific reviewers, he must express himself in a language that is comprehensible to virtually anyone. And that imposes severe limitations on what he can do and what he can say. He cannot assume that his readers know the basic tenets of physics; he cannot assume that they know relativity or quantum mechanics or string theory. He cannot even assume that they know calculus (which is, after all, the bedrock of modern physics). The upshot is that he can only dance around the key ideas. He cannot truly address any of them.

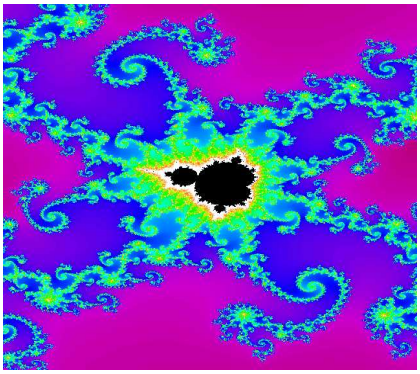


Figure 9.11. First fractal.

And that is the lesson for the readers of this book. Writing and reading strict logical proofs, or any kind of rigorous reasoning, is hard work. But, in the end, the hard work pays off. For a formal proof is comprehensible to anyone with the proper training. And it travels well. It will be just as valid, and just as comprehensible, in 100 years as it is today. Wolfram's ideas garnered some short-term attention. But they are already fading into the sands of time.

9.5 Benoit Mandelbrot and Fractals

Certainly one of the phenomena of modern mathematics is the development of fractal theory. Spawned by Benoit Mandelbrot (1924–) in 1982, fractal geometry is touted to be a new blueprint for understanding the world around us. Some pictures of fractals are shown in Figures 9.11, 9.12, and 9.13.

Mandelbrot begins his famous book [MAN1] by pointing out that

> Clouds are not spheres, mountains are not cones, coastlines are not circles, and bark is not smooth, nor does lightning travel in a straight line.

Instead, he posits, the objects that occur in nature are very complicated. Typically they have very wiggly boundaries. The geometric phenomenon on which he focuses is an object such that, if you take a piece of it and blow it up (i.e., dilate it), it looks the same. See Figure 9.14.

Fractal geometry is now a huge industry, and there are those who use fractals to create denoising filters, to implement image-compression algorithms,

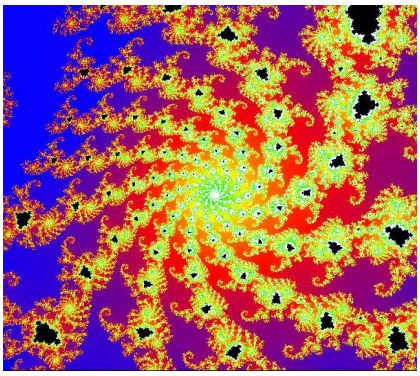


Figure 9.12. Second fractal.

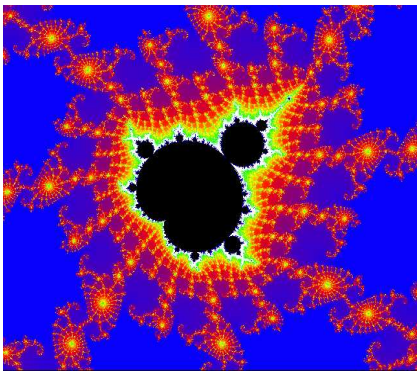


Figure 9.13. The Mandelbrot set.

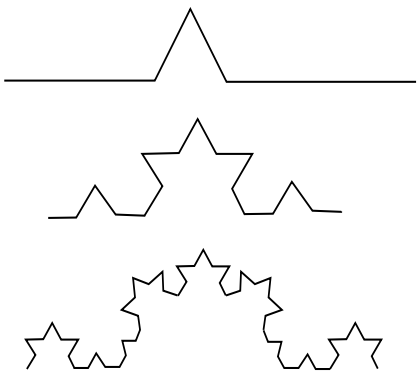


Figure 9.14. Scalability of fractals.

and to analyze investment strategies.

Mandelbrot claims not only to have invented a new branch of mathematics (in point of fact he did not actually invent the Mandelbrot set—it appears earlier in [BRM]). He also asserts that he has created a new way of *doing* mathematics. He has been known to deride traditional mathematicians who do mathematics in the traditional fashion (see [MAN2]). Mandelbrot's new methodology, reminiscent of Wolfram's study of cellular automata, is phenomenological. The fractal geometer typically does not enunciate theorems and prove them. Instead he/she generates computer graphics and describes them and endeavors to draw conclusions.

One can easily imagine that fractal geometry is extremely popular among mathematical amateurs. High school teachers like fractals because they feel that they can show their students some cutting-edge mathematics—just by showing them pictures of fractals. For a number of years Mandelbrot (an employee of IBM) was the research *gestalt* of International Business Machines. Their primetime television ads proudly featured Mandelbrot and pictures of fractals. And IBM rewarded Mandelbrot accordingly. The company used its financial clout to get Mandelbrot a teaching position at Yale University and the prestigious Barnard Prize, among other encomia.

It must be said that there are many branches of mathematics and science in which fractals are considered to be extremely useful. Physicists are always looking for new ways to model nature, and fractals give them a new language for formulating the laws of our world. More than half the submissions to some of the major physics journals concern fractals. Statisticians in Australia use fractals to model semi-permeable membranes. Mandelbrot himself used fractal ideas to help AT&T with a denoising problem.

Fractal geometry is both a new way to conceive of shape and form and also a new way to think about mathematics. It does *not* conform to the Euclidean paradigm of theorem and proof. It in fact epitomizes a new experimental approach to mathematics that is complimentary to the more familiar and traditional methodologies.

9.6 Roger Penrose and *The Emperor's New Mind*

In the mid-1980s, distinguished British physicist Roger Penrose of Oxford University was watching the "telly"—absorbing a show about artificial intelligence. The show seemed to be asserting that soon machines will be able

to think just as human beings do. These assertions pushed all of Roger Penrose's button—all the *wrong* buttons. He went marginally cataleptic, set aside all his research projects, and decided that he had to write a book refuting these claims. The result was *The Emperor's New Mind* [PEN1], a book that has had a profound impact on modern thinking. It was a bestseller to boot, in spite of the fact that it is so profoundly technical that there are few university professors who can read it straight through.

Penrose is a remarkable scholar. His advanced training is in mathematics, and he has made profound contributions to modern mathematical thought. But his primary work for the past several decades has been in theoretical physics. He is a close associate of Stephen Hawking. Penrose's expertise spans several academic disciplines. And his book reveals his erudition in considerable detail.

To oversimplify his thesis, Penrose says in his book that a computer can only execute a procedure that is given by an algorithm—typically a *computer program*. But a mathematician, for example, finds mathematical truths by way of insight, intuition, and sometimes even leaps of faith. These mental operations *cannot* be achieved nor implemented by means of an algorithm. Penrose might have gone further to note that the creation of artwork—whether it be poetry or music or literature or sculpture or dance or painting—is not generally achieved by way of an algorithm. It is generally done heuristically, with a basis in the artist's life experience.

The intellectual, scholarly embodiment of the thesis that machines can, or some day will be able to, think is the subject of "Artificial Intelligence" or AI. We have only enjoyed AI at our universities for forty years or so—thanks largely to vigorous support from the military establishment. It is easy to see why military leaders, those who think about armaments and warfare, would like to have a machine that can look down the road and tell what is coming. They would like to have a machine that can evaluate a battle situation and determine how to deploy troops or armaments. They would like to have a machine that can produce optimal bombing strategies. For a good many years (stretching back to the days of Sir Walter Raleigh), the aiming and disposition of artillery has been one of the prime motivators for various mathematical calculations, for the construction of a variety of computing machines, and for the development of certain branches of mathematics (such as spherical trigonometry and geodesy). Now we are in a much more sophisticated age, and the demands of technology are considerably more recondite.

It can be argued, and is frequently proposed by the skeptics and the nay-

sayers, that AI has achieved little. To be sure, we now have some rudimentary robots. We have a machine that can walk through the room without bumping into the furniture. We have a "robot" that can vacuum your living room or mow your lawn. Much modern manufacturing—the sort of rote work that used to be performed by assembly-line workers—is now done by robots. But there is no robot that can tie a shoe.

The field of AI has spawned some other fields that play a notable role in modern theoretical computer science; "Expert Systems" is one of these. But there are no machines today, nor will there be any in the foreseeable future, that can *think*.

The discourse in which Roger Penrose engages is quite tricky. For, reasoning like a mathematician, one must decide in advance what it means to "think". Questions about Gödel's incompleteness theorem and Turing machines are natural to consider. Church's thesis, a venerable tenet of modern logic, is that any effectively computable function is a recursive function. What does this mean? A *recursive function*—this is an important idea of Kurt Gödel—is a function that is built up from very basic steps—steps that a machine can perform. There is a technical, very precise, mathematical definition of recursive function that can be found in textbooks—see [WOLF] as well as [KRA4] and references therein. It is rather more difficult to say what an effectively computable function is—much of the modern discourse on Church's thesis concerns trying to come up with the right definition of this concept. Roughly speaking, an *effectively computable function* is one that a machine can calculate. The intuition behind Church's thesis is that any "effectively computable function" must be one that can be broken down into simple, algorithmic steps. Hence it must be recursive.

One approach to the questions being considered here is that any effectively computable function, or any procedure prescribed by a Turing machine, is an instance of human thought. And it is clear by their very definitions that these are in fact operations that a computer could perform. But Penrose would argue that this is an extremely limited perception of what human thought actually is. Beethoven's Ninth Symphony, *War and Peace*, even the invention of the computer itself, would never have been achieved by way of such mechanized or precedural thinking. Creative thought does not proceed from algorithms.

It is also natural to consider Penrose's concerns from the point of view of the Gödel incompleteness theorem. In any logical system that is at least as complex as arithmetic, there will be true statements that we cannot prove

(inside the system). Thus a computer will always be limited, no matter what language or logical system it is using, in what it can achieve. A human being can deal with the "Gödel sentence" heuristically, or by amassing evidence, or by offering a plausibility or perhaps a probabilistic argument. The human being can step outside the logical system and exploit whatever tools may be needed to obtain a proof. The computer cannot deal with the problem at all. Computer scientists (see [MCC]) like to point out that the programming language `Lisp` is very good at handling the Gödel phenomenon. And we can also construct systems that will simply avoid Gödel sentences. Nevertheless, Gödel's incompleteness argument carries genuine philosophical weight.

It is easy to imagine that many classically trained, theoretical mathematicians are more than anxious to embrace Penrose's ideas. There are few of us who want to believe that we shall some day be supplanted by a room full of machines running on silicon chips. But an equally enthusiast cadre from the computer science community are quick to oppose Penrose. The rather articulate and entertaining article [MCC] is an instance of the kind of reasoning that was aligned against Penrose.

The AI community continues to flourish—MIT and Cal Tech, for example, have vigorous artificial intelligence groups. There are obviously many interesting questions to consider in this discipline. But it is noteworthy to attend a gathering of AI investigators and to see how much of their time they spend arguing over definitions. As mathematicians, we of course appreciate the importance of laying down the right definitions so that everything else will follow rigorously therefrom. But we only allot a finite amount of time to that effort and then we move on. The subject of artificial intelligence has a more organic component to it—after all, it is an effort to apply mathematical principles to the way that the human mind functions—and therefore it is subject to forces and pressures from which theoretical mathematics is immune. Roger Penrose certainly touched a nerve with his thinking about the efforts of the AI community. His ideas continue to reverberate throughout the mathematical sciences.

9.7 The P/NP Problem

Complexity theory is a device for measuring how complicated a problem is (from a computational point of view), or how much computer time it will take to solve it. As this book has made clear, computation (and the theory

of computation) looms large in modern mathematical science. Of particular interest are which problems can be solved in a reasonable amount of time, and which will require an inordinate or impractical amount of time. Problems of the latter sort are termed *computationally expensive*. Problems of the former type are called *feasible*. Just as an instance, the problem of factoring a whole number with 200 digits is computationally expensive. This could actually take years, even on a fast digital computer. But if instead I say to you

> Here is a 200-digit number N. I claim that these two 100-digit numbers p and q are its prime factors. Please verify this assertion.

then all you need to do is to multiply p and q together. This will just take a few seconds (in fact typing in the numbers is the part of the problem that takes longest).

The issues being considered here are part of the **NP**-Completeness problem. Many people consider this to be the most important problem in mathematics and theoretical computer science. It has far-reaching consequences for what we can do with computers, or what we can hope to do in the future. Many of the fundamental ideas in cryptography depend on questions of computational complexity. We shall discuss some of these ideas in the present section.

9.7.1 The Complexity of a Problem

Complexity theory is a means of measuring how complicated it is, or how much computer time it will take, to solve a problem. We measure complexity theory in the following way: Suppose that the formulation of an instance of a problem involves n pieces of data. Then how many steps[5] will it take (as a function of n) to solve the problem? Can we obtain an effective bound on that number of steps, that is valid for asymptotically large values of n?

Consider dropping n playing cards on the floor. Your job is to put them back in order. How many steps will this take?

In at most n steps (just by examining each card), you can locate the first card. In at most another $(n-1)$ steps, you can locate the second card. In at most another $(n-2)$ steps, you can locate the third card, and so forth.

[5]Here a "step" is an elementary action that a computer or robot could take. It should be a "no brainer".

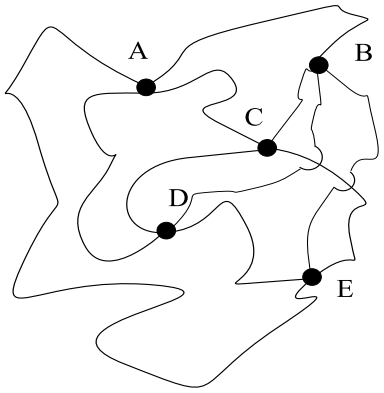


Figure 9.15

In summary, it will require at most

$$n + (n-1) + (n-2) + \cdots + 1 = \frac{n(n+1)}{2}$$

steps to put the deck of cards back in order (this summation formula is usually attributed to Carl Friedrich Gauss (1777–1855)). Because the expression $n[n+1]/2 \leq 2n^2$, we say that this problem has *polynomial complexity of degree (at most) 2*.

Now contrast that first example with the celebrated Traveling Salesman Problem: There are given n cities and there is a road of known length connecting each pair (see Figure 9.15). The problem is to find the shortest route that will enable the salesman to visit each city and to return to his starting place.

On the level of effective computability we see that a search for the solution of the Traveling Salesman problem amounts to examining each possible ordering of cities (because clearly the salesman can visit the cities in any order). There are $n! = n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot 3 \cdot 2 \cdot 1$ such orderings. According to Stirling's formula,

$$n! \approx \left(\frac{n}{e}\right)^n \cdot \sqrt{2\pi n}.$$

Thus the problem cannot be solved (in the obvious way) in a polynomial number of steps. We say that the problem (potentially) has *exponential complexity*. It actually requires some additional arguments to establish this fact.

