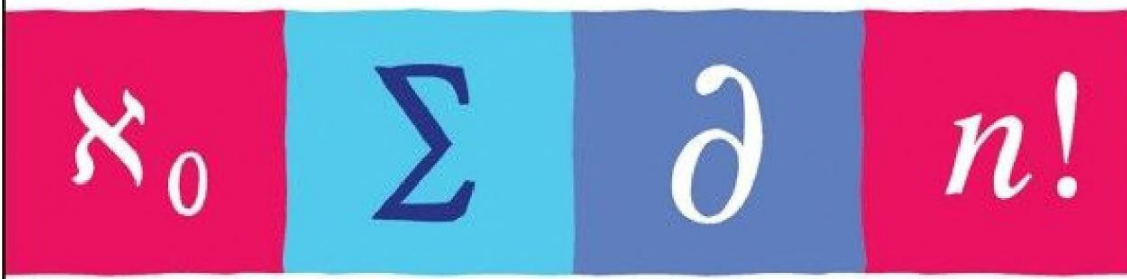
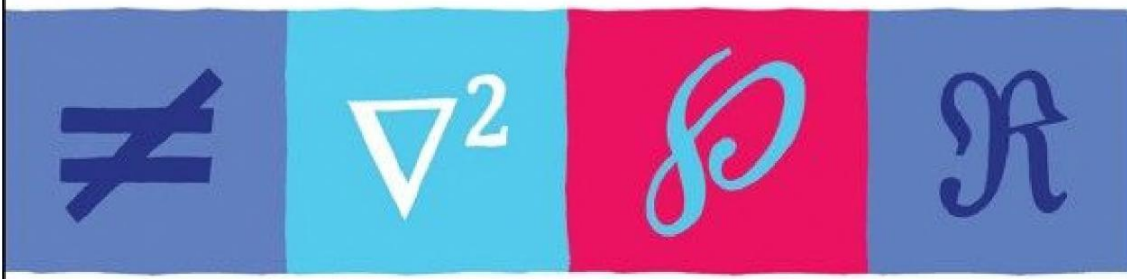
'Britain's most brilliant and prolific populariser of mathematics' *Guardian*

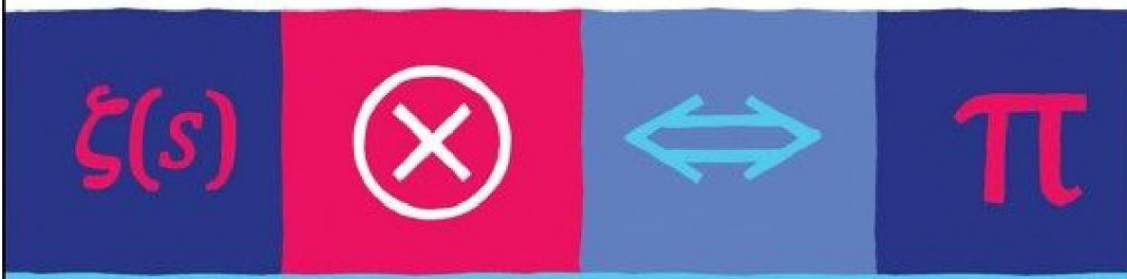
THE GREAT



MATHEMATICAL



PROBLEMS



IAN STEWART

with Terry Pratchett and Jack Cohen

The Science of Discworld

The Science of Discworld II: The Globe

The Science of Discworld III: Darwin's Watch

with Jack Cohen

The Collapse of Chaos

Figments of Reality

Evolving the Alien/What Does a Martian Look Like?

Wheelers (science fiction)

Heaven (science fiction)

First published in Great Britain in 2013 by

PROFILE BOOKS LTD

3A Exmouth House

Pine Street

London EC1R 0JH

www.profilebooks.com

1 3 5 7 9 10 8 6 4 2

Printed and bound in Great Britain by Clays, Bungay, Suffolk

A CIP catalogue record for this book is available from the British Library.

ISBN 978 1 84668 1998

eISBN 978 184765 3512

Contents

We must know. We shall know.
David Hilbert

Speech about mathematical problems in 1930, on the occasion of his honorary citizenship of Königsberg.[1]

Preface

Mathematics is a vast, ever-growing, ever-changing subject. Among the innumerable questions that mathematicians ask, and mostly answer, some stand out from the rest: prominent peaks that tower over the lowly foothills. These are the really big questions, the difficult and challenging problems that any mathematician would give his or her right arm to solve. Some remained unanswered for decades, some for centuries, a few for millennia. Some have yet to be conquered. Fermat's last theorem was an enigma for 350 years until Andrew Wiles dispatched it after seven years of toil. The Poincaré conjecture stayed open for over a century until it was solved by the eccentric genius Grigori Perelman, who declined all academic honours and a million-dollar prize for his work. The Riemann hypothesis continues to baffle the world's mathematicians, impenetrable as ever after 150 years.

The Great Mathematical Problems contains a selection of the really big questions that have driven the mathematical enterprise in radically new directions. It describes their origins, explains why they are important, and places them in the context of mathematics and science as a whole. It includes both solved and unsolved problems, which range over more than two thousand years of mathematical development, but its main focus is on questions that either remain open today, or have been solved within the past fifty years.

A basic aim of mathematics is to uncover the underlying simplicity of apparently complicated questions. This may not always be apparent, however, because the mathematician's conception of 'simple' relies on many technical and difficult concepts. An important feature of this book is to emphasise the deep simplicities, and avoid – or at the very least explain in straightforward terms – the complexities.