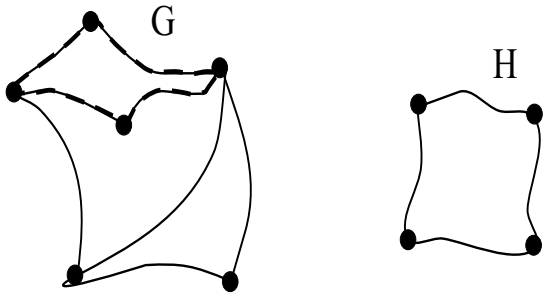


Figure 9.16

Another famous problem that is exponentially complex is the *subgraph problem*. Let G be a graph; that is, G is a collection of vertices together with certain edges that connect certain pairs of the vertices. Let H be another graph. The question is whether G contains a sub-graph that is isomorphic to H (Figure 9.16). (Here two graphs are isomorphic if there is a combinatorial mapping matching up vertices and edges.)

9.7.2 Comparing Polynomial and Exponential Complexity

From the point of view of theoretical computer science, it is a matter of considerable interest to know whether a problem is of polynomial or of exponential complexity, for this information gives an indication of how computationally expensive a certain procedure will be. In the area of computer games (for example), this will translate into whether a certain action can be executed with realistic speed, or sluggishly.

In the discussion that follows, we restrict attention to the so-called "decision problems." These are problems with yes/no answers. An example of a problem that is *not* a decision problem is an optimization problem (e.g., find the configuration that maximizes something). But in fact most optimization problems can be converted to decision problems by introducing an auxiliary parameter that plays the role of an upper bound. It is not a severe restriction to treat only decision problems, and it makes the exposition much cleaner.

9.7.3 Polynomial Complexity

A problem is said to be of "Class **P**" if there is a polynomial p and a (deterministic) Turing machine (DTM) for which every input of length n comes to a halt, with a yes/no answer, after at most $p(n)$ steps (see Section 6.1 for a discussion of Turing machines). The word "deterministic" is used here to denote an effectively computable process, with no guessing.

Problems of Class **P** are considered to be tractable. A problem that is not of class **P**—that is, for which there is no polynomial solution algorithm—is by definition *intractable*. Class **P** problems give solutions in a reasonable amount of time, and *they always give solutions*. It will never happen that the machine runs forever. The problem of ordering a deck of cards (already discussed), the problem of finding one particular marble in a jar of marbles, and the problem of matching up husbands and wives at a party are all problems of Class **P**.

9.7.4 Assertions that Can Be Verified in Polynomial Time

More significant for the theoretical development of this chapter is that, for certain problems that are otherwise intractable, the *verification of a solution* can be a problem of Class **P**. We describe the details of this assertion below.

For example, the problem of finding the prime factorization of a given natural number N with n digits is believed to be of exponential complexity (see [SCL]). More precisely, the complexity—according to the best currently known algorithm—is about of size $10^{\sqrt{n \ln n}}$, which means that the computation would take about that many steps. But the verification procedure is of polynomial complexity: If N is given and a putative factorization $p_1, \ldots, p_k$ is given, then it is obviously (just by inspection of the rules of arithmetic) a polynomial time problem to calculate $p_1 \cdot p_2 \cdots p_k$ and verify (or not) that it equals N.

The problem of factoring a large integer is believed by all the experts to be of exponential complexity (that is, if the integer has n digits then it will take on the order of 2^n steps to find the prime factorization).[6] The

[6]It is an amazing recent result of Agrawal and his group [AGA] that there is a polynomial time algorithm for determining whether a positive integer is prime or composite. But Agarwal's method will *not* produce the prime factorization. It only provides a "yes" or "no" answer.

celebrated RSA encryption method—one of the cutting-edge techniques in modern cryptography—is premised on this hypothesis. That is to say, if we could find a way to factor large integers in polynomial time then RSA-encrypted messages could be rapidly decrypted.[7] At this time it is not known whether the factorization of large integers is a polynomial time problem. The best algorithms that we have are exponential time algorithms.

Likewise, the subgraph problem is known to be of exponential complexity. But, if one is given a graph G, a subgraph K, and another graph H, then it is a problem of only polynomial complexity to confirm (or not) that H is graph-theoretically isomorphic to K.

The considerations in the last three paragraphs will play a decisive role in our development of the concept of problems of "Class **NP**."

9.7.5 Nondeterministic Turing Machines

A nondeterministic Turing machine (NDTM) is a Turing machine with an extra write-only head that generates an initial guess for the Turing machine to evaluate (see [GAJ] for a rigorous definition of the concept of nondeterministic Turing machine). We say that a problem is of Class **NP** if there is a polynomial p and a nondeterministic Turing machine with the property that, for any instance of the problem, the Turing machine will generate a guess of some length n and come to a halt, generating an answer of "yes" or "no" (i.e., that this guess *is* a solution or *is not* a solution) in at most $p(n)$ steps.

It first should be noted that $\mathbf{P} \subseteq \mathbf{NP}$ (that is, every problem of class **P** is also of class **NP**). This assertion is obvious, because if Π is a problem of class **P**, then we can let the guess be vacuous to see that Π is then of class **NP**. It is natural, therefore, to consider $\mathbf{NP} \setminus \mathbf{P}$ (i.e., the problems that are of class **NP** but not of class **P**).

The first interesting question in this subject is whether $\mathbf{NP} \setminus \mathbf{P}$ is nonempty. If it is, then any problem in that set-theoretic difference has the property that it is *not* deterministically of polynomial complexity, but if it is given a guess for a solution, then it can evaluate that guess in polynomial time. Any problem that can be established to lie in this set-theoretic difference will be considered to be intractable.

It turns out that it is quite difficult to determine whether $\mathbf{NP} \setminus \mathbf{P}$ is

[7] The amusing film *Sneakers*, starring Robert Redford, is in fact precisely about such an eventuality.

nonempty. Thus far, no such problem has been identified. So some alternative questions have been formulated; these seem to be more within our reach. In particular, there are relatively straightforward proof techniques for addressing these alternative formulations.

9.7.6 Foundations of NP-Completeness

The first of these separate questions, which is the foundational question of **NP**-completeness, is the following. Suppose that we are given a problem Π. Can we establish the following syllogism?

$$\text{If } \mathbf{NP} \setminus \mathbf{P} \text{ is nonempty (i.e., } \mathbf{P} \neq \mathbf{NP}\text{), then } \Pi \in \mathbf{NP} \setminus \mathbf{P}.$$

See [GAJ] for sample problems that may be addressed using this slightly modified problem.

The most important substitute question is that of **NP**-completeness. We shall address it in the Section 9.6.8.

9.7.7 Polynomial Equivalence

We say that two problems Π_1 and Π_2 are *polynomially equivalent* if there is a polynomially complex translation of the language in which Π_1 is expressed into the language in which Π_2 is expressed and vice versa. That is to say, there is a polynomial q such that any statement about Π_1 with n characters can be translated into a statement about Π_2 with at most $q(n)$ characters and conversely.

9.7.8 Definition of NP-Completeness

Let Π be a problem in **NP**. We say that Π is **NP**-*Complete* if Π is polynomially equivalent to every other problem in **NP**. Now suppose that Π is an **NP**-complete problem. If it turns out that Π can be solved with a polynomial time algorithm, then it follows that every other problem in **NP** can be solved with a polynomial time algorithm. On the other hand, if it turns out that *any* problem in **NP** is intractable (i.e., lies in $\mathbf{NP} \setminus \mathbf{P}$), then Π is intractable. The **NP**-complete problems are considered to be the hardest problems in **NP**.

9.7.9 Intractable Problems and NP-Complete Problems

Although we do not know whether **NP** \ **P** is nonempty (i.e., whether there are any intractable problems), we do know—and can explicitly identify—a great many **NP**-complete problems. It is considered to be one of the great unsolved problems in theoretical computer science and mathematics to determine whether **NP**-complete problems are intractable. Thus far, little substantial progress has been made on the question.

9.7.10 Examples of NP-Complete Problems

Here we enumerate a few examples, from several different branches of mathematics and theoretical computer science, of problems that are known to be **NP**-complete. Bear in mind that if any of these problems is determined to be intractable, then they all are. Our source for this material is the Appendix of [GAJ].

A Problem from Diophantine Equations

A *Diophantine equation* is a polynomial equation with integer coefficients for which we seek integer solutions. Fermat's last problem, discussed elsewhere in this book, is a Diophantine equation. The *quadratic Diophantine problem* is as follows.

Q: Let a, b, c be integers. Does the equation

$$ax^2 + by = c$$

have a pair (x, y) of integer solutions?

Reference: [WOLF].

A Problem from Graph Theory

In what follows, if G is a graph, we write $G = (V, E)$, where V denotes the collection of vertices and E denotes the collection of edges of the graph. The notation $|V|$ denotes the number of vertices in the graph. A graph is said to be *connected* if it is not the disjoint union of two subgraphs.

Complete Subgraphs: The *complete graph on k vertices* is a graph with k vertices such that *every* pair of vertices is connected by an edge. Thus the complete graph on k vertices has $\binom{k}{2} = \frac{k(k-1)}{2}$ edges.

> **Q**: Given a graph $G = (V, E)$ and a natural number $k \leq |V|$, does G contain a subgraph that is isomorphic to the complete graph on k vertices?

Reference: [COO].

Problems from the Theory of Sets and Partitions

Set Packing Problem:

> **Q**: Let $\mathcal{C}$ be a collection of finite sets and k a natural number with $k \leq |\mathcal{C}|$. Here, as is standard, $|\mathcal{C}|$ denotes the number of elements in $\mathcal{C}$. Does $\mathcal{C}$ contain at least k mutually disjoint sets?

Reference: [KAR].

Storage and Retrieval Problems

Bin Packing Problem:

> **Q**: Let U be a finite set of items. Let $s : U \to \mathbb{N}$ be a function that assigns a size to each element of U. Let $B, k \in \mathbb{N}$. Is there a partition $U = U_1 \cup \cdots \cup U_k$ such that $\sum_{u \in U_j} s(u) \leq B$ for each j?

References: [DAN], [HOS], and [LAW].

9.8 Andrew Wiles and Fermat's Last Theorem

Pierre de Fermat (1601-1665) was one of the most remarkable mathematicians who ever lived. He spent his entire adult life as a magistrate or judge in the city of Toulouse, France. His career was marked by prudence, honesty, and scrupulous fairness. He led a quiet and productive life. His special

passion was for mathematics. Fermat was perhaps the most talented amateur mathematician in history.

Fermat is remembered today by a large statue that is in the basement of the Hôtel de Ville in Toulouse. The statue depicts Fermat, dressed in formal attire, and seated. There is a sign, etched in stone and part of the statue, that says, "Fermat, the father of differential calculus." Seated in Fermat's lap is a naked muse showing her ample appreciation for Fermat's mental powers.

Pierre Fermat had a brother and two sisters and was almost certainly brought up in the town (Beaumont-de-Lomagne) of his birth. Although there is little evidence concerning his school education, it must have been at the local Franciscan monastery.

He attended the University of Toulouse before moving to Bordeaux in the second half of the 1620s. In Bordeaux he began his first serious mathematical researches and in 1629 he gave a copy of his restoration of Apollonius's *Plane loci* to one of the mathematicians there. Certainly in Bordeaux he was in contact with Beaugrand and during this time he produced important work on maxima and minima which he gave to Étienne d'Espagnet, who clearly shared mathematical interests with Fermat.

From Bordeaux, Fermat went to Orléans where he studied law at the University. He received a degree in civil law and he purchased the offices of councillor at the parliament in Toulouse. So by 1631 Fermat was a lawyer and government official in Toulouse and because of the office he now held he became entitled to change his name from Pierre Fermat to Pierre de Fermat.

For the remainder of his life he lived in Toulouse but, as well as working there, he also worked in his home town of Beaumont-de-Lomagne and a nearby town of Castres. The plague struck the region in the early 1650s; as a result, many of the older men died. Fermat himself was struck down by the plague and in 1653 his death was wrongly reported, then corrected:

> I informed you earlier of the death of Fermat. He is alive, and we no longer fear for his health, even though we had counted him among the dead a short time ago.

The period from 1643 to 1654 was one when Fermat was out of touch with his scientific colleagues in Paris. There are a number of reasons for this. Firstly pressure of work kept him from devoting so much time to mathematics. Secondly the Fronde, a civil war in France, took place and from 1648 Toulouse was greatly affected. Finally there was the plague of 1651 which

must have had great consequences both on life in Toulouse and of course its near fatal consequences on Fermat himself. However it was during this time that Fermat worked on the theory of numbers.

Fermat is best remembered for his work in number theory, in particular for Fermat's Last Theorem. This theorem states that the equation

$$x^n + y^n = z^n$$

has no non-zero integer solutions x, y and z when the integer exponent $n > 2$. Fermat wrote, in the margin of Bachet's translation of Diophantus's *Arithmetica*,

> I have discovered a truly remarkable proof which this margin is too small to contain.

These marginal notes only became known after Fermat's death, when his son Samuel published an edition of Bachet's translation of Diophantus's *Arithmetica* with his father's notes in 1670.

It is now believed that Fermat's "proof" was wrong although it is impossible to be completely certain. The truth of Fermat's assertion was proved in June, 1993 by the British mathematician Andrew Wiles at Princeton University, but Wiles withdrew the claim when problems emerged later in 1993. In November, 1994 Wiles again claimed to have a correct proof which has now been accepted. Unsuccessful attempts to prove the theorem over a 300 year period led to the discovery of commutative ring theory and a wealth of other mathematical developments.

Fermat's correspondence with the Paris mathematicians restarted in 1654 when Blaise Pascal, Étienne Pascal's son, wrote to him to ask for confirmation about his ideas on probability. Blaise Pascal knew of Fermat through his father, who had died three years before, and was well aware of Fermat's outstanding mathematical abilities. Their short correspondence set up the theory of probability and from this they are now regarded as joint founders of the subject.

It was Fermat's habit to solve problems and then pose them (without proof or solution) to the community of mathematicians. Some of these were quite deep and difficult, and people found them aggravating. One problem that he posed was that the sum of two cubes cannot be a cube (a special case of Fermat's Last Theorem which may indicate that by this time Fermat realised that his proof of the general result was incorrect), that there are

exactly two integer solutions of $x^2+4 = y^3$, and that the equation $x^2+2 = y^3$ has only one integer solution. He posed certain problems directly to the English—just as a cross-cultural challenge. The mathematical community failed to see that Fermat had been hoping his specific problems would lead them to discover, as he had done, deeper theoretical results.

Fermat has been described by some historical scholars as

> Secretive and taciturn, he did not like to talk about himself and was loath to reveal too much about his thinking. ... His thought, however original or novel, operated within a range of possibilities limited by that time [1600 -1650] and that place [France].

Fermat's last problem (or last theorem) is one of the grand old ladies of mathematics. It has been considered to be a problem of central and lasting importance ever since Pierre de Fermat wrote his famous marginal comment. Many an excellent mathematician—from Sophie Germain (1776–1831) to Leonhard Euler (1707-1783) to Emil Artin (1898–1962)—worked on the problem and obtained important partial results. It was a tremendous event when, in 1993, Princeton mathematician Andrew Wiles announced that he had a proof of Fermat's Last Theorem. In fact he did so in a way that was uncharacteristically melodramatic for a mathematician.

The Isaac Newton Institute for Mathematics was constructed at Cambridge University in 1992. Of course Isaac Newton had been a professor at Cambridge for most of his professional scholarly life. He occupied the Lucasian Chair of Mathematics. In fact the oft-told Cinderella story (which is verifiably true) is that Newton's teacher, Isaac Barrow, stepped down from that position so that Newton could occupy it.[8] An extraordinary act of scholarly charity. Certainly Newton was the greatest scientist, and one of the three greatest mathematicians, who ever lived. He is remembered with reverence by the British, especially the British intellectuals. So it is natural that they would build a mathematics institute in his honor.

Among the activities at the Newton Institute are a variety of mathematics conferences. In 1993 there was to be a conference on algebraic number theory. Certainly Andrew Wiles, a tenured Professor at Princeton University, was one

[8]But there were other factors at play in Barrow's life. He had been offered a position at court, and he needed to resign his professorship in order to accept it.

It may be noted that Stephen Hawking occupies the Lucasian Chair today. Rare book collector John Fry owns a first edition of Newton's *Principia* that has been autographed by Hawking.

of the most prominent and accomplished number theorists in the world. He had been extraordinarily quiet for several years, not producing any papers nor giving many talks. But his reputation was still strong, and he was invited to give one of the plenary talks at this conference.

The first unusual thing that Wiles did was write to the organizers to say that he had some important and substantial things to say, and he couldn't fit his message into a single lecture. He needed three lectures to exposit his ideas. Such *chutzpah* is almost never seen in mathematics, and certainly never from such a quiet and polite individual as Andrew Wiles. But the organizers were great admirers of Wiles's earlier work, and they acceded to his request.

It might be mentioned that, up to this point in time, Wiles had been quite secretive about his work on Fermat's last problem. He felt that what he had been doing was rather daring, and he did not want to start a spate of gossip. He also, quite frankly, did not want other people horning in on his work. Perhaps more importantly, he just wanted to be left alone. Even after he had a manuscript written up, he would not show the work in its entirety (and its entirety was about 200 manuscript pages!) to even his most trusted colleagues and associates. He showed different parts to different people, and he asked for their criticism and feedback.

When the appointed time came around, and Wiles delivered his thoughts to a packed audience at the Newton Institute, there was a definite sense that something exciting was happening. Of course Wiles had not told anyone the details of what he was up to, nor what his main message was going to be. But, as he proceeded from the first lecture, to the second lecture, to the third lecture (on successive days), the tension and excitement were definitely building. By the third lecture it was clear that something new had been added, for there were a number of members of the press in the room. Clearly Andrew Wiles was going to do something seminal, perhaps of historical significance.

And, indeed, as he wound down to the end of his third lecture, Andrew Wiles wrote on the board that a certain elliptic curve was modular. Then he drew a double arrow and wrote "*FLT*". Of course this last is number theorists' slang for "Fermat's Last Theorem". Andrew Wiles had just announced to the world that Fermat's Last Theorem was proved.

Of course this was a huge event. The news instantly went out by e-mail all around the world. Soon everyone knew that Andrew Wiles had proved Fermat's last theorem. We have already mentioned that Wiles had written

almost no papers and given almost no talks for the preceding seven years. He also had cut off communication with most of his professional friends and associates. The reason was now clear: He had been up in the attic of his house, sitting at an old desk and working on Fermat's Last Theorem.

There was a huge champagne celebration at the Newton Institute, and everyone hastened to tell Andrew Wiles that he was a hail fellow well met. It should be borne in mind that Wiles was at this time a number theorist of the highest reputation. If he announced something to be true then it was felt that it must certainly be true. Mathematicians of this caliber do not make mistakes.

At this point perhaps a little further background should be provided. Wiles had, rather cautiously, been sharing some of his ideas with his colleagues (and former colleagues) at Princeton—notably Nick Katz, Peter Sarnak, and Fields Medalist Gerd Faltings. In fact he had given each of these mathematicians *portions* of his manuscript to check. Nobody got to see the whole thing! And they dutifully slugged their way through the dense mathematics—much of it very original and unfamiliar—in order to aid Wiles with his work. Thus, when Wiles made his famous announcement at the Newton Institute, he could be fairly confident that what he was saying was correct and verifiable.

The conference in Cambridge lasted about one week, and then Wiles returned to Princeton. Of course now the cat was out of the bag and pandemonium reigned in Princeton University's Math Department. It is safe to say that no event of this magnitude had ever taken place in august Fine Hall. Most breakthroughs in pure mathematics tend to be of little interest to nonspecialists. For the most part, they are too technical and recondite for anyone but another mathematician to understand. But Fermat's Last Theorem was different. First of all, *anyone* could understand the problem. Secondly, the problem was over 300 years old and Fermat had posed it in such a charming and eccentric way. Wiles's achievement seemed, in the public mind, to represent the pinnacle of the (scholarly) single combat warrior defeating the giant Goliath of mathematics. Princeton's Fine Hall was awash with reporters and news cameras and press conferences.[9] Ordinarily taciturn and painfully shy mathematicians found themselves before the camera making demure state-

[9]Andrew Wiles was even often the opportunity to appear in a *Gap* commercial. Barbara Walters wanted to have him on her show. When Wiles revealed that he had no idea who Walters was (nor did he much care), she chose Clint Eastwood instead.

ments about the value and beauty of Wiles's work. Of course Wiles was in seventh heaven. Fermat's Last Theorem had been his pet problem since childhood. By solving the problem, he had fulfilled a lifelong ambition.[10]

Of course Wiles at this stage participated in the ordinary process of scholarly interchange. That is to say, he circulated copies of his (quite long) paper all over the world. And you can bet that people *read it*. After all, this was the event of the season. The work was extremely technical and difficult to plow through, but good number theorists could run seminars and work their way through all the delicate reasoning and calculations.

And then disaster struck. Professors Katz and Illusie had been checking a certain portion of Wiles's argument, and they found an error. A very serious error. Wiles had claimed (without proof!) that a certain group was finite. But in fact there was no way to confirm this assertion. Thus Wiles's celebrated proof of *FLT* had developed a leak. Now he was a single combat warrior of a new stripe. For nobody was going to step up to the plate and fix Wiles's theorem *for* him. *He* had to do it.

For a while the fact of the gap in Wiles's proof was kept "in house". Nobody wanted to embarrass Andrew Wiles. And there was definitely hope that he would be able to fix it. Now it should be noted that Wiles's teacher in England had been John Coates. Over time, a certain amount of polite professional rivalry had developed between the two. When Coates got wind of the problems with Wiles's proof of Fermat's last theorem, he was only too happy to circulate the news. Thus the cat was finally let out of the bag.

Once again, the story of Andrew Wiles emphasizes the traditional "single combat warrior" view of mathematics that has been prevalent for over two thousand years: A single individual comes up with the ideas, a single individual writes the paper, a single individual defends it. The mathematical community at large studies the work and decides whether it is worthwhile and correct. If the work passes muster, then that single individual reaps the benefits and rewards. If the work is found to have an error, then it is the responsibility of that single point man to fix it.

Princeton University has extraordinary resources, and an extraordinary

[10] An interesting sidenote on all the publicity that accompanied Wiles's accomplishment is this. The mathematicians at Princeton are no wilting flowers. They all have a pretty good opinion of themselves. And they started to feel that if Andy Wiles deserved all this publicity then perhaps they did as well. Before long, other Princeton faculty were announcing that *they too* had solved famous open problems—including the **P**/**NP** problem discussed above. Unfortunately none of these claims ever panned out.

dedication to its faculty and their work. The institution gave Wiles a year off, with no duties, so that he could endeavor to fix his proof. Poor Wiles. He had made such an extravagant and bold claim, and received *such* notoriety. His picture had appeared on the front page of every newspaper in the world. And now his program was at serious risk. Unless he could fix it.

Andrew Wiles is a devoted scholar. He believes in his mathematics, and he believes in himself, and he believed in his proof of Fermat's Last Theorem. He set himself to work on patching up his now-famous-but-incorrect proof. Since he did not know how to show that that particular group was finite, he developed other approaches to the problem. Time after time, they all failed. As the year drew to a close, Wiles was on the brink of despair. But then he had a stroke of luck.

Andrew Wiles had been sharing his thoughts with his graduate student Richard Taylor. Taylor was extremely gifted (he is now a Professor at Harvard), and could understand all the intricacies of Wiles's approach to the *FLT*. He recommended that Wiles return to his original approach to the problem, and he contributed some ideas of his own on how to deal with the difficulty of that group being finite. Thus together Wiles and Taylor re-tackled the original program that Wiles had set for proving Fermat's Last Theorem. And finally they succeeded. The benchmark was that they gave Fields Medalist Gerd Faltings the *entire new manuscript*, and he read it straight through. At the end Faltings declared, "Wiles and Taylor have done it. *FLT* is proved." And that was it. This is how the mathematical method operates in our world. One good mathematician asserts that he can prove a new theorem. He writes down a proof in the accepted argot and shows this recorded proof to one or more expert colleagues. Then they assess the work and pass judgment. There is now an entire issue of the *Annals of Mathematics*—in volume 141, 1995—containing the proof of FLT. This consists of a long paper by Wiles alone and a somewhat shorter paper by Taylor and Wiles. Although Wiles generally receives the lion's share of the credit for this great accomplishment, Taylor's contribution must be recorded as being key to the solution.

One of the reasons that *mathematicians* were so excited by Wiles's work on Fermat's Last Theorem is that it built on so many of the important ideas of number theory that had been developed in the preceding fifty years. One of the keys to what Wiles did was Ken Ribet's (Professor of Mathematics at U. C. Berkeley) result that the famous Taniyama-Shimura-Weil Conjecture implies *FLT*. For many years Wiles had kept his lifelong interest in Fermat's

Last Theorem something of a secret. It was considered to be one of those impossibly difficult problems—three hundred years old and something that was perhaps best worked around—and he would have been embarrassed to let people know that he was spending time on it. But Ribet's result put the *FLT* back in the mainstream of number theory. It made *FLT* a matter of current interest. It hooked the great unsolved problem up with key ideas in the main flow of modern number-theoretic research.

The key upshot of the chain of thought in the last paragraph is that if Wiles could prove the Taniyama-Shimura-Weil Conjecture then Fermat's Last Theorem would follow. In the end, Wiles did not succeed in doing that. Instead he ended up proving a *modified* or limited version of Taniyama-Shimura-Weil, and that was sufficient to prove *FLT*.

By today a great many people have checked Andrew Wiles's proof. There have been simplifications to some of the steps, and the problem has also been generalized (the famous "ABC Conjecture" is one such generalization that has become quite famous). Certainly Fermat's Last Theorem is now considered to be a genuine theorem, and Wiles's ideas have become an important part of the fabric of number theory.

Of course Wiles's work has placed the spotlight on all the ideas that contributed to his victory over Fermat's last problem. In particular, the Taniyama-Shimura-Weil conjecture shared some of the limelight. As a result, that conjecture has now been fully proved as well. This is an elegant example of how mathematics works in practice.

9.9 The Elusive Infinitesimal

When calculus was first invented by Newton and Leibniz, people were not quick to embrace these new ideas. The techniques of calculus were evidently powerful, and could be used to solve a host of problems that were long considered to have been intractable. To be sure, Newton successfully studied and analyzed

- motion
- gravity
- refraction of light
- the motions of the planets

- mechanics

But there were a number of theoretical underpinnings of calculus that were not well understood—even by the subject's inventors. Among these were the concept of limit—which actually would not be properly understood until Augustin-Louis Cauchy (1789-1857) gave a rigorous definition in 1821. But the real sticking point in the subject was infinitesimals. An infinitesimal to Isaac Newton was a number that was positive, but smaller than any ordinary real number. Thus an infinitesimal is a *positive* number that is smaller than 1/10, smaller than 1/1000, smaller than 10^{-10}, smaller than the radius of an atom. How could such a quantity exist?

No less an eminence than Bishop Berkeley (1685–1753)—who was himself a considerable scholar—weighed in with a skeptical view of the calculus. One of Berkeley's broadsides reads:

> All these points, I say, are supposed and believed by certain rigorous exactors of evidence in religion, men who pretend to believe no further than they can see. That men who have been conversant only about clear points should with difficulty admit obscure ones might not seem altogether unaccountable. But he who can digest [infinitesimals] need not, methinks, be squeamish about any point of divinity.

The Bishop succeeded in casting a pall over calculus that was not to be lifted for nearly two hundred years.

Well, the truth is that nobody—not even Newton himself—really knew what an infinitesimal was. For Newton infinitesimals were a terrific convenience—they enabled him to perform incisive physical and mathematical reasoning that led to daring and seemingly correct conclusions. Yet the logical foundations for what he was doing seemed to be suspect.

Cauchy's theory of limits, which came much later, showed us how to "work around" infinitesimals. But the infinitesimals were still there, and physicists and others loved to reason using infinitesimals—they were compelling and heuristically appealing and led to useful and powerful conclusions. But nobody could say what they were. Enter Abraham Robinson.

Robinson was a mathematician remarkable for his breadth and depth. He studied both very applied problems and very abstract and pure problems. He was universally admired and respected for his many and varied contributions to mathematics. In 1963 Robinson created a new subject area

called "nonstandard analysis". The most important feature of this new creation is that it is a number system that contains all the usual real numbers $\mathbb{R}$ that mathematicians routinely use—the whole numbers, the rational numbers or fractions, and the irrational numbers (like $\sqrt{2}$ and π)—but it also contains infinitesimals. Yes, Abraham Robinson's number system $\mathbb{R}^*$ contained the elusive figments that Isaac Newton had conceived three hundred years earlier.

What is important to understand here is that Robinson's construction of the nonstandard reals is *completely rigorous*. Using subtle techniques from algebra and logic—like the idea of an ultrafilter—Robinson gives an explicit and irrefutable *construction* of this new number system. And the infinitesimals are plainly and explicitly exhibited. What is more, this new number system also contains numbers that are infinitely large; we call these *infinitary numbers*. An infinitary number has the extraordinary property that it is larger than 10, larger than 1000, larger than 1000000, larger than 10^{100}, larger than the diameter of the sun measured in microns—in fact larger than *every* ordinary real number.