Mathematics is newer, and more diverse, than most of us imagine. At a rough estimate, the world's research

mathematicians number about a hundred thousand, and they produce more than *two million* pages of new mathematics every year. Not 'new numbers', which are not what the enterprise is really about. Not 'new sums' like existing ones, but bigger – though we do work out some pretty big sums. One recent piece of algebra, carried out by a team of some 25 mathematicians, was described as 'a calculation the size of Manhattan'. That wasn't quite true, but it erred on the side of conservatism. The *answer* was the size of Manhattan; the calculation was a lot bigger. That's impressive, but what matters is quality, not quantity. The Manhattan-sized calculation qualifies on both counts, because it provides valuable basic information about a symmetry group that seems to be important in quantum physics, and is definitely important in mathematics. Brilliant mathematics can occupy one line, or an encyclopaedia – whatever the problem demands.

When we think of mathematics, what springs to mind is endless pages of dense symbols and formulas. However, those two million pages generally contain more words than symbols. The words are there to explain the background to the problem, the flow of the argument, the meaning of the calculations, and how it all fits into the evergrowing edifice of mathematics. As the great Carl Friedrich Gauss remarked around 1800, the essence of mathematics is 'notions, not notations'. Ideas, not symbols. Even so, the usual language for expressing mathematical ideas is symbolic. Many published research papers do contain more symbols than words. Formulas have a precision that words cannot always match.

However, it is often possible to explain the ideas while leaving out most of the symbols. *The Great Mathematical Problems* takes this as its guiding principle. It illuminates what mathematicians do, how they think, and why their subject is interesting and important. Significantly, it shows how today's mathematicians are rising to the challenges set by their predecessors, as one by one the great enigmas of the past surrender to the powerful techniques of the present, which changes the mathematics and science of the future. Mathematics ranks among humanity's greatest achievements,

and its great problems – solved and unsolved – have guided and stimulated its astonishing power for millennia, both past and yet to come.

Coventry, June 2012

Figure Credits

Fig. 31 http://random.mostlymaths.net.

Fig. 33 Carles Simó. From: *European Congress of Mathematics, Budapest 1996*, Progress in Mathematics 168, Birkhäuser, Basel.

Fig. 43 Pablo Mininni.

Fig. 46 University College, Cork, Ireland.

Fig. 50 Wolfram MathWorld.

1
Great problems

TELEVISION PROGRAMMES ABOUT MATHEMATICS are rare, good ones rarer. One of the best, in terms of audience involvement and interest as well as content, was Fermat's last theorem. The programme was produced by John Lynch for the British Broadcasting Corporation's flagship popular science series *Horizon* in 1996. Simon Singh, who was also involved in its making, turned the story into a spectacular bestselling book.[2] On a website, he pointed out that the programme's stunning success was a surprise:

> It was 50 minutes of mathematicians talking about mathematics, which is not the obvious recipe for a TV blockbuster, but the result was a programme that captured the public imagination and which received critical acclaim. The programme won the BAFTA for best documentary, a Priz Italia, other international prizes and an Emmy nomination – this proves that mathematics can be as emotional and as gripping as any other subject on the planet.

I think that there are several reasons for the success of both the television programme and the book and they have implications for the stories I want to tell here. To keep the discussion focused, I'll concentrate on the television documentary.

Fermat's last theorem is one of the truly great mathematical problems, arising from an apparently innocuous remark which one of the leading mathematicians of the seventeenth century wrote in the margin of a classic textbook. The problem became notorious because no one could prove what Pierre de Fermat's marginal note claimed, and this state of affairs continued for more than 300 years despite strenuous efforts by extraordinarily clever people. So when the British mathematician Andrew Wiles finally cracked the problem in 1995, the magnitude of his achievement was obvious to anyone. You didn't even need to know what the problem was,

let alone how he had solved it. It was the mathematical equivalent of the first ascent of Mount Everest.

In addition to its significance for mathematics, Wiles's solution also involved a massive human-interest story. At the age of ten, he had become so intrigued by the problem that he decided to become a mathematician and solve it. He carried out the first part of the plan, and got as far as specialising in number theory, the general area to which Fermat's last theorem belongs. But the more he learned about real mathematics, the more impossible the whole enterprise seemed. Fermat's last theorem was a baffling curiosity, an isolated question of the kind that any number theorist could dream up without a shred of convincing evidence. It didn't fit into any powerful body of technique. In a letter to Heinrich Olbers, the great Gauss had dismissed it out of hand, saying that the problem had 'little interest for me, since a multitude of such propositions, which one can neither prove nor refute, can easily be formulated'.[3] Wiles decided that his childhood dream had been unrealistic and put Fermat on the back burner. But then, miraculously, other mathematicians suddenly made a breakthrough that linked the problem to a core topic in number theory, one on which Wiles was already an expert. Gauss, uncharacteristically, had underestimated the problem's significance, and was unaware that it could be linked to a deep, though apparently unrelated, area of mathematics.

With this link established, Wiles could now work on Fermat's enigma *and* do credible research in modern number theory at the same time. Better still, if Fermat didn't work out, anything significant that he discovered while trying to prove it would be publishable in its own right. So off the back burner it came, and Wiles began to think about Fermat's problem in earnest. After seven years of obsessive research, carried on in private and in secret – an unusual precaution in mathematics – he became convinced that he had found a solution. He delivered a series of lectures at a prestigious number theory conference, under an obscure title that fooled no one.[4] The exciting news broke, in the media as well as the halls of academe: Fermat's last theorem had been proved.