Robinson's new idea came as a bolt from the blue. Now one had an entirely new universe in which to practice mathematics. And it actually came to pass that mathematicians could discover new facts, new truths, new theorems in the nonstandard real world that in fact had never been known before in the ordinary world of the real numbers. Best of all, after a new theorem was proved in the nonstandard context, it usually could be pushed back down into the ordinary world of the real numbers. So one actually obtained a new theorem in the context of traditional mathematics.

Thus nonstandard analysis—particularly the infinitesimals that Abraham Robinson created—have given us a new, rigorous, precise tool for doing mathematical analysis. One leaves the ordinary world of the real numbers and enters the ethereal (but definitely genuine, and rigorously constructed) world of the nonstandard reals. There one can perform amazing feats—ones that would be impossible in the prosaic world of the ordinary reals. Having achieved what was once thought to be impossible, one then goes back down into the ordinary real numbers and has a new theorem.

It is a bit like looking for a lost child in the wilderness. One climbs into an aircraft and searches in a new way—that would be inconceivable if one were confined to the ground. One can see further, and one can peer deeply into the forest in ways that would be impossible for someone on foot. Finally one spots the child and radios a ground crew that can go quickly to the location

and rescue the lost one. The point is that the aircraft is an intermediary tool that has nothing to do with the ultimate goal: to find the child. The airplane gives one additional powers that make the solution of the problem more expeditious and more sure. But, once the problem is solved, the aircraft is put back into the hangar and life resumes its ordinary form. So it is with the nonstandard reals $\mathbb{R}^*$. They are a tool that gives one greater power, and enables one to see things that otherwise would be infeasible. Once the infinitesimals have done their work, they are put back on the shelf. The problem is solved and one can resume life in the ordinary real numbers.

9.10 A Miscellany of Misunderstood Proofs

This author once wrote (see [KRA])

> Being a mathematician is a bit like being a manic depressive: you spend your life alternating between giddy elation and black despair. You will have difficulty being objective about your own work: before a problem is solved, it seems to be mightily important; after it is solved, the whole matter seems trivial and you wonder how you could have spent so much time on it.

Part of the genesis of the "black despair" is the difficulty of solving mathematical problems, and the time that one must invest to **(i)** *understand* and **(ii)** *solve* such a problem is certainly a source of frustration and unhappiness. Once you have your paper written up and ready to show the world, then you have to worry about whether anyone else will appreciate your effort. You have invested perhaps two or more years in this project; will you garner a commensurate amount of adulation? This last issue can manifest itself in a couple of different forms:

The world may think that your paper is correct. The referee at the journal to which you submit the paper may sign off on it, and recommend it for publication. But nobody may think it is very interesting, or very important, or very original, or very new. The great mathematician John E. Littlewood said that a good mathematical paper should be *new*, *interesting*, and *correct* (see [LIT]). These are what we in the business call necessary conditions but not sufficient. What this means is that a paper lacking any of these qualities will end up on the trash heap of time. But a paper that has all these qualities is not guaranteed to ring the bell, to garner a lot of attention

and praise, to be dubbed "an important piece of work." It is a candidate for being a seminal piece of work; it has all the right attributes. But it depends on the jury of the world's mathematicians to determine whether this is work of lasting value or just another drop in the bucket.

It all comes down to the adage that it is the mathematical community that decides what is worthwhile. And that decision is based on intrinsic merit but also on fashion. This may seem odd to the uninitiated, but mathematics has vogues and cliques and prejudices just like the music or literature or the clothing industries do. Given any subject area, there will be an intense period of excitement when hot new ideas are developed and any new paper on the topic is considered to be of *a priori* interest. After a time the novelty wears off, it all starts to look like the same old same old, and people move on to something else.

The big shots at the top math departments—Harvard, Princeton, MIT, Berkeley, Paris, Göttingen—have a lot of influence on the directions that mathematics may take, and on what topics are considered to be important. But a mathematician at most any university can have a big breakthrough that attracts everyone's attention. And that will set a new direction for the field.[11] For the time being, the subject of this big discovery will be the most important thing in sight. And as long as significant new ideas are forthcoming, the area will be hot. But after a while it will fade from view. It can be revived if someone comes along with a new idea, and this often happens. But just as often it does not.

So if your new paper, the product of your intense labor over a protracted period of time, is in an area that is of broad interest, then you will find that you will earn much appreciative commentary, you will be invited to speak at conferences, and you will enjoy some of the warmest encomia of the profession. In the other case your paper will be published, put on the shelf, and more or less forgotten.

What we have just described is the ordinary course of life, not just for the mathematician but for academics in general and for people who are engaged in any line of creative work—whether it be music or painting or dance or cooking. As a mathematician matures, he or she realizes that one of the main reasons for doing academic work is to satisfy one's own intellectual

[11] As an instance, Terence Tao of UCLA just won the Fields Medal. He has had a great many important new ideas. Now UCLA is a good school, but it only ranks #26 in the *U.S. News and World Report* college ranking. Great ideas and brilliant minds can reside anywhere.

curiosity, and to work out one's own ideas. For most of us that is more than enough reason to continue to pursue the holy grail of the next theorem.

Of course, as we have indicated elsewhere, it also can happen that your new paper has a mistake in it. If it is a small mistake, then it is likely you can fix it and press on. If it is a big mistake, then you may have to shelve the paper and find another pursuit to justify your existence. It is occasionally possible to rectify even this dire situation. Andrew Wiles made a big mistake in his first proof of Fermat's Last Theorem (see Section 9.1). It took him a full year to fix it, but he finally did. And then he reaped the rewards.

9.10.1 Frustration and Misunderstanding

Probably the most frustrating, and often unfixable, situation is when you are quite sure that your new paper is correct but the world does not accept it. That is to say, you believe that you have given a *bona fide* proof of a significant new result, but the community of mathematicians does not believe your proof. If you are in such a situation, and you are lucky, someone will sit down with you and show you where your error is. But many times the world just ignores a paper that it does not believe, and simply forges ahead.

And where does that leave you? You can go on the lecture circuit and try to convince people that you really have the goods. But the fact is that mathematics is vetted and verified in a very special and precise way: individual mathematicians sit down and read your work. They either believe it or they do not. And in the latter case you are left in an extremely awkward situation.

We have provided this extensive prolegomena to set the stage for now listing a few famous instances of people who have written significant papers or books purporting to solve important problems, but who have found that the mathematics profession has not been willing to validate and accept the work. [Refer once again to the case of Bill Thurston and the great geometrization program described in Section 8.4.] In most of these cases the progenitors have experienced bitterness, unhappiness, and frustration. In some instances the experience changed the direction of a person's career. In other instances, the person quit doing mathematics altogether.

- In 1969 Alfred Adler of the State University of New York at Stony Brook published a paper in the *American Journal of Mathematics* [ADL] purporting to show that there is no complex manifold structure

on the six-dimensional sphere. Complex manifolds are surfaces that have certain important algebraic and analytic properties. These play a significant role in differential geometry and mathematical physics. But it is difficult to come up with concrete examples of complex manifolds. The six-dimensional sphere would be an interesting and significant example if indeed it had a complex structure. Adler asserted that it did not. And he got his paper published in a top-ranked journal, so one may infer that a solid referee with strict standards agreed.

But people have not accepted Adler's proof. For thirty-seven years Adler has maintained that his proof is correct. He is now retired, and no longer participating in the discourse. But he produced no further published work after his 1969 paper. And, for a good many of those thirty-seven years, nobody was able to put their finger on where the mistake was. Finally, about five years ago, Yum-Tong Siu of Harvard wrote a paper in which he explained precisely where the error was. But Adler's paper has never been repaired, and nobody knows to this day how to prove the assertion either true or false.

- In 1993 Wu-Yi Hsiang published his solution of the Kepler sphere-packing problem (see the detailed discussion in Section 8.3). He published the paper in a journal of which he was an editor, and perhaps he received some friendly treatment from the journal. That is to say, the paper may not have been subjected to the most rigorous refereeing. Even before the publication of the paper, the ranking experts had mounted some objections to Hsiang's arguments. He had claimed in the paper that he had identified the "worst case scenario" for sphere packing, and he then proceeded to analyze that particular case. The experts did not find this argument to be either complete or compelling. But Hsiang believed in his methodology and proceeded to publish the work. And there it stands. Hsiang's methodology was traditional mathematics—he used methods of spherical trigonometry. Anyone who chooses to do so can read the paper and check the work. The experts did so, and in the end did not accept it as a proof.

  Meanwhile Thomas Hales (again see Section 8.3) has produced a new proof of the Kepler sphere-packing conjecture. It is being published in the prestigious *Annals of Mathematics*, and was subjected to rigorous refereeing over a long period of time. But Hales's arguments

rely heavily on intense and protracted computer calculation. They cannot possibly be checked by a human being. There are now techniques for having a second computer check the work of the first computer, and Hales has organized a project to carry out such a verification. He anticipates that it will take at least twenty years.

- In 1978 Wilhelm Klingenberg published a book [KLI] in which he asserted that any closed, bounded, surface in space has infinitely many distinct closed geodesics. Here a *geodesic* is a curve of least length. For example, in the plane, the curve of least length connecting two points is a straight line. On a sphere (like the earth), the curve of least length connecting two points is a great circle. Airlines follow great circle paths when flying from Los Angeles to Paris, for example. Klingenberg is a very distinguished mathematician who had spent his entire career studying geodesics. He had made many important contributions, and was the leader in the field. But people did not accept his proof.

  The distinguished mathematician Friedrich Hirzebruch used to hold yearly conferences in Bonn that were dubbed the *Arbeitstagungen*. These were quite unusual events. The first day of the conference was spent with *all* the participants debating who should be allowed to speak and why. After Klingenberg announced his result, he was certainly asked to speak at the Arbeitstangung. But the audience really gave him a hard time, peppered him with questions, and cast doubt upon various parts of his proof. To this day there is no consensus on whether Klingenberg's proof of this important result is correct.

- We have already told the story, in Section 8.4, of William P. Thurston's geometrization program. This is surely one of the dazzling new ideas in geometric topology, and it would imply the celebrated Poincaré conjecture. Thurston has never been able to produce a proof that others consider to be complete or compelling. Although there are many who believe, both on *a priori grounds* and because they are sold on Thurston's outline proof, that Thurston has a *bona fide* theorem, the world at large is still waiting for the details. Grisha Perelman's new arguments (see Section 8.5) may settle the matter once and for all. But it will be some time before we know this matter for certain.

- When Frederick Almgren died in 1997, he left behind a 1728-page manuscript purporting to prove an important regularity theorem for

minimal surfaces. These are surfaces that model soap films, polymers, and other important artifacts of nature. Now it should be stressed that Almgren was a very distinguished mathematician, a pioneer in his field, and a Professor at Princeton University. *He did not make mistakes.* On *a priori* grounds, people are inclined to believe Almgren's theorem. But it is essentially impossible for any human being to digest the 1728 pages of dense and technical mathematics that Almgren produced. Two of his students, Vladimir Scheffer and Jean Taylor (also Almgren's wife) published Almgren's manuscript as a book [ALM]. But it may be a long time before this work has gone through the usual vetting and has been given the official imprimatur.

- We conclude this section by relating the story of a proof that caused a huge fight to occur. The incident is more than fifty years old, yet there remain bitter feelings. The two chief participants are very distinguished mathematicians. But everyone took a side in this matter, and the raw feelings still fester. The issue here is *not* whether a given proof is correct. Rather, the question is whose proof it is.

  In 1948 Paul Erdős was in residence at the Institute for Advanced Study and of course Atle Selberg (1917–) was a permanent member. Selberg got a promising idea for obtaining an elementary proof of the prime number theorem (that is, a proof that does not use complex analysis or other tools outside of elementary number theory). See Section 9.1 for background. Paul Erdős was able to supply an important step that Selberg needed. Later, Selberg was able to modify his proof so that it did not require Erdős's idea. Unfortunately, a terrible priority dispute erupted between Erdős and Selberg. Since he was able to contribute an important step to the argument, Erdős just assumed that he and Selberg would write a joint paper on the result. Selberg had other ideas.

  One version of the tale is that Selberg was visiting at another university, sitting in somebody's office and having a chat. Another mathematician walked in and said, "I just got a postcard saying that Erdős and some Norwegian guy that I never heard of have found an elementary proof of the prime number theorem." This really set off Selberg, and he was then determined to write up the result all by himself.

  Dorian Goldfeld (1947–) has taken pains to interview all survivors

who participated in or witnessed the feud between Erdős and Selberg. It is clear that nobody was wrong and nobody was right. Both Erdős and Selberg contributed to this important discovery, but they had a significant clash of egos and of styles. Irving Kaplansky (1917–) was in residence at the Institute in those days and witnessed the feud in some detail. He tells me that at one point he went to Erdős and said, "Paul, you always say that mathematics is part of the public trust. Nobody owns the theorems. They are out there for all to learn and to develop. So why do you continue this feud with Selberg? Why don't you just let it go?" Erdős's reply was, "Ah, but this is the prime number theorem."

Chapter 10

John Horgan and "The Death of Proof?"

I have often maintained, and even committed to paper on some occasions, the view that mathematics is a science, which, in analogy with physics, has an experimental and a theoretical side, but operates in an intellectual world of objects, concepts and tools. Roughly, the experimental side is the investigation of special cases, either because they are of interest in themselves or because one hopes to get a clue to general phenomena, and the theoretical side is the search for general theorems.

Armand Borel

Mathematics is the science of necessary conclusions.

C. S. Pierce

If, as they say, some dust thrown in my eyes
Will keep my talk from getting overwise,
I'm not the one for putting off the proof.
Let it be overwhelming.

Robert Frost

We are talking here about theoretical physics, and therefore of course mathematical rigor is irrelevant and impossible.

Edmund Landau

There are writings on the wall that, now that the silicon savior has arrived, a new testament is going to be written. Although there will always be a small group of "rigorous" old-style mathematicians . . . who will insist that the true religion is theirs and that the computer is a false Messiah, they may be viewed by future mainstream mathematicians as a fringe sect of harmless eccentrics, as mathematical physicists are viewed by regular physicists today.

Doron Zeilberger

The folly of mistaking a paradox for a discovery, a metaphor for a proof, a torrent of verbiage for a spring of capital truths, and oneself for an oracle, is

inborn in us.

Paul Valéry

Ah, Why, ye Gods, should two and two make four?

Alexander Pope

To foresee the future of mathematics, the true method is to study its history and its present state.

Henri Poincaré

The question of the foundations and the ultimate meaning of mathematics remains open; we do not know in what direction it will find its final solution or even whether a final objective answer can be expected at all. "Mathematizing" may well be a creative activity of man, like language or music, of primary originality, whose historical decisions defy complete objective rationalization.

Hermann Weyl

10.1 Horgan's Thesis

In 1993 John Horgan, a staff writer for *Scientific American*, published an article called *The Death of Proof?* [HOR1]. In this piece the author claimed that mathematical proof no longer had a valid role in modern thinking. There were several components of his reasoning, and they are well worth considering here.

First of all, Horgan was a student of literary critic Harold Bloom at Yale University. One of Bloom's principal theses—in the context of *literature*—is that there is nothing new under the sun. Any important and much-praised piece of modern literature is derivative from one of the classics by Shakespeare or Chaucer or Spencer or Dryden or some other great figure from 400 years ago. Horgan endeavored to shoehorn modern scientific work into a similar mold. He claimed that Isaac Newton and the other great minds of 300 years ago had all the great ideas, and all we are doing now is turning the crank and producing derivative thoughts.

A second vector in Horgan's reasoning is that computers can do much more effectively what human beings have done traditionally. Which is to *think*. Put in slightly different words, a mathematical dinosaur (such as this author) might claim that the deep insights of mathematics can only be discovered and produced and verified (i.e., proved) by a human being

with traditional training in mathematics. John Horgan would claim that a computer can do it better and more effectively and certainly much faster.

The third component of John Horgan's reasoning was that mathematical proofs have become so complicated (witness Andrew Wiles's proof of Fermat's Last Theorem, which takes up an entire issue of the *Annals of Mathematics*) that nobody can understand them anyway. So how could they possibly be playing any significant role in the development of modern mathematical science?

The trouble with Horgan's arguments is that he is thinking like a literary critic, and reasoning by analogy. One may agree or disagree with Harold Bloom's thesis (for instance, from what classical work is Alan Ginsburg's *Howl* derivative? Or James Joyce's *Finnegan's Wake*?). But there is no evidence that it applies to modern scientific progress. Certainly Isaac Newton, to take just one example, is the greatest scientist who ever lived. First of all, he created the modern scientific method. Second of all, he created mathematical physics. Thirdly, he (along with Leibniz) invented calculus and created the most powerful body of scientific/analytic tools ever devised. There may never be another scholar to equal Newton. But each generation of scientific work builds atop the earlier work. It would be quite difficult to argue that relativity theory is derivative from the work of 300 years ago. Or that quantum theory is derivative from ideas of Maxwell. Or that string theory is derivative from ideas of Fermat.

Second, computers are terrific at manipulating data accurately and rapidly. But they cannot think—at least not in the way that humans think. Artificial intelligence software is an effort to make the computer perform tasks that the human brain can perform. Some notable successes have been achieved, but they are rather elementary. They don't begin to approach or emulate the power and depth of something like Andrew Wiles's proof of Fermat's Last Theorem.

Lastly, we cannot help but agree with Horgan that mathematical proofs have become rather complicated. Two hundred years ago, published mathematics papers tended to be fairly short, and arguments were rarely more than a few pages. Today we are considerably more sophisticated, mathematics is a much larger and more complex and sophisticated enterprise, and yes, indeed, many mathematics papers run to fifty pages and have enormously complex proofs. But it is incorrect to assert that other mathematicians are helpless to check them. In fact they do. It is an enormous amount of work to do so, but if a result is important, and if the proof introduces significant new

techniques, then people will put in the time to study the work and validate it. Many people have studied Andrew Wiles's proof, and it is quite certain that it is correct. Moreover, people have simplified several parts of the proof. And there are important generalizations of Fermat's Last Theorem, such as the *ABC* Conjecture, which put the entire problem into a new perspective and will no doubt lead to further simplifications.

Mathematics is a process, marked by milestones such as solutions of the great problems and discoveries of marvelous new theorems. But, over time, the big, complicated ideas get digested and worked over and built into the infrastructure of the subject. In the end they are simplified and rendered natural and understandable. Gauss actually anticipated the discovery (by Bolyai and Lobachevsky) of non-Euclidean geometry. He did not share or publish his ideas because he felt that they were too subtle; nobody would understand them. Today we teach non-Euclidean geometry to high school students. The Cauchy integral, Galois theory, the Lebesgue integral, cohomology theory, and many other key ideas of modern mathematics were considered to be impossibly recondite when they were first discovered. Today we teach them to undergraduates.

John Horgan's article created quite a stir. In the mathematics community, this author became the point man for responding to the piece; the rebuttal appears in [KRA1]. In my article, I make the points just enunciated, but in considerably more detail. It is safe to say that the mathematics community has evaluated Horgan's arguments and rejected them.

But Horgan in fact leveraged his thesis into something more substantial. He subsequently published a book called *The End of Science* [HOR2]. This work carried on many of the themes introduced in the *Scientific American* article. In this book the author claims that scientific research has come to an end. All the great ideas were discovered prior to 1900 and what we are doing now is a sham—just a way of scamming National Science Foundation grant money from the government.

This new claim by Horgan is reminiscent of repeated efforts in the nineteenth and early twentieth centuries to close the U. S. Patent Office—with the claim that everything has already been invented and there is nothing more to do. Even in the past 25 years the number of new inventions has been astonishing—ranging from the personal computer to the cellular telephone to the Segway transportation device. Again, Horgan is guilty in this book of reasoning by analogy. He advocates, for instance, thinking about how it looks when you are driving a car into a brick wall. In the last few

moments, it looks as though the wall is approaching ever faster. Just so, Horgan reasons, it looks as though science is making progress at an exponentially increasing rate because in fact it is running into a brick wall (of no progress).

It is difficult to take such arguments seriously. Reasoning like this, which may be most apposite in the context of literary theory, make little sense in the milieu of science. It seems irrefutable that science is in a golden age. The genome project has opened up many new doors. String theory is really changing the face of physics. There are so many new directions in engineering—including biomedical engineering—that it would be impossible to describe them all. Chemistry changes our lives every day with new polymers and new synthetic products.

10.2 Will "Proof" Remain the Benchmark for Mathematical Progress?

The great advantage of a traditional mathematical proof is that it is a bulletproof means of verifying and confirming an assertion. A statement that is proved today by the methods introduced by Euclid in his *Elements* will still be valid 1000 years from now. Changing fashions, changing values, and changing goals do not affect the validity of a mathematical proof. New discoveries do not invalidate old ones when the old ones were established by proof.

It is for this reason that mathematical proof will continue to be a standard to which we can all aspire. But we can at the same time tolerate and appreciate other approaches to the matter. Mathematical proofs do not traditionally carry a great deal of weight in a physics or engineering department. Heuristic reasoning traditionally does not hold much water in a mathematics department. As indicated in the beginning of this book, a proof is a psychological device for convincing somebody that something is true. If that "somebody" is a classically trained mathematician, then most likely the proof that he/she wants to see is a traditional, Euclid-style, logical argument. If instead that "somebody" is a modern numerical analyst, then he/she may be more convinced by a long and very precise computer calculation.

Note that neither of the two scenarios painted at the end of the last paragraph supplants the other. They serve two different purposes; their

intent is to bring two different types of people on board to a certain program. In fact, properly viewed, they complement each other.

One of the main points of the present book is that "proof"—the traditional, Euclid-inspired, tightly knit chain of logical reasoning leading inexorably to a precise conclusion—is immortal. It was invented to serve a very specific purpose, and it does so admirably and well. But this mode of reasoning, and of verifying assertions, is now put into a new context. It is a very robust, and supportive, and stimulating context that only *adds* to what we can do. It augments our understanding. It increases our arsenal of weapons. It makes us all stronger.

Chapter 11

Methods of Mathematical Proof

. . . it was said that one person asked another, "What are you working on?"; "String theory"; "Oh, didn't you know, that went out of fashion last week."

Saunders Mac Lane

This [negative effects of speculation] has been true when incorrect or speculative material is presented as known and reliable, and credit is claimed by the perpetrator. Sometimes this is an "honest mistake," sometimes the result of non-standard conceptions of what constitutes proof. Straightforward mistakes are less harmful.

Arthur Jaffe and Frank Quinn

We are now looking for what might be called metamathematical structures. We remove the mathematics from its original context and isolate it, trying to detect new structures. It is impossible to collect only the relevant information that will lead to the new discovery. One collects objects (theorems, statistics, conjectures, etc.) that have a reasonable degree of similarity and familiarity and then attempts to eliminate the irrelevant or the untrue (counterexamples). We are preparing for some form of eliminative induction. In this context, it is not unreasonable to introduce objects without being sure of their truth, for all the objects, whether proved or not, will be subject to the same degree of scrutiny. Moreover, if these probably true objects fall into the class of desired objects (i.e., they fit the new conjecture), it may be possible to find a legitimate proof in the new context.

J. Borwein, P. Borwein, R. Girgensohn, and S. Parnes

I believe that elementary number theory and the rest of mathematics should be pursued more in the spirit of experimental science, and that you should be willing to adopt new principles. I believe that Euclid's statement that an axiom is a self-evident truth is a big mistake. The Schrödinger equation certainly isn't a self-evident truth! And the Riemann hypothesis isn't self-evident either, but it's very useful. A physicist would say that there is ample

experimental evidence for the Riemann hypothesis and would go ahead and take it as a working assumption.

G. J. Chaitin

After all, the probability of an error in the proof [of the classification of finite simple groups] is one.

Michael Aschbacher

The book thus far has provided a detailed introduction to the concept of proof, what it is for and how it is produced. No book on mathematical proofs would be complete without some discussion of particular techniques of mathematical proof. This chapter discusses direct proof, proof by contradiction, and proof by induction. It is rather more technical than the rest of the book, and the reader less interested in the nuts and bolts of mathematics may wish to skip ahead to the concluding chapter. But he/she may also find it amusing to browse through this chapter for a sense of what the practice of mathematics is really like.

11.1 Direct Proof

In this section and the next two we are concerned with form rather than substance. We are not interested in proving anything profound, but rather in showing you what a proof looks like. We shall assume now that you are familiar with the positive integers, or *natural numbers*. This number system $\{1, 2, 3, \ldots\}$ is denoted by the symbol $\mathbb{N}$. For now we will take the elementary arithmetic properties of $\mathbb{N}$ for granted. We shall formulate various statements about natural numbers and we shall prove them. We begin with a definition.

Definition 11.1.1 A natural number n is said to be *even* if it can be divided by 2, with integer quotient and no remainder.

Definition 11.1.2 A natural number n is said to be *odd* if, when it is divided by 2, the remainder is 1.

You may have never before considered, at this level of precision, what is the meaning of the terms "odd" or "even". But your intuition should confirm these definitions. A good definition should be precise, but it should also appeal to your heuristic idea about the concept that is being defined.

Notice that, according to these definitions, any natural number is either even or odd. For if n is any natural number, and if we divide it by 2, then the remainder will be either 0 or 1—there is no other possibility (according to the Euclidean algorithm—see [HER]). In the first instance, n is even; in the second, n is odd.

In what follows we will find it convenient to think of an even natural number as one having the form $2m$ for some natural number m. We will think of an odd natural number as one having the form $2k - 1$ for some natural number k. Check for yourself that, in the first instance, division by 2 will result in a quotient of m and a remainder of 0; in the second instance it will result in a quotient of $k - 1$ and a remainder of 1. Now let us formulate a statement about the natural numbers and prove it.

Proposition 11.1.3 *The square of an even natural number is even.*

Proof: We may reformulate our statement as "If n is even then $n \cdot n$ is even." This statement makes a promise. Refer to the definition of "even" to see what that promise is:

If n can be written as twice a natural number then $n \cdot n$ can be written as twice a natural number.

The hypothesis of the assertion is that $n = 2 \cdot m$ for some natural number m. But then

$$n^2 = n \cdot n = (2m) \cdot (2m) = 4m^2 = 2(2m^2).$$

Our calculation shows that n^2 is twice the natural number $2m^2$. So n^2 is also even.

We have shown that the hypothesis that n is twice a natural number entails the conclusion that n^2 is twice a natural number. In other words, if n is even then n^2 is even. That is the end of our proof. □

A companion result is the next proposition.

Proposition 11.1.4 *The square of an odd natural number is odd.*

Proof: We follow the model laid down in the proof of the previous proposition.

Assume that n is odd. Then $n = 2k - 1$ for some natural number k. But then

$$n^2 = n \cdot n = (2k-1) \cdot (2k-1) = 4k^2 - 4k + 1 = 2(2k^2 - 2k + 1) - 1.$$

We see that n^2 is $2k' - 1$, where $k' = 2k^2 - 2k + 1$. In other words, according to our definition, n^2 is odd. □

Both of the proofs that we have presented are examples of "direct proof." A direct proof proceeds according to the statement being proved; for instance, if we are proving a statement about a square then we calculate that square. If we are proving a statement about a sum then we calculate that sum. Here are some additional examples:

EXAMPLE 11.1.5 Prove that, if n is a positive integer, then the quantity $n^2 + 3n + 2$ is even.

Proof: Denote the quantity $n^2 + 3n + 2$ by K. Observe that

$$K = n^2 + 3n + 2 = (n+1)(n+2)\,.$$

Thus K is the product of two successive integers: $n + 1$ and $n + 2$. One of those two integers must be even. So it is a multiple of 2. Therefore K itself is a multiple of 2. Hence K must be even.

Proposition 11.1.6 *The sum of two odd natural numbers is even.*

Proof: Suppose that p and q are both odd natural numbers. According to the definition, we may write $p = 2r - 1$ and $q = 2s - 1$ for some natural numbers r and s. Then

$$p + q = (2r - 1) + (2s - 1) = 2r + 2s - 2 = 2(r + s - 1).$$

We have realized $p + q$ as twice the natural number $r + s - 1$. Therefore $p + q$ is even. □

Remark 11.1.7 If we did mathematics solely according to what sounds good, or what appeals intuitively, then we might reason as follows: "If the sum of two odd natural numbers is even then it must be that the sum of two even natural numbers is odd." This is incorrect. For instance 4 and 6 are each even but their sum $4 + 6 = 10$ is *not* odd.

Intuition definitely plays an important role in the development of mathematics, but all assertions in mathematics must, in the end, be proved by rigorous methods.

Proposition 11.1.8 *Let n be a natural number. Then either $n > 6$ or $n < 9$.*

Proof: If you draw a picture of a number line then you will have no trouble convincing yourself of the truth of the assertion. What we want to learn here is to organize our thoughts so that we may write down a rigorous proof.

Fix a natural number n. If $n > 6$ then the 'or' statement is true and there is nothing to prove. If $n \not> 6$, then we must check that $n < 9$. But the statement $n \not> 6$ means that $n \leq 6$ so we have

$$n \leq 6 < 9.$$

That is what we wished to prove. □

EXAMPLE 11.1.9 Prove that every even integer may be written as the sum of two odd integers.