The proof was impressive and elegant, full of good ideas. Unfortunately, experts quickly discovered a serious gap in its logic. In attempts to demolish great unsolved problems of mathematics, this kind of development is depressingly common, and it almost always proves fatal. However, for once the Fates were kind. With assistance from his former student Richard Taylor, Wiles managed to bridge the gap, repair the proof, and complete his solution. The emotional burden involved became vividly clear in the television programme: it must have been the only occasion when a mathematician has burst into tears on screen, just recalling the traumatic events and the eventual triumph.

You may have noticed that I haven't told you what Fermat's last theorem *is*. That's deliberate; it will be dealt with in its proper place. As far as the success of the television programme goes, it doesn't actually matter. In fact, mathematicians have never greatly cared whether the theorem that Fermat scribbled in his margin is true or false, because nothing of great import hangs on the answer. So why all the fuss? Because a huge amount hangs on the inability of the mathematical community to *find* the answer. It's not just a blow to our self-esteem: it means that existing mathematical theories are missing something vital. In addition, the theorem is very easy to state; this adds to its air of mystery. How can something that seems so simple turn out to be so hard?

Although mathematicians didn't really care about the answer, they cared deeply that they didn't know what it was. And they cared even more about finding a method that could solve it, because that must surely shed light not just on Fermat's question, but on a host of others. This is often the case with great mathematical problems: it is the methods used to solve them, rather than the results themselves, that matter most. Of course, sometimes the actual result matters too: it depends on what its consequences are.

Wiles's solution is much too complicated and technical for television; in fact, the details are accessible only to specialists.[5] The proof does involve a nice mathematical story, as we'll see in due course, but any attempt to explain that on

television would have lost most of the audience immediately. Instead, the programme sensibly concentrated on a more personal question: what is it like to tackle a notoriously difficult mathematical problem that carries a lot of historical baggage? Viewers were shown that there existed a small but dedicated band of mathematicians, scattered across the globe, who cared deeply about their research area, talked to each other, took note of each other's work, and devoted a large part of their lives to advancing mathematical knowledge. Their emotional investment and social interaction came over vividly. These were not clever automata, but real people, engaged with their subject. That was the message.

Those are three big reasons why the programme was such a success: a major problem, a hero with a wonderful human story, and a supporting cast of emotionally involved people. But I suspect there was a fourth, not quite so worthy. The majority of non-mathematicians seldom hear about new developments in the subject, for a variety of perfectly sensible reasons: they're not terribly interested anyway; newspapers hardly ever mention anything mathematical; when they do, it's often facetious or trivial; and nothing much in daily life seems to be affected by whatever it is that mathematicians are doing behind the scenes. All too often, school mathematics is presented as a closed book in which every question has an answer. Students can easily come to imagine that new mathematics is as rare as hen's teeth.

From this point of view, the big news was not that Fermat's last theorem had been proved. It was that at last *someone had done some new mathematics*. Since it had taken mathematicians more than 300 years to find a solution, many viewers subconsciously concluded that the breakthrough was the first important new mathematics discovered in the last 300 years. I'm not suggesting that they *explicitly* believed that. It ceases to be a sustainable position as soon as you ask some obvious questions, such as 'Why does the Government spend good money on university mathematics departments?' But subconsciously it was a common default assumption, unquestioned and unexamined. It made the magnitude of

Wiles's achievement seem even greater.

One of the aims of this book is to show that mathematical research is thriving, with new discoveries being made all the time. You don't hear much about this activity because most of it is too technical for non-specialists, because most of the media are wary of anything intellectually more challenging than *The X Factor*, and because the applications of mathematics are deliberately hidden away to avoid causing alarm. 'What? My iPhone depends on advanced mathematics? How will I log in to Facebook when I failed my maths exams?'

Historically, new mathematics often arises from discoveries in other areas. When Isaac Newton worked out his laws of motion and his law of gravity, which together describe the motion of the planets, he did not polish off the problem of understanding the solar system. On the contrary, mathematicians had to grapple with a whole new range of questions: yes, we know the laws, but what do they imply? Newton invented calculus to answer that question, but his new method also has limitations. Often it rephrases the question instead of providing the answer. It turns the problem into a special kind of formula, called a differential equation, whose *solution* is the answer. But you still have to solve the equation. Nevertheless, calculus was a brilliant start. It showed us that answers were possible, and it provided one effective way to seek them, which continues to provide major insights more than 300 years later.

As humanity's collective mathematical knowledge grew, a second source of inspiration started to play an increasing role in the creation of even more: the internal demands of mathematics itself. If, for example, you know how to solve algebraic equations of the first, second, third, and fourth degree, then you don't need much imagination to ask about the fifth degree. (The degree is basically a measure of complexity, but you don't even need to know what it is to ask the obvious question.) If a solution proves elusive, as it did, that fact *alone* makes mathematicians even more determined to find an answer, whether or not the result has useful

applications.

I'm not suggesting applications don't matter. But if a particular piece of mathematics keeps appearing in questions about the physics of waves – ocean waves, vibrations, sound, light – then it surely makes sense to investigate the gadget concerned in its own right. You don't need to know ahead of time exactly how any new idea will be used: the topic of waves is common to so many important areas that significant new insights are bound to be useful for something. In this case, those somethings included radio, television, and radar.[6] If somebody thinks up a new way to understand heat flow, and comes up with a brilliant new technique that unfortunately lacks proper mathematical support, then it makes sense to sort the whole thing out *as a piece of mathematics*. Even if you don't give a fig about how heat flows, the results might well be applicable elsewhere. Fourier analysis, which emerged from this particular line of investigation, is arguably the most useful single mathematical idea ever found. It underpins modern telecommunications, makes digital cameras possible, helps to clean up old movies and recordings, and a modern extension is used by the FBI to store fingerprint records.[7]

After a few thousand years of this kind of interchange between the external uses of mathematics and its internal structure, these two aspects of the subject have become so densely interwoven that picking them apart is almost impossible. The mental attitudes involved are more readily distinguishable, though, leading to a broad classification of mathematics into two kinds: pure and applied. This is defensible as a rough-and-ready way to locate mathematical ideas in the intellectual landscape, but it's not a terribly accurate description of the subject itself. At best it distinguishes two ends of a continuous spectrum of mathematical styles. At worst, it misrepresents which parts of the subject are useful and where the ideas come from. As with all branches of science, what gives mathematics its power is the *combination* of abstract reasoning and inspiration from the outside world, each feeding off the other. Not only is it impossible to pick the two strands apart: it's pointless.

Most of the really important mathematical problems, the great problems that this book is about, have arisen within the subject through a kind of intellectual navel-gazing. The reason is simple: they are *mathematical* problems. Mathematics often looks like a collection of isolated areas, each with its own special techniques: algebra, geometry, trigonometry, analysis, combinatorics, probability. It tends to be taught that way, with good reason: locating each separate topic in a single well-defined area helps students to organise the material in their minds. It's a reasonable first approximation to the structure of mathematics, especially long-established mathematics. At the research frontiers, however, this tidy delineation often breaks down. It's not just that the boundaries between the major areas of mathematics are blurred. It's that they don't really exist.

Every research mathematician is aware that, at any moment, suddenly and unpredictably, the problem they are working on may turn out to require ideas from some apparently unrelated area. Indeed, new research often combines areas. For instance, my own research mostly centres on pattern formation in dynamical systems, systems that change over time according to specific rules. A typical example is the way animals move. A trotting horse repeats the same sequence of leg movements over and over again, and there is a clear pattern: the legs hit the ground together in diagonally related pairs. That is, first the front left and back right legs hit, then the other two. Is this a problem about patterns, in which case the appropriate methods come from group theory, the algebra of symmetry? Or is it a problem about dynamics, in which case the appropriate area is Newtonian-style differential equations?

The answer is that, by definition, it has to be both. It is not their intersection, which would be the material they have in common – basically, nothing. Instead, it is a new 'area', which straddles two of the traditional divisions of mathematics. It is like a bridge across a river that separates two countries; it links the two, but belongs to neither. But this bridge is not a thin strip of roadway; it is comparable in size to each of the countries. Even more vitally, the methods involved are not limited to those two areas. In fact, virtually every course in mathematics

that I have ever studied has played a role somewhere in my research. My Galois theory course as an undergraduate at Cambridge was about how to solve (more precisely, why we can't solve) an algebraic equation of the fifth degree. My graph theory course was about networks, dots joined by lines. I never took a course in dynamical systems, because my PhD was in algebra, but over the years I picked up the basics, from steady states to chaos. Galois theory, graph theory, dynamical systems: three separate areas. Or so I assumed until 2011, when I wanted to understand how to detect chaotic dynamics in a network of dynamical systems, and a crucial step depended on things I'd learned 45 years earlier in my Galois theory course.

Mathematics, then, is not like a political map of the world, with each speciality neatly surrounded by a clear boundary, each country tidily distinguished from its neighbours by being coloured pink, green, or pale blue. It is more like a natural landscape, where you can never really say where the valley ends and the foothills begin, where the forest merges into woodland, scrub, and grassy plains, where lakes insert regions of water into every other kind of terrain, where rivers link the snow-clad slopes of the mountains to the distant, low-lying oceans. But this ever-changing mathematical landscape consists not of rocks, water, and plants, but of ideas; it is tied together not by geography, but by logic. And it is a dynamic landscape, which changes as new ideas and methods are discovered or invented. Important concepts with extensive implications are like mountain peaks, techniques with lots of uses are like broad rivers that carry travellers across the fertile plains. The more clearly defined the landscape becomes, the easier it is to spot unscaled peaks, or unexplored terrain that creates unwanted obstacles. Over time, some of the peaks and obstacles acquire iconic status. These are the great problems.

What makes a great mathematical problem great? Intellectual depth, combined with simplicity and elegance. Plus: it has to be *hard*. Anyone can climb a hillock; Everest is another matter

entirely. A great problem is usually simple to state, although the terms required may be elementary or highly technical. The statements of Fermat's last theorem and the four colour problem make immediate sense to anyone familiar with school mathematics. In contrast, it is impossible even to state the Hodge conjecture or the mass gap hypothesis without invoking deep concepts at the research frontiers – the latter, after all, comes from quantum field theory. However, to those versed in such areas, the statement of the question concerned is simple and natural. It does not involve pages and pages of dense, impenetrable text. In between are problems that require something at the level of undergraduate mathematics, if you want to understand them in complete detail. A more general feeling for the essentials of the problem – where it came from, why it's important, what you could do if you possessed a solution – is usually accessible to any interested person, and that's what I will be attempting to provide. I admit that the Hodge conjecture is a hard nut to crack in that respect, because it is very technical and very abstract. However, it is one of the seven Clay Institute millennium mathematics problems, with a million-dollar prize attached, and it absolutely must be included.