Proof: Let the even integer be $K = 2m$, for m an integer. If m is odd then we write

$$K = 2m = m + m$$

and we have written K as the sum of two odd integers. If, instead, m is even, then we write

$$K = 2m = (m - 1) + (m + 1).$$

Since m is even then both $m-1$ and $m+1$ are odd. So again we have written K as the sum of two odd integers. □

EXAMPLE 11.1.10 Let N be a positive integer. Prove that

$$1 + 2 + \cdots + N = \frac{N \cdot (N+1)}{2}\,.$$

Proof: Let us write

$$S = 1 + 2 + \cdots + N\,.$$

Our device is to write S twice in a clever way:

$$\begin{array}{ccccccccc} S & = & 1 & + & 2 & + & \cdots & + & N \\ S & = & N & + & N-1 & + & \cdots & + & 1 \end{array}$$

Now we add each column vertically to obtain

$$2S = (N+1) + (N+1) + \cdots (N+1)\,.$$

Of course there are N copies of $(N+1)$ on the right. So we have derived

$$2S = N \cdot (N+1)$$

or

$$S = \frac{N \cdot (N+1)}{2}\,.$$

□

11.2 Proof by Contradiction

Aristotelian logic dictates that every sensible statement has a truth value: TRUE or FALSE. If we can demonstrate that a statement **A** could not possibly be false, then it must be true. On the other hand, if we can demonstrate that **A** could not be true, then it must be false. Here is a dramatic example of this principle. In order to present it, we shall assume for the moment that you are familiar with the system $\mathbb{Q}$ of rational numbers. These are numbers that may be written as the quotient of two integers (without dividing by zero, of course).

Theorem 11.2.1 (Pythagoras) *There is no rational number x with the property that $x^2 = 2$.*

Remark: We presented this result, and its proof, in Section 1.3 in another context. But this theorem has such great historic significance, and the argument is so profound, that we feel it is well worth repeating. This time we shall concentrate in more detail on the logical aspects of the argument.

Proof: Let us assume the statement to be false. So there is a rational number x with $x^2 = 2$. Since x is rational we may write $x = p/q$, where p and q are integers.

We may as well suppose that both p and q are positive and non-zero. After reducing the fraction, we may suppose that it is in lowest terms—so p and q have no common factors.

Now our hypothesis asserts that

$$x^2 = 2$$

or

$$\left(\frac{p}{q}\right)^2 = 2.$$

We may write this out as

$$p^2 = 2q^2. \qquad (*)$$

Observe that this equation asserts that p^2 is an even number. But then p must be an even number (p cannot be odd, for that would imply that p^2 is odd by Proposition 11.1.4). So $p = 2r$ for some natural number r.

Substituting this assertion into equation $(*)$ now yields that

$$(2r)^2 = 2q^2.$$

Simplifying, we may rewrite our equation as

$$2r^2 = q^2.$$

This new equation asserts that q^2 is even. But then q itself must be even (if q were odd then 11.1.4 would tell us that q^2 is odd).

We have proven that both p and q are even. But that means that they have a common factor of 2. This contradicts our starting assumption that p and q have no common factor.

Let us pause to ascertain what we have established: the assumption that a rational square root x of 2 exists, and that it has been written in lowest

terms as $x = p/q$, leads to the conclusion that p and q have a common factor and hence are *not* in lowest terms. What does this entail for our logical system?

We cannot allow a statement of the form $\mathbf{C}$ = "$\mathbf{A}$ and $\sim \boldsymbol{A}$" (in the present context the statement $\mathbf{A}$ is "$x = p/q$ in lowest terms"). For such a statement $\mathbf{C}$ must be false.

But if x exists then the statement $\mathbf{C}$ is true. No statement (such as $\mathbf{A}$) can have two truth values. In other words, the statement $\mathbf{C}$ must be false. The only possible conclusion is that x does not exist. That is what we wished to establish. □

Remark 11.2.2 In practice, we do not include the last three paragraphs in a proof by contradiction. We provide them now because this is our first detailed exposure to such a proof, and we want to make the reasoning absolutely clear. The point is that the assertions $\mathbf{A}$ and $\sim \mathbf{A}$ cannot both be true. An assumption that leads to this eventuality cannot be valid. That is the essence of proof by contradiction.

Historically, Theorem 11.2.1 was extremely important. Prior to Pythagoras the ancient Greeks (following Eudoxus) believed that all numbers (at least all numbers that arise in real life) are rational. However, by the Pythagorean theorem, the length of the diagonal of a unit square is a number whose square is 2. And our theorem asserts that such a number cannot be rational. We now know that there are many non-rational, or irrational numbers.

Here is a second example of a proof by contradiction:

Theorem 11.2.3 (Dirichlet) *Suppose that $n + 1$ pieces of mail are delivered to n mailboxes. Then some mailbox contains at least two pieces of mail.*

Proof: Suppose that the assertion is false. Then each mailbox contains either zero or one piece of mail. But then the total amount of mail in all the mailboxes cannot exceed

$$\underbrace{1 + 1 + \cdots + 1}_{n \text{ times}}.$$

In other words, there are at most n pieces of mail. That conclusion contradicts the fact that there are $n + 1$ pieces of mail. We conclude that some mailbox contains at least two pieces of mail. □

This last theorem, due to Gustav Lejeune Dirichlet (1805-1859), was classically known as the *Dirichletscher Schubfachschluss*. This German name translates to "Dirichlet's drawer inference principle." Today, at least in this country, it is more commonly known as "the pigeonhole principle." Since pigeonholes are no longer a common artifact of everyday life, we have illustrated the idea using mailboxes.

EXAMPLE 11.2.4 Draw the unit interval I in the real line. Now pick 11 points at random from that interval (imagine throwing darts at the interval, or dropping ink drops on the interval). Then some pair of the points has distance not greater than 0.1 inch.

To see this, write

$$I = [0, 0.1] \cup [0.1, 0.2] \cup \cdots [0.8, 0.9] \cup [0.9, 1].$$

Here we have used standard interval notation. Think of each of these subintervals as a mailbox. We are delivering 11 letters (that is, the randomly selected points) to these ten mailboxes. By the pigeonhole principle, some mailbox must receive two letters.

We conclude that some subinterval of I, having length 0.1, contains two of the randomly selected points. Thus their distance does not exceed 0.1 inch. □

11.3 Proof by Induction

The logical validity of the method of proof by induction is intimately bound up with the construction of the natural numbers, with ordinal arithmetic, and with the so-called well ordering principle. We cannot treat all these logical niceties here. As with any good idea in mathematics, we shall nonetheless

be able to make it intuitively clear that the method is a valid and useful one. So no confusion should result.

Consider a statement $P(n)$ about the natural numbers. For example, the statement might be "The quantity $n^2 + 5n + 6$ is always even." If we wish to prove this statement, we might proceed as follows:

(1) Prove the statement $P(1)$.

(2) Prove that $P(k) \Rightarrow P(k+1)$ for every $k \in \{1, 2, \ldots\}$.

Let us apply the syllogism *modus ponendo ponens* from Subsection 0.3.2 to determine what we will have accomplished. We know $P(1)$ and, from **(2)** with $k = 1$, that $P(1) \Rightarrow P(2)$. We may therefore conclude $P(2)$. Now **(2)** with $k = 2$ says that $P(2) \Rightarrow P(3)$. We may then conclude $P(3)$. Continuing in this fashion, we may establish $P(n)$ for every natural number n.

Notice that this reasoning applies to any statement $P(n)$ for which we can establish **(1)** and **(2)** above. Thus **(1)** and **(2)** taken together constitute a method of proof. It is a method of establishing a statement $P(n)$ for every natural number n. The method is known as *proof by induction*.

EXAMPLE 11.3.1 Let us use the method of induction to prove that, for every natural number n, the number $n^2 + 5n + 6$ is even.

Solution: Our statement $P(n)$ is

$$\text{The number } n^2 + 5n + 6 \text{ is even.}$$

We now proceed in two steps:

$\mathbf{P(1)}$ is true. When $n = 1$ then

$$n^2 + 5n + 6 = 1^2 + 5 \cdot 1 + 6 = 12,$$

and this is certainly even. We have verified $P(1)$.

$\mathbf{P(n) \Rightarrow P(n+1)}$. We are proving an implication in this step. We *assume* $P(n)$ and *use it* to establish $P(n+1)$. Thus we are assuming that

$$n^2 + 5n + 6 = 2m$$

for some natural number m. Then, to check $P(n+1)$, we calculate

$$\begin{aligned}(n+1)^2+5(n+1)+6 &= [n^2+2n+1]+[5n+5]+6\\ &= [n^2+5n+6]+[2n+6]\\ &= 2m+[2n+6].\end{aligned}$$

Notice that in the last step we have *used our hypothesis* that n^2+5n+6 is even, that is that $n^2+5n+6=2m$. Now the last line may be rewritten as

$$2(m+n+3).$$

Thus we see that $(n+1)^2+5(n+1)+6$ is twice the natural number $m+n+3$. In other words, $(n+1)^2+5(n+1)+6$ is even. But that is the assertion $P(n+1)$.

In summary, assuming the assertion $P(n)$ we have established the assertion $P(n+1)$. That completes Step **(2)** of the method of induction. We conclude that $P(n)$ is true for every n. □

Here is another example to illustrate the method of induction.

Proposition 11.3.2 *If n is any natural number then*

$$1+2+\cdots+N=\frac{N\cdot(N+1)}{2}.$$

Remark: Of course we already gave a direct proof of this result in Section 11.1. Now we are learning proof by induction, so we give a new proof.

Proof: The statement $P(N)$ is

$$1+2+\cdots+N=\frac{N\cdot(N+1)}{2}.$$

Now let us follow the method of induction closely.

P(1) is true. The statement $P(1)$ is

$$1=\frac{1(1+1)}{2}.$$

This is plainly true.

$\mathbf{P(N) \Rightarrow P(N+1)}$. We are proving an implication in this step. We *assume* $P(N)$ and *use it* to establish $P(N+1)$. Thus we are assuming that

$$1 + 2 + \cdots + N = \frac{N \cdot (N+1)}{2}. \qquad (*)$$

Let us add the quantity $(N+1)$ to both sides of $(*)$. We obtain

$$1 + 2 + \cdots + N + (N+1) = \frac{N \cdot (N+1)}{2} + (N+1).$$

The left side of this last equation is exactly the left side of $P(N+1)$ that we are trying to establish. That is the motivation for our last step.

Now the righthand side may be rewritten as

$$\frac{N \cdot (N+1) + 2 \cdot (N+1)}{2}.$$

This simplifies to

$$\frac{(N+1) \cdot (N+2)}{2}.$$

In conclusion, we have established that

$$1 + 2 + \cdots + N + (N+1) = \frac{(N+1) \cdot (N+2)}{2}.$$

This is the statement $P(N+1)$.

Assuming the validity of $P(N)$, we have proved the validity of $P(N+1)$. That completes the second step of the method of induction, and establishes $P(N)$ for all N. □

Some problems are formulated in such a way that it is convenient to begin the induction with some value of n other than $n = 1$. The next example illustrates this notion:

EXAMPLE 11.3.3 Let us prove that, for $n \geq 4$, we have the inequality

$$3^n > 2n^2 + 3n.$$

Solution: The statement $P(n)$ is

$$3^n > 2n^2 + 3n.$$

$\mathbf{P(4)}$ is true. Observe that the inequality is false for $n = 1, 2, 3$. However for $n = 4$ it is certainly the case that

$$3^4 = 81 > 44 = 2 \cdot 4^2 + 3 \cdot 4.$$

$\mathbf{P(n) \Rightarrow P(n+1)}$. Now assume that $P(n)$ has been established and let us use it to prove $P(n+1)$. We are hypothesizing that

$$3^n > 2n^2 + 3n.$$

Multiplying both sides by 3 gives

$$3 \cdot 3^n > 3(2n^2 + 3n)$$

or

$$3^{n+1} > 6n^2 + 9n.$$

But now we have

$$\begin{aligned} 3^{n+1} &> 6n^2 + 9n \\ &= 2(n^2 + 2n + n) + (4n^2 + 3n) \\ &> 2(n^2 + 2n + 1) + (3n + 3) \\ &= 2(n+1)^2 + 3(n+1). \end{aligned}$$

This inequality is just $P(n+1)$, as we wished to establish. That completes step **(2)** of the induction, and therefore completes the proof. □

Chapter 12

Closing Thoughts

Physics has provided mathematics with many fine suggestions and new initiatives, but mathematics does not need to copy the style of experimental physics. Mathematics rests on proof—and proof is eternal.

Saunders Mac Lane

. . . we have ample experimental evidence for the truth of our identity and we may want to take it as something more than just a working assumption. We may want to formally introduce it into our mathematical system. What we need to avoid is the haphazard introduction of new axioms.

J. Borwein, P. Borwein, R. Girgensohn, and S. Parnes

Mathematics seeks to reduce complexity to a manageable level and also to impose structure where no structure is apparent.

Michael Aschbacher

I suppose you are two fathoms deep in mathematics and if you are, then God help you, for so am I, only with this difference. I stick fast in the mud at the bottom and there I shall remain.

Charles Darwin

All that is missing is a proof.

John Milnor

Never try to teach a pig to sing. It frustrates you and irritates the pig.

Anonymous

Life is good for only two things, discovering mathematics and teaching mathematics.

Simeon Poisson

I advise my students to listen carefully the moment they decide to take no more mathematics courses. They might be able to hear the sound of closing doors.

Anonymous

It looks simple at first sight, but reveals its subtle horrors to those who try to solve it.

S. Singh

. . . it is impossible to write out a very long and complicated argument without error, so is such a 'proof' really a proof?

Michael Aschbacher

12.1 Why Proofs are Important

Before proofs, about 2600 years ago, mathematics was a heuristic and phenomenological subject. Spurred largely (though not entirely) by practical considerations of land surveying, commerce, and counting, there seemed to be no real need for any kind of theory or rigor. It was only with the advent of abstract mathematics—or mathematics for its own sake—that it began to become clear why proofs are important. Indeed, proofs are central to the way that we view our discipline.

Today, there are tens of thousands of mathematicians all over the world. Just as an instance, the *Notices of the American Mathematical Society* has a circulation of about 30,000. [This is the news organ, and the journal of record, for the American Mathematical Society.] And abstract mathematics is a well-established discipline. There are few with any advanced knowledge of mathematics who would argue that proof no longer has a place in our subject. Proof is at the heart of the subject; it is what makes mathematics tick. Just as hand-eye coordination is at the heart of hitting a baseball, and practical technical insight is at the heart of being an engineer, and a sense of color and aesthetics is at the heart of being a painter, so an ability to appreciate and to create proofs is at the heart of being a mathematician.