Great problems are creative: they help to bring new mathematics into being. In 1900 David Hilbert delivered a lecture at the International Congress of Mathematicians in Paris, in which he listed 23 of the most important problems in mathematics. He didn't include Fermat's last theorem, but he mentioned it in his introduction. When a distinguished mathematician lists what he thinks are some of the great problems, other mathematicians pay attention. The problems wouldn't be on the list unless they were important, and hard. It is natural to rise to the challenge, and try to answer them. Ever since, solving one of Hilbert's problems has been a good way to win your mathematical spurs. Many of these problems are too technical to include here, many are open-ended programmes rather than specific problems, and several appear later in their own right. But they deserve to be mentioned, so I've put a brief summary in the notes.[8]

That's what makes a great mathematical problem great. What makes it problematic is seldom deciding what the answer should be. For virtually all great problems, mathematicians have a very clear idea of what the answer ought to be – or had one, if a solution is now known. Indeed, the statement of the problem often includes the expected answer. Anything described as a conjecture is like that: a plausible guess, based on a variety of evidence. Most well-studied conjectures eventually turn out to be correct, though not all. Older terms like hypothesis carry the same meaning, and in the Fermat case the word 'theorem' is (more precisely, was) abused – a theorem requires a proof, but that was precisely what was missing until Wiles came along.

Proof, in fact, is the requirement that makes great problems problematic. Anyone moderately competent can carry out a few calculations, spot an apparent pattern, and distil its essence into a pithy statement. Mathematicians demand more evidence than that: they insist on a complete, logically impeccable proof. Or, if the answer turns out to be negative, a disproof. It isn't really possible to appreciate the seductive allure of a great problem without appreciating the vital role of proof in the mathematical enterprise. Anyone can make an educated guess. What's hard is to prove it's right. Or wrong.

The concept of mathematical proof has changed over the course of history, with the logical requirements generally becoming more stringent. There have been many highbrow philosophical discussions of the nature of proof, and these have raised some important issues. Precise logical definitions of 'proof' have been proposed and implemented. The one we teach to undergraduates is that a proof begins with a collection of explicit assumptions called axioms. The axioms are, so to speak, the rules of the game. Other axioms are possible, but they lead to different games. It was Euclid, the ancient Greek geometer, who introduced this approach to mathematics, and it is still valid today. Having agreed on the axioms, a proof of some statement is a series of steps, each of which is a logical consequence of either the axioms, or previously proved statements, or both. In effect, the mathematician is exploring a

logical maze, whose junctions are statements and whose passages are valid deductions. A proof is a path through the maze, starting from the axioms. What it proves is the statement at which it terminates.

However, this tidy concept of proof is not the whole story. It's not even the most important part of the story. It's like saying that a symphony is a sequence of musical notes, subject to the rules of harmony. It misses out all of the creativity. It doesn't tell us how to find proofs, or even how to validate other people's proofs. It doesn't tell us which locations in the maze are significant. It doesn't tell us which paths are elegant and which are ugly, which are important and which are irrelevant. It is a formal, mechanical description of a process that has many other aspects, notably a human dimension. Proofs are discovered by people, and research in mathematics is not just a matter of step-by-step logic.

Taking the formal definition of proof literally can lead to proofs that are virtually unreadable, because most of the time is spent dotting logical i's and crossing logical t's in circumstances where the outcome already stares you in the face. So practising mathematicians cut to the chase, and leave out anything that is routine or obvious. They make it clear that there's a gap by using stock phrases like 'it is easy to verify that' or 'routine calculations imply'. What they don't do, at least not consciously, is to slither past a logical difficulty and to try to pretend it's not there. In fact, a competent mathematician will go out of his or her way to point out exactly those parts of the argument that are logically fragile, and they will devote most of their time to explaining how to make them sufficiently robust. The upshot is that a proof, in practice, is a mathematical story with its own narrative flow. It has a beginning, a middle, and an end. It often has subplots, growing out of the main plot, each with its own resolution. The British mathematician Christopher Zeeman once remarked that a theorem is an intellectual resting point. You can stop, get your breath back, and feel you've got somewhere definite. The subplot ties off a loose end in the main story. Proofs resemble narratives in other ways: they often have one or more central characters – ideas

rather than people, of course – whose complex interactions lead to the final revelation.

As the undergraduate definition indicates, a proof starts with some clearly stated assumptions, derives logical consequences in a coherent and structured way, and ends with whatever it is you want to prove. But a proof is not just a list of deductions, and logic is not the sole criterion. A proof is a story told to and dissected by people who have spent much of their life learning how to read such stories and find mistakes or inconsistencies: people whose main aim is to prove the storyteller *wrong*, and who possess the uncanny knack of spotting weaknesses and hammering away at them until they collapse in a cloud of dust. If any mathematician claims to have solved a significant problem, be it a great one or something worthy but less exalted, the professional reflex is not to shout 'hurray!' and sink a bottle of champagne, but to try to shoot it down.

That may sound negative, but proof is the only reliable tool that mathematicians have for making sure that what they say is correct. Anticipating this kind of response, researchers spend a lot of their effort trying to shoot their own ideas and proofs down. It's less embarrassing that way. When the story has survived this kind of critical appraisal, the consensus soon switches to agreement that it is correct, and at that point the inventor of the proof receives appropriate praise, credit, and reward. At any rate, that's how it usually works out, though it may not always seem that way to the people involved. If you're close to the action, your picture of what's going on may be different from that of a more detached observer.

How do mathematicians solve problems? There have been few rigorous scientific studies of this question. Modern educational research, based on cognitive science, largely focuses on education up to high school level. Some studies address the teaching of undergraduate mathematics, but those are relatively few. There are significant differences between learning and teaching existing mathematics and creating new mathematics. Many of us can play a musical instrument, but

far fewer can compose a concerto or even write a pop song.

When it comes to creativity at the highest levels, much of what we know – or think we know – comes from introspection. We ask mathematicians to explain their thought processes, and seek general principles. One of the first serious attempts to find out how mathematicians think was Jacques Hadamard's *The Psychology of Invention in the Mathematical Field*, first published in 1945.[9] Hadamard interviewed leading mathematicians and scientists of his day and asked them to describe how they thought when working on difficult problems. What emerged, very strongly, was the vital role of what for lack of a better term must be described as intuition. Some feature of the subconscious mind guided their thoughts. Their most creative insights did not arise through step by step logic, but by sudden, wild leaps.

One of the most detailed descriptions of this apparently illogical approach to logical questions was provided by the French mathematician Henri Poincaré, one of the leading figures of the late nineteenth and early twentieth centuries. Poincaré ranged across most of mathematics, founding several new areas and radically changing many others. He plays a prominent role in several later chapters. He also wrote popular science books, and this breadth of experience may have helped him to gain a deeper understanding of his own thought processes. At any rate, Poincaré was adamant that conscious logic was only part of the creative process. Yes, there were times when it was indispensable: deciding what the problem really was, systematically verifying the answer. But in between, Poincaré felt that his brain was often working on the problem without telling him, in ways that he simply could not fathom.

His outline of the creative process distinguished three key stages: preparation, incubation, and illumination. Preparation consists of conscious logical efforts to pin the problem down, make it precise, and attack it by conventional methods. This stage Poincaré considered essential: it gets the subconscious going and provides raw materials for it to work with. Incubation takes place when you stop thinking about the problem and go

off and do something else. The subconscious now starts combining ideas with each other, often quite wild ideas, until light starts to dawn. With luck, this leads to illumination: your subconscious taps you on the shoulder and the proverbial light bulb goes off in your mind.

This kind of creativity is like walking a tightrope. On the one hand, you won't solve a difficult problem unless you make yourself familiar with the area to which it seems to belong – along with many other areas, which may or may not be related, just in case they are. On the other hand, if all you do is get trapped into standard ways of thinking, which others have already tried, fruitlessly, then you will be stuck in a mental rut and discover nothing new. So the trick is to know a lot, integrate it consciously, put your brain in gear for weeks … and then set the question aside. The intuitive part of your mind then goes to work, rubs ideas against each other to see whether the sparks fly, and notifies you when it has found something. This can happen at any moment: Poincaré suddenly saw how to solve a problem that had been bugging him for months when he was stepping off a bus. Srinivasa Ramanujan, a self-taught Indian mathematician with a talent for remarkable formulas, often got his ideas in dreams. Archimedes famously worked out how to test metal to see if it were gold when he was having a bath.

Poincaré took pains to point out that without the initial period of preparation, progress is unlikely. The subconscious, he insisted, needs to be given plenty to think about, otherwise the fortuitous combinations of ideas that will eventually lead to a solution cannot form. Perspiration begets inspiration. He must also have known – because any creative mathematician does – that this simple three-stage process seldom occurs just once. Solving a problem often requires more than one breakthrough. The incubation stage for one idea may be interrupted by a subsidiary process of preparation, incubation, and illumination for something that is needed to make the first idea work. The solution to any problem worth its salt, be it great or not, typically involves many such sequences, nested inside each other like one of Benoît Mandelbrot's intricate fractals. You

solve a problem by breaking it down into subproblems. You convince yourself that if you can solve these subproblems, then you can assemble the results to solve the whole thing. Then you work on the subproblems. Sometimes you solve one; sometimes you fail, and a rethink is in order. Sometimes a subproblem itself breaks up into more pieces. It can be quite a task just to keep track of the plan.

I described the workings of the subconscious as ‘intuition’. This is one of those seductive words like ‘instinct’, which is widely used even though it is devoid of any real meaning. It’s a name for something whose presence we recognise, but which we do not understand. Mathematical intuition is the mind’s ability to sense form and structure, to detect patterns that we cannot consciously perceive. Intuition lacks the crystal clarity of conscious logic, but it makes up for that by drawing attention to things we would never have consciously considered. Neuroscientists are barely starting to understand how the brain carries out much simpler tasks. But however intuition works, it must be a consequence of the structure of the brain and how it interacts with the external world.

Often the key contribution of intuition is to make us aware of weak points in a problem, places where it may be vulnerable to attack. A mathematical proof is like a battle, or if you prefer a less warlike metaphor, a game of chess. Once a potential weak point has been identified, the mathematician’s technical grasp of the machinery of mathematics can be brought to bear to exploit it. Like Archimedes, who wanted a firm place to stand so that he could move the Earth, the research mathematician needs some way to exert leverage on the problem. One key idea can open it up, making it vulnerable to standard methods. After that, it’s just a matter of technique.

My favourite example of this kind of leverage is a puzzle that has no intrinsic mathematical significance, but drives home an important message. Suppose you have a chessboard, with 64 squares, and a supply of dominoes just the right size to cover two adjacent squares of the board. Then it’s easy to cover the entire board with 32 dominoes. But now suppose that two

diagonally opposite corners of the board have been removed, as in Figure 1. Can the remaining 62 squares be covered using 31 dominoes? If you experiment, nothing seems to work. On the other hand, it's hard to see any obvious reason for the task to be impossible. Until you realise that however the dominoes are arranged, each of them must cover one black square and one white square. This is your lever; all you have to do now is to wield it. It implies that any region covered by dominoes contains the same number of black squares as it does white squares. But diagonally opposite squares have the same colour, so removing two of them (here white ones) leads to a shape with two more black squares than white. So no such shape can be covered. The observation about the combination of colours that *any* domino covers is the weak point in the puzzle. It gives you a place to plant your logical lever, and push. If you were a medieval baron assaulting a castle, this would be the weak point in the wall – the place where you should concentrate the firepower of your trebuchets, or dig a tunnel to undermine it.