If one were to remove "proof" from mathematics then all that would remain is a descriptive language. We could examine right triangles, and congruences, and parallel lines and attempt to learn something. We could look at pictures of fractals and make descriptive remarks. We could generate computer printouts and offer witty observations. We could let the computer crank out reams of numerical data and attempt to evaluate those data. We could post beautiful computer graphics and endeavor to assess them. *But we would not be doing mathematics.* Mathematics is **(i)** coming up with new ideas and **(ii)** validating those ideas by way of proof.[1] The timelessness and

[1]Mathematics is this and much more. This book has endeavored to portray mathe-

intrinsic value of the subject come from the methodology, and that methodology is proof.

And again we must emphasize that what sets "proof" apart from the methodologies of other disciplines is its timelessness. A great idea in computer science could easily be rendered "old school" in a couple of years. The languages `SNOBOL` and `COBOL` were hot in the 1960s, but hardly anyone uses them anymore. In the 1970s, `Fortran` was the definitive computer language for scientific computation. Today it has been superseded by `C` and `C++`.[2]

And so it is in medicine. You may recall that radial keratotomy was for a short time the hottest thing around for correcting vision through laser surgery. This lasted just a couple of years, until medical scientists realized that they could not accurately predict the long-term effects of the procedure. Now it has been replace by radial laser-assisted *in situ* keratomileusis or LASIK surgery. Now there are questions about the long-term effects of LASIK, so that methodology is being re-evaluated.

Artificial hearts were developed because patients were rejecting transplanted hearts. But now doctors have figured out how to get patients to accept the transplants; so artificial hearts are of less interest. It used to be that X-ray was the definitive diagnostic tool. Of course X-rays are still useful, but in many applications they are replaced by magnetic resonance imaging or one of the many other new imaging technologies that have been developed.

We have described some of these events in detail just to emphasize that this sort of thing *never* happens in mathematics. Certainly mathematics has its fashions and its prejudices. But, in mathematics, once correct is always correct. For a time, everyone in college studied spherical trigonometry; it was just part of the curriculum, like calculus is today. Then it fell out of fashion. But, in 1993, Wu-Yi Hsiang used spherical trigonometry to attempt a solution of the venerable Kepler sphere-packing problem. So this renewed interest in spherical trigonometry. Hyperbolic geometry was something of a relic of classical Riemannian geometry until Bill Thurston came along in the 1980s and made it part of his program to classify all 3-manifolds. Now everybody is studying hyperbolic geometry.

---

matics as a multi-faceted beast. In today's world, mathematics is proofs and algorithms (both proved and heuristic), theories, methodologies, approaches, conjectures, models, and much, much more.

[2]It should be noted that `Fortran` has the advantage of *not* being a structured programming language. So that it is still useful for programming parallel processing machines.

But it must be emphasized that spherical trigonometry and hyperbolic geometry were never declared *wrong*. There was never any danger of that happening. It's just that the world passed these subjects by for a while. Everybody knew they were there, and what they were good for. They simply did not attract any interest. There were too many new and exciting things to spend time on. But now, because of some good ideas of some excellent mathematicians, they have been brought to the fore again.

Proofs remain important in mathematics because they are our bellwether for what we can believe in, and what we can depend on. They are timeless and rigid and dependable. They are what hold the subject together, and what make it one of the glories of human thought.

12.2 Why It Is Important for our Notion of Proof to Evolve and Change

As described in this book, our concept of what proof *is* was developed and enunciated by the ancient Greeks 2500 years ago. They set for us a remarkable and profound paradigm from which we have not waivered in the intervening millenia. But ideas change and develop. Ultimately our goal is to reach the truth, to understand that truth, to verify that truth, and to disseminate and teach that truth. For a good part of our professional history mathematicians lived inside their heads. Their primary interest was in discovering new mathematics and validating it. Sharing it with others was of secondary interest. Now our view, and our value system, has changed.

Today there are more mathematicians than there were during the entire period 500 B.C.E. to 1950 C.E. Of course a good many mathematicians work at colleges and universities. There are 2850 colleges and universities in the United States alone, and many more thousands of institutions of higher learning around the world. There are also thousands of mathematicians employed in industry, in government research facilities (such as Los Alamos and Oak Ridge), and in research laboratories of all kinds. Because mathematics is so *diverse* and *diffuse*, it has become essential that there be more communication among mathematicians.

As we have described in the pages of this book, the notion of what it means to be a mathematician has grown and developed over time. Not long ago, a mathematician was an expert in Euclid's geometry and Newton's calculus

and Gibbs's vector analysis and a few other well worn and crusty subject areas. Today there are many different types of mathematicians with many different sorts of backgrounds. Some mathematicians work on the Genome Project. Some mathematicians work for NASA. Some work for Aerospace Corporation. Some work for the National Security Agency. Some work for financial firms on Wall Street. They often speak different languages and have different value systems. In order for mathematicians with different pedigrees to communicate effectively, we must be consciously aware of how different types of mathematical scientists approach their work. What sorts of problems do they study? What types of answers do they seek? How do they validate their work? What tools do they use?

It is for these reasons that it is essential that the mathematical community have a formal recognition of the changing and developing nature of mathematical proof. Certainly the classical notion of proof, taught to us by Euclid and Pythagoras, is the bedrock of our analytical thinking procedures. Nobody is advocating that we abandon or repudiate the logical basis for our subject. What *is* true is that many different points of view, many different processes, many different types of calculation, many different sorts of evidence may contribute to the development of our thoughts. And we should be welcoming to them all. One never knows where the next idea will come from, or how it may come to fruition. Since good ideas are so precious, and so hard to come by, we should not close any doors or turn away any opportunities.

Thus our notion of "proof" will develop and change. We may learn a lot from this evolution of mathematical thought, and we should. The advent of high-speed digital computers has allowed us to see things that we could never have seen before (using computer graphics and computer imaging) and to do "what if" calculations that were never before feasible. The development and proliferation of mathematical collaboration—both within and without the profession— has created new opportunities and taught us new ways to communicate. And, as part of the process, we have learned to speak new languages. Learning to talk to engineers is a struggle, but one side benefit of the process is that we gain the opportunity of learning many new problems. Likewise for physics and theoretical computer science and biology and medicine.

The great thing about going into mathematics in the twenty-first century is that it opens many doors and closes few of them. The world has become mathematized, and everyone is now conscious of this fact. People also appreciate that mathematicians have critical thinking skills, and are real problem

solvers. Law schools, medical schools, and many other postgraduate programs favor undergraduate math majors because they know that these are people who are trained to think. The ability to analyze mathematical arguments (i.e., *proofs*) and to solve mathematical problems is a talent that travels well and finds applications in many different contexts.

12.3 What Will Be Considered a Proof in 100 Years?

It is becoming increasingly evident that the delinations among "engineer" and "mathematician" and "physicist" are becoming ever more vague. The widely proliferated collaboration among these different groups is helping to erase barriers and to open up lines of communication. Although "mathematician" has historically been a much-honored and respected profession, one that represents the pinnacle of human thought, we may now fit that model into a broader context.

It seems plausible that in 100 years we will no longer speak of mathematicians as such but rather of *mathematical scientists*. This will include mathematicians to be sure, but also a host of others who use mathematics for analytical purposes. It would not be at all surprising if the notion of "Department of Mathematics" at the college and university level gives way to "Division of Mathematical Sciences".

In fact we already have a role model for this type of thinking at the California Institute of Technology (Caltech). For Caltech does not have departments at all. Instead it has divisions. There is a Division of Physical Sciences, which includes physics, mathematics, and astronomy. There is a Division of Life Sciences that includes Biology, Botany, and several other fields. The philosophy at Caltech is that departmental divisions tend to be rather artificial, and tend to cause isolation and lack of communication among people who would benefit distinctly from cross-pollination. This is just the type of symbiosis that we have been describing for mathematics in the preceding paragraphs.

So what will be considered a "proof" in the next century? There is every reason to believe that the traditional concept of pure mathematical proof will live on, and will be designated as such. But there will also be computer proofs, and proofs by way of physical experiment, and proofs by

way of numerical calculation. This author has participated in a project—connected with NASA's space shuttle program—that involved mathematicians, engineers, and computer scientists. The contributions from the different groups—some numerical, some analytical, some graphical—reinforced each other, and the end result was a rich tapestry of scientific effort. The end product is published in [CHE1] and [CHE2]. This type of collaboration, while rather the exception today, is likely to become ever more common as the field of applied mathematics grows, and as the need for interdisciplinary investigation proliferates.

Today many mathematics departments contain experts in computer graphics, experts in engineering problems, experts in numerical analysis, and experts in partial differential equations. These are all people who thrive on interdisciplinary work. And the role model that they play will influence those around them.

The Mathematics Department that is open to interdisciplinary work is one that is enriched and fulfilled in a pleasing variety of ways. Colloquium talks will cover a broad panorama of modern research. Visitors will come from a variety of backgrounds, and represent many different perspectives. Mathematicians will direct Ph.D. theses for students from engineering and physics and computer science and other disciplines as well. Conversely, mathematics students will find thesis advisors in many other departments. One already sees this happening with students studying wavelets and harmonic analysis and numerical analysis. The trend will broaden and continue.

So the answer to the question is that "proof" will live on, but it will take on new and varied meanings. The traditional idea of proof will prosper because it will interact with other types of verification and affirmation. And other disciplines, ones that do not traditionally use mathematical proof, will come to appreciate the value of this mode of intellectual discourse.[3] The end result will be a richer tapstry of mathematical science and mathematical work. We will all benefit as a result.

[3]We must repeat that some of these "other disciplines" are already impressively appreciative of the mathematical method. Both the *IEEE Transactions on Signal Processing* and the *IEEE Transactions on Image Processing* contain copious mathematics, and many results that are proved.

References

[ADL] A. Adler, The second fundamental forms of S^6 and $P^n(C)$, *Am. J. Math.* 91(1969), 657–670.

[AGA] Agrawal, primality in polynomial time, xxxxx

[ALM] F. J. Almgren, *Almgren's Big Regularity Paper. Q-Valued Functions Minimizing Dirichlet's Integral and the Regularity of Area-Minimizing Rectifiable Currents up to Codimension 2*, World Scientific Publishing Company, River Edge, NJ, 1200.

[APH1] K. Appel and W. Haken, A proof of the four color theorem, *Discrete Math.* 16(1976), 179–180.

[APH2] K. Appel and W. Haken, The four color proof suffices, *Math. Intelligencer* 8(1986), 10–20.

[ASC] M. Aschbacher, Highly complex proofs and implications of such proofs, *Phil. Trans. R. Soc.* A 363(2005), 2401–2406.

[ASM] M. Aschbacher and S. Smith, *The Classification of Quasithin Groups*, I and II, American Mathematical Society, Providence, RI, 2004.

[ASW] T. Aste and D. Weaire, *The Pursuit of Perfect Packing*, Institute of Physics Publishing, Bristol, U.K., 2000.

[ATI] M. Atiyah, Responses to Jaffe and Quinn 1994, *Bulletin of the AMS* 30(1994), 178–207.

[ATI2] M. Atiyah, Mathematics: Art and Science, *Bulletin of the AMS* 43(2006), 87–88.

[**AXK1**] J. Ax and S. Kochen, Diophantine problems over local fields. I., *Amer. J. Math.* 87(1965), 605–630.

[**AXK3**] J. Ax and S. Kochen, Diophantine problems over local fields. II. A complete set of axioms for p-adic number theory, *Amer. J. Math.* 87(1965), 631–648.

[**AXK3**] J. Ax and S. Kochen, Diophantine problems over local fields. III. Decidable fields, *Annals of Math.* 83(1966), 437–456.

[**BIS**] E. Bishop, *Foundations of Constructive Analysis*, McGraw-Hill, New York, 1967.

[**BIB**] E. Bishop and D. Bridges, *Constructive Analysis*, Springer-Verlag, New York, 1985.

[**BOO**] G. Boolos, Frege's theorem and the Peano postulates, *Bulletin of Symbolic Logic* 1(1995), 317–326.

[**BBGP**] J. Borwein, P. Borwein, R. Girgensohn, and S. Parnes, Making sense of experimental mathematics, *The Mathematical Intelligencer*18(1996), 12–18.

[**BRE**] D. Bressoud, *Proofs and Confirmations: The Story of the Alternating Sign Matrix Conjecture*, Cambridge University Press, Cambridge, 1999.

[**BRM**] R. Brooks and J. P. Matelski, The dynamics of 2-generator subgroups of $\mathrm{PSL}(2,\mathbb{C})$, *Riemann Surfaces and Related Topics*, Proceedings of the 1978 Stony Brook Conference, pp. 65-71, *Ann. of Math. Studies* 97, Princeton University Press, Princeton, NJ, 1981.

[**BMAM**] A. Bundy, D. MacKenzie, M. Atiyah, and A. MacIntyre, eds., The nature of mathematical proof, Proceedings of a Royal Society discussion meeting, *Phil. Trans. R. Soc. A* 363(2005), to appear.

[**CAZ**] H.-D. Cao and X.-P. Zhu, A complete proof of the Poincaré and geometrization conjectures—application of the Hamilton-Perelman theory of the Ricci flow, *Asian Journal of Mathematics* 10(2006), 165-498.

[**CJW**] J. Carlson, A. Jaffe, and A. Wiles, *The Millenium Prize Problems*, American Mathematical Society, Providence, RI, 2006.

[CHA1] G. J. Chaitin, *Algorithmic Information Theory*, Cambridge University Press, Cambridge and New York, 1992.

[CHA2] G. J. Chaitin, *The Limits of Mathematics: A Course on Information Theory and the Limits of Formal Reasoning*, Springer-Verlag, London, 2003.

[CHE1] G. Chen, S. G. Krantz, D. Ma, C. E. Wayne, and H. H. West, The Euler-Bernoulli beam equation with boundary energy dissipation, in *Operator Methods for Optimal Control Problems* (Sung J. Lee, ed.), Marcel Dekker, New York, 1988, 67-96.

[CHE2] G. Chen, S. G. Krantz, C. E. Wayne, and H. H. West, Analysis, designs, and behavior of dissipative joints for coupled beams, *SIAM Jr. Appl. Math.*, 49(1989), 1665-1693.

[CHO] S.-C. Chou and W. Schelter, Proving geometry theorems with rewrite rules, *Journal of Automated Reasoning* 2(1986), 253–273.

[CON] J. H. Conway, C. Goodman-Strauss, and N. J. A. Sloane, Recent progress in sphere packing,*Current Developments in Mathematics*, 1999 (Cambridge, MA), 37–76, International Press, Somerville, MA, 1999.

[COO] S. A. Cook, The complexity of theorem-proving procedures, *Proceedings of the 3rd Annual ACM Symposium on Theory of Computing*, Association for Computing Machinery, New York, 1971, 151–158.

[DAN] G. B. Dantzig, On the significance of solving linear programming problems with some integer variables, *Econometrica* 28 (1957), 30–44.

[DEB1] L. de Branges, *Hilbert Spaces of Entire Functions*, Prentice-Hall, Englewood Cliffs, NJ, 1968.