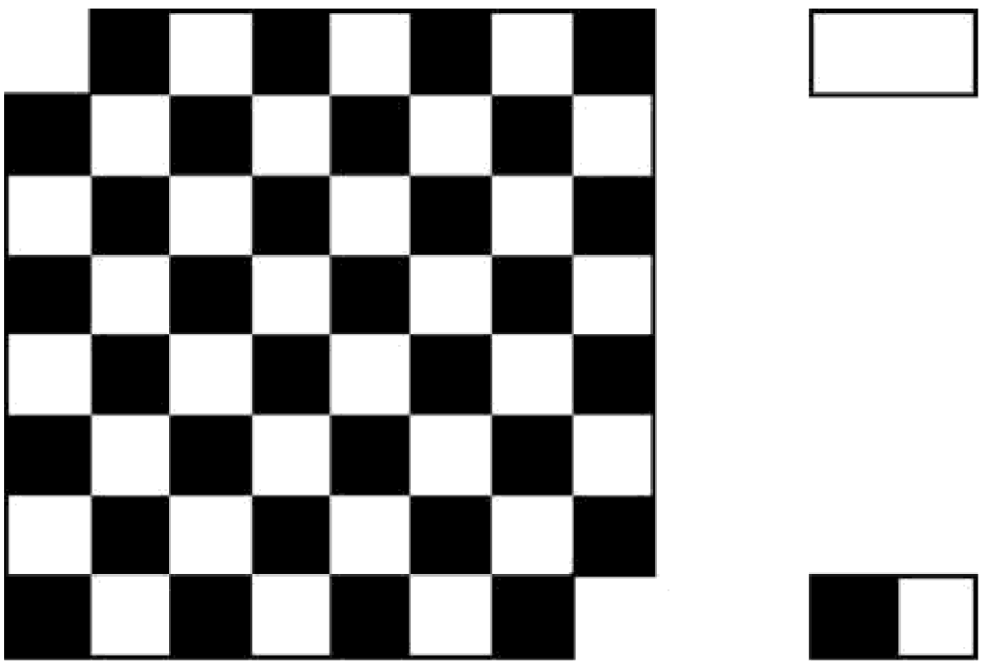


Fig 1 Can you cover the hacked chessboard with dominoes, each

covering two squares (top right)? If you colour the domino (bottom right) and count how many black and white squares there are, the answer is clear.

Mathematical research differs from a battle in one important way. Any territory you once occupy remains yours for ever. You may decide to concentrate your efforts somewhere else, but once a theorem is proved, it doesn't disappear again. This is how mathematicians make progress on a problem, even when they fail to solve it. They establish a new fact, which is then available for anyone else to use, in any context whatsoever. Often the launchpad for a fresh assault on an age-old problem emerges from a previously unnoticed jewel half-buried in a shapeless heap of assorted facts. And that's one reason why new mathematics can be important for its own sake, even if its uses are not immediately apparent. It is one more piece of territory occupied, one more weapon in the armoury. Its time may yet come – but it certainly won't if it is deemed 'useless' and forgotten, or never allowed to come into existence because no one can see what it is *for*.

2
Prime territory
Goldbach Conjecture

SOME OF THE GREAT PROBLEMS show up very early in our mathematical education, although we may not notice. Soon after we are taught multiplication, we run into the concept of a prime number. Some numbers can be obtained by multiplying two smaller numbers together; for example, $6 = 2 \times 3$. Others, such as 5, cannot be broken up in this manner; the best we can do is $5 = 1 \times 5$, which doesn't involve two *smaller* numbers. Numbers that can be broken up are said to be composite; those that can't are prime. Prime numbers seem such simple things. As soon as you can multiply whole numbers together you can understand what a prime number is. Primes are the basic building blocks for whole numbers, and they turn up all over mathematics. They are also deeply mysterious, and they seem to be scattered almost at random. There's no doubting it: primes are an enigma. Perhaps this is a consequence of their definition – not so much what they are as what they are not. On the other hand, they are fundamental to mathematics, so we can't just throw up our hands in horror and give up. We need to come to terms with primes, and ferret out their innermost secrets.

A few features are obvious. With the exception of the smallest prime, 2, all primes are odd. With the exception of 3, the sum of their digits can't be a multiple of 3. With the exception of 5, they can't end in the digit 5. Aside from these rules, and a few subtler ones, you can't look at a number and immediately spot whether it is prime. There do exist formulas for primes, but to a great extent they are cheats: they don't provide useful new information about primes; they are just clever ways to encode the definition of 'prime' in a formula. Primes are like people: they are individuals, and they don't conform to standard rules.

Over the millennia, mathematicians have gradually increased their understanding of prime numbers, and every so often another big problem about them is solved. However, many questions still remain unanswered. Some are basic and easy to state; others are more esoteric. This chapter discusses what we do and don't know about these infuriating, yet fundamental, numbers. It begins by setting up

some of the basic concepts, in particular, prime factorisation – how to express a given number by multiplying primes together. Even this familiar process leads into deep waters as soon as we start asking for genuinely effective methods for finding a number's prime factors. One surprise is that it seems to be relatively easy to test a number to determine whether it is prime, but if it's composite, finding its prime factors is often much harder.

Having sorted out the basics, we move on to the most famous unsolved problem about primes, the 250-year-old Goldbach conjecture. Recent progress on this question has been dramatic, but not yet decisive. A few other problems provide a brief sample of what is still to be discovered about this rich but unruly area of mathematics.

Prime numbers and factorisation are familiar from school arithmetic, but most of the interesting features of primes are seldom taught at that level, and virtually nothing is proved. There are sound reasons for that: the proofs, even of apparently obvious properties, are surprisingly hard. Instead, pupils are taught some simple methods for working with primes, and the emphasis is on calculations with relatively small numbers. As a result, our early experience of primes is a bit misleading.