[DEB2] L. de Branges, A proof of the Bieberbach conjecture, *Acta Math.* 154(1985), 137–152.

[DEV1] K. Devlin, *The Millenium Problems: The Seven Greatest Unsolved Mathematical Puzzles of Our Time*, Basic Books, New York, 2003.

[EKZ] S. B. Ekhad and D. Zeilberger, A high-school algebra, "formal calculus", proof of the Bieberbach conjecture [after L. Weinstein], Jerusalem

Combinatorics '93, 113–115, *Contemporary Math.* 178, American Mathematical Society, Providence, RI, 1994.

[FTO] L. Fejes-Tóth, On close-packings of sphere in spaces of constant curvature, *Publ. Math. Debrecen* 3(1953), 158–167.

[FIP] C. Fitzgerald and C. Pommerenke, The de Branges theorem on univalent functions, *Trans. American Math. Society* 290(1985), 683–690.

[FRE1] G. Frege, *Begriffsschrift und andere Aufsätze*, Hildesheim, G. Olms, 1964.

[FRE2] G. Frege, *Grundgesetze der Arithmetik*, two volumes in one, Hildesheim, G. Olms, 1964.

[GAR] M. Gardner, *The Second Scientific American Book of Mathematical Puzzles and Diversions*, University of Chicago Press, Chicago, IL, 1987.

[GAJ] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman and Company, San Francisco, CA, 1991.

[GLS] D. Gorenstein, R. Lyons, and R. Solomon, *The Classification of the Finite Simple Groups*, American Mathematical Society, Providence, RI, 1994.

[GRA] J. Gray, *The Hilbert Challenge*, Oxford University Press, New York, 2000.

[GRT] B. Green and T. Tao, The primes contain arbitrarily long arithmetic progressions, *Annals of Math.*, to appear.

[GRE] M. J. Greenberg, *Euclidean and Non-Euclidean Geometries*, 2nd ed., W. H. Freeman, New York, 1980.

[HAL1] T. Hales, The status of the Kepler conjecture, *Math. Intelligencer* 16(1994), 47–58.

[HAL2] T. Hales, A proof of the Kepler conjecture, *Annals of Math.* 162(2005), 1065-1185.

[HAL] M. Hall, *The Theory of Groups*, MacMillan, New York, 1959.

[**HAR**] G. H. Hardy, *A Mathematician's Apology*, Cambridge University Press, London, 1967.

[**HER**] I. M. Herstein, *Topics in Algebra*, Xerox, Lexington, 1975.

[**HIN**] G. Higman and B. H. Neumann, Groups as gropoids with one law, *Publicationes Mathematicae Debreceu* 2(1952), 215–227.

[**HIA**] D. Hilbert and W. Ackermann, *Grundzüge der theoretischen Logik*, Springer-Verlag, Berlin, 1928.

[**HIL**] D. Hilbert, *Grundlagen der Geometrie*, Teubner, Leipzig, 1899.

[**HHM**] D. Hoffman, The computer-aided discovery of new embedded minimal surfaces, *Math. Intelligencer* 9(1987), 8–21.

[**HOR1**] J. Horgan, The Death of Proof?, *Scientific American* 269(1993), 93–103.

[**HOR2**] J. Horgan, *The End of Science*, Broadway Publishers, New York, 1997.

[**HOS**] E. Horowitz and S. Sahni, Exact and approximate algorithms for scheduling nonidentical processors, *J. Assoc. Comput. Mach.* 23 (1976), 317–327.

[**HRJ**] K. Hrbacek and T. Jech, *Introduction to Set Theory*, 3rd ed., Marcel Dekker, New York, 1999.

[**HSI1**] W.-Y. Hsiang, On the sphere packing problem and the proof of Kepler's conjecture, *Internat. J. Math.* 4(1993), 739–831.

[**HSI2**] W.-Y. Hsiang, Sphere packings and spherical geometry—Kepler's conjecture and beyond, Center for Pure and Applied Mathematics, U. C. Berkeley, July, 1991.

[**HSI3**] W.-Y. Hsiang, *Least Action Principle of Crystal Formation of Dense Packing Type and Kepler's Conjecture*, World Scientific, River Edge, NJ, 2001.

[**JAF**] Arthur Jaffe, Proof and the evolution of mathematics, *Synthese* 111(1997), 133–146.

[**JAQ**] Arthur Jaffe and F. Quinn, "Theoretical mathematics": toward a cultural synthesis of mathematics and theoretical physics, *Bulletin of the A.M.S.* 29(1993), 1–13.

[**JEC**] T. Jech, *The Axiom of Choice*, North-Holland, Amsterdam, 1973.

[**KAP1**] R. Kaplan and E. Kaplan, *The Nothing That Is: A Natural History of Zero*, Oxford University Press, Oxford, 2000.

[**KAP2**] R. Kaplan and E. Kaplan, *The Art of the Infinite: The Pleasures of Mathematics*, Oxford University Press, Oxford, 2003.

[**KAR**] R. M. Karp, Reducibility among combinatorial problems, in R. E. Miller and J. W. Thatcher, eds., *Complexity of Computer Computations*, Plenum Press, New York, 1972, 85–103.

[**KLL**] B. Kleiner and J. Lott, Notes on Perelman's papers, `arXiv:math.DG/0605667`.

[**KLI**] W. Klingenberg, *Lectures on Closed Geodesics*, Grundlehren der Mathematischen Wissenschaften, v. 230, Springer-Verlag, Berlin-New York, 1978.

[**KLN**] M. Kline, *Mathematics, The Loss of Certainty*, Oxford University Press, New York, 1980.

[**KRA1**] S. G. Krantz, The immortality of proof, *Notices of the AMS*, 41(1994), 10-13.

[**KRA2**] S. G. Krantz, *A Primer of Mathematical Writing*, American Mathematical Society, Providence, RI, 1996.

[**KRA3**] S. G. Krantz, *Mathematical Publishing: A Guidebook*, American Mathematical Society, Providence, RI, 2005.

[**KRA4**] S. G. Krantz, *Handbook of Logic and Proof Techniques for Computer Science*, Birkhäuser, Boston, 2002.

[**KRA5**] S. G. Krantz, *The Elements of Advanced Mathematics*, 2^{nd} ed., CRC Press, Boca Raton, FL, 2002.

[**KRA6**] S. G. Krantz, Review of *A New Kind of Science*, *Bull. AMS* 40(143–150).

[KUH] T. S. Kuhn, *The Structure of Scientific Revolutions*, 2nd ed., University of Chicago Press, Chicago, IL, 1970.

[KUN] K. Kunen, Single axioms for groups, *J. Automated Reasoning* 9(1992), 291–308.

[LAW] E. L. Lawler, A pseudopolynomial algorithm for sequencing jobs to minimize total tardiness, *Ann. Discrete Math.* 1(1977), 331–342.

[LRKF] J. K. Lenstra, A. H. G. Rinnooy Kan, and M. Florian, Deterministic production planning: algorithms and complexity, unpublished manuscript, 1978.

[LIT] J. E. Littlewood, *A Mathematician's Miscellany*, Methuen, London, 1953.

[LOV] L. Lovasz, Coverings and colorings of hypergraphs, *Proceedings of the 4th Southeastern Conference on Combinatorics, Graph Theory, and Computing*, Utilitas Mathematica Publishing, Winnipeg, 1973, 3–12.

[MAC] S. Mac Lane, Mathematical models, *Am. Math. Monthly* 88(1981), 462–472.

[MAN1] B. Mandelbrot, *The Fractal Geometry of Nature*, Freeman, New York, 1977.

[MAN2] B. Mandelbrot, Responses to "Theoretical mathematics: toward a cultural synthesis of mathematics and theoretical physics," by A. Jaffe and F. Quinn, *Bulletin of the AMS* 30(1994), 193–196.

[MAA] K. Manders and L. Adleman, NP-complete decision problems for binary quadratics, *J. Comput. System Sci.* 16(1978), 168–184.

[MAN] A. L. Mann, A complete proof of the Robbins conjecture, preprint.

[MCC] J. McCarthy, *Review of The Emperor's New Mind by Roger Penrose*, *Bull. AMS* 23(1990), 606–616.

[MCC] W. McCune, Single axioms for groups and abelian groups with various operations, *J. Automated Reasoning* 10(1993), 1–13.

[**MOT**] J. W. Morgan and G. Tian, *Ricci Flow and the Poincaré Conjecture*, The Clay Institute of Mathematics, American Mathematical Society, Providence, RI, to appear.

[**NAG**] S. Nasar and D. Gruber, Manifold Destiny, *The New Yorker*, August 28, 2006.

[**PEN1**] R. Penrose, *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*, Oxford University Press, Oxford, 1989.

[**PEN2**] R. Penrose, *The Road to Reality: A Complete Guide to the Laws of the Universe*, Jonathan Cape, London, 2004.

[**PER1**] G. Perelman, The entropy formula for the Ricci flow and its geometric applications, `arXiv:math.DG/0211159v1`.

[**PER2**] G. Perelman, Ricci flow with surgery on three-manifolds, `arXiv:math.DG/0303109v1`.

[**PER3**] G. Perelman, Finite extinction time for the solutions to the Ricci flow on certain three-manifolds, `arXiv:math.DG/0307245v1`.

[**POP**] K. R. Popper, *Conjectures and Refutations: The Growth of Scientific Knowledge*, Basic Books, New York, 1962.

[**RIE**] B. Riemann, On the number of primes less than a given magnitude, *Monthly Reports of the Berlin Academy*, 1859.

[**ROB**] A. Robinson, *Nonstandard Analysis*, North Holland, Amsterdam, 1966.

[**RUD**] W. Rudin, *Principles of Real Analysis*, 3rd ed., McGraw-Hill, New York, 1976.

[**RUS**] B. Russell, *History of Western Philosophy*, Routledge, London, 2004.

[**SAB**] Karl Sabbagh, *The Riemann Hypothesis: The Greatest Unsolved Problem in Mathematics*, Farrar, Straus, & Giroux, New York, 2003.

[**SAV**] M. vos Savant, *The World's Most Famous Math Problem*, St. Martin's Press, New York, 1993.

[**SCHA**] T. J. Schaefer, Complexity of some two-person perfect-infor- mation games, *J. Comput. Syst. Sci.* 16(1978), 185–225.

[SCL] C. P. Schnorr and H. W. Lenstra, Jr., A Monte Carlo factoring algorithm with linear storage, *Math. Comput.* 43(1984), 289–311,

[SEY] P. Seymour, Progress on the four-color theorem, *Proceedings of the ICM* (Zürich, 1994), 183–195, Birkhäuser, Basel, 1995.

[SMA] S. Smale, Review of E. C. Zeeman: Catastrophe Theory, Selected Papers 1972–1977, *Bulletin AMS* 84(1978), 1360–1368.

[SMU1] R. Smullyan, *Forever Undecided: A Puzzle Guide to Gödel*, Alfred Knopf, New York, 1987.

[SMU2] R. Smullyan, *The Lady or the Tiger? and Other Logic Puzzles*, Times Books, New York, 1992.

[STI] M. Stickel, A case study of theorem proving by the Knuth-Bendix method: discovering that $x^3 = x$ implies ring commutativity, *Proceedings of the Seventh International Conference on Automated Deduction*, R. E. Shostak, ed., Springer=Verlag, New York, 1984, pp. 248–258.

[STM] L. J. Stockmeyer and A. R. Meyer, Word problems requiring exponential time, *Proceedings of the* 5*th Annual ACM Symposium on Theory of Computing*, Association for Computing Machinery, New York, 1973, 1–9.

[STR] R. S. Strichartz, letter to the editor, *Notices of the American Mathematical Society* 53(2006), 406.

[STR1] W. R. Stromquist, Some Aspects of the Four Color Problem, Ph.D. thesis, Harvard University, 1975.

[STR2] W. R. Stromquist, The four-color thoerem for small maps, *J. Combinatorial Theory* 19(1975), 256–268.

[STRZ] Pawel Strzelecki, The Poincaré conjecture?, *American Mathematical Monthly* 113(2006), 75–78.

[SUP] P. Suppes, *Axiomatic Set Theory*, Van Nostrand, Princeton, 1972.

[THU1] W. P. Thurston, On proof and progress in mathematics, *Bull. AMS* 30(1994), 161–177.

[**THU2**] W. P. Thurston, *Three-Dimensional Geometry and Topology*, Vol. 1, Princeton University Press, Princeton, 1997.

[**THU3**] W. P. Thurston, *The Geometry and Topology of Three-Manifolds*, notes, Princeton University, 1980, 502 pp.

[**WAG**] S. Wagon, *The Banach-Tarski Paradox*, Cambridge University Press, Cambridge and New York, 1985.

[**WEIL1**] A. Weil, *Basic Number Theory*, 2nd ed., Springer-Verlag, New York, 1973.

[**WEIL2**] A. Weil, *The Apprenticeship of a Mathematician*, Birkhäuser, Boston, MA, 1992.

[**WEI**] L. Weinstein, The Bieberbach conjecture, *International Math. Res. Notices* 5(1991), 61–64.

[**WIG**] E. Wigner, The unreasonable effectiveness of mathematics in the natural sciences, *Comm. Pure App. Math.* 13(1960), 1–14.

[**WRU**] A. N. Whitehead and B. Russell, *Principia Mathematica*, Cambridge University Press, Cambridge, 1910.

[**WIL**] A. Wiles, Modular elliptic curves and Fermat's last theorem, *Annals of Math.* 141(1995), 443–551.

[**WILS**] L. Wilson, *The Academic Man, A Study in the Sociology of a Profession*, Oxford University Press, London, 1942.

[**WOLF**] R. S. Wolf, *A Tour Through Mathematical Logic*, A Carus Monograph of the Mathematical Association of America, Washington, D.C., 2005.

[**WOL**] S. Wolfram, *A New Kind of Science*, Wolfram Media, Inc., Champaign, IL, 2002.

[**WOS1**] L. Wos, *Automated Reasoning: Introduction and Applications*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

[**WOS2**] L. Wos, *Automation of Reasoning: An Experimenter's Notebook with Otter Tutorial*, Academic Press, New York, 1996.

[WOS3] L. Wos, Automated reasoning answers open questions, *Notices of the AMS* 40(1993), 15–26.

[YAN] B. H. Yandell, *The Honors Class*, A. K. Peters, Natick, MA, 2002.

[ZEE] E. C. Zeeman, Controversy in science: on the ideas of Daniel Bernoulli and René Thom, The 1992/3 Johann Bernoulli Lecture, Gröningen, *Nieuw Archief van de Wiskunde* 11(1993), 257–282.

[ZEI] D. Zeilberger, Theorems for a price: Tomorrow's semi-rigorous mathematical culture, *Notices of the AMS* 40(1993), 978–981.