The ancient Greeks knew some of the basic properties of primes, and they knew how to prove them. Primes and factors are the main topic of Book VII of Euclid's *Elements*, the great geometry classic. This particular book contains a geometric presentation of division and multiplication in arithmetic. The Greeks preferred to work with lengths of lines, rather than numbers as such, but it is easy to reformulate their results in the language of numbers. Euclid takes care to prove statements that may seem obvious: for example, Proposition 16 of Book VII proves that when two numbers are multiplied together, the result is independent of the order in which they are taken. That is, ab, = ba, a basic law of algebra.

In school arithmetic, prime factors are used to find the greatest common divisor (or highest common factor) of two numbers. For instance, to find the greatest common divisor of 135 and 630, we factorise them into primes:

$135 = 3^3 \times 5 \quad 630 = 2 \times 3^2 \times 5 \times 7$

Then, for each prime, we take the largest power that occurs in both factorisations, obtaining $3^2 \times 5$. Multiply out to get 45: this is the

greatest common divisor. This procedure gives the impression that prime factorisation is needed to find greatest common divisors. Actually, the logical relationship goes the other way. Book VII Proposition 2 of the *Elements* presents a method for finding the greatest common divisor of two whole numbers without factorising them. It works by repeatedly subtracting the smaller number from the larger one, then applying a similar process to the resulting remainder and the smaller number, and continuing until there is no remainder. For 135 and 630, a typical example using smallish numbers, the process goes like this. Subtract 135 repeatedly from 630:

$630-135=495$

$495-135=360$

$360-135=225$

$225-135=90$

Since 90 is smaller than 135, switch to the two numbers 90 and 135:

$135-90=45$

Since 45 is smaller than 90, switch to 45 and 90:

$90-45=45$

$45-45=0$

Therefore the greatest common divisor of 135 and 630 is 45.

This procedure works because at each stage it replaces the original pair of numbers by a simpler pair (one of the numbers is smaller) that has the same greatest common divisor. Eventually one of the numbers divides the other exactly, and at that stage we stop. Today's term for an explicit computational method that is guaranteed to find an answer to a given problem is 'algorithm'. So Euclid's procedure is now called the Euclidean algorithm. It is logically prior to prime factorisation. Indeed, Euclid uses his algorithm to prove basic properties about prime factors, and so do university courses in mathematics today.

Euclid's Proposition 30 is vital to the whole enterprise. In modern terms, it states that if a prime divides the product of two numbers – what you get by multiplying them together – then it must divide one of them. Proposition 32 states that either a number is prime or it has a prime factor. Putting the two together, it is easy to deduce that every number is a product of prime factors, and that this expression is unique apart from the order in which the factors are written. For example,

$$60 = 2 \times 2 \times 3 \times 5 = 2 \times 3 \times 2 \times 5 = 5 \times 3 \times 2 \times 2$$

and so on, but the only way to get 60 is to rearrange the first factorisation. There is no factorisation, for example, looking like 60 = 7 × *something*. The existence of the factorisation comes from Proposition 32. If the number is prime, stop. If not, find a prime factor, divide to get a smaller number, and repeat. Uniqueness comes from Proposition 30. For example, if there were a factorisation 60 = 7 × *something*, then 7 must divide one of the numbers 2, 3, or 5, but it doesn't.

At this point I need to clear up a small but important point: the exceptional status of the number 1. According to the definition as stated so far, 1 is clearly prime: if we try to break it up, the best we can do is 1 = 1 × 1, which does not involve smaller numbers. However, this interpretation causes problems later in the theory, so for the last century or two, mathematicians have added an extra restriction. The number 1 is so special that it should be considered as neither prime nor composite. Instead, it is a third manner of beast, a unit. One reason for treating 1 as a special case, rather than a genuine prime, is that if we call 1 a prime then uniqueness fails. In fact, 1 × 1 = 1 already exhibits the failure, and 1 × 1 × 1 × 1 × 1 × 1 × 1 × 1 = 1 rubs our noses in it. We could modify uniqueness to say 'unique except for extra 1s', but that's just another way to admit that 1 is special.

Much later, in Proposition 20 of Book IX, Euclid proves another key fact: 'Prime numbers are more than any assigned multitude of prime numbers.' That is, the number of primes is infinite. It's a wonderful theorem with a clever proof, but it opened up a huge can of worms. If the primes go on for ever, yet seem to have no pattern, how can we describe what they look like?

We have to face up to that question because we can't ignore the primes. They are essential features of the mathematical landscape. They are especially common, and useful, in number theory. This area of mathematics studies properties of whole numbers. That may sound a bit elementary, but actually number theory is one of the deepest and most difficult areas of mathematics. We will see plenty of evidence for that statement later. In 1801 Gauss, the leading number theorist of his age – arguably one of the leading mathematicians of all time, perhaps even the greatest of them all – wrote an advanced textbook of number theory, the *Disquisitiones Arithmeticae* ('Investigations in arithmetic'). In among the high-level

topics, he pointed out that we should not lose sight of two very basic issues: ‘The problem of distinguishing prime numbers from composite numbers and of resolving the latter into their prime factors is known to be one of the most important and useful in arithmetic.’

At school, we are usually taught exactly one way to find the prime factors of a number: try all possible factors in turn until you find something that goes exactly. If you haven’t found a factor by the time you reach the square root of the original number – more precisely, the largest whole number that is less than or equal to that square root – then the number is prime. Otherwise you find a factor, divide out by that, and repeat. It’s more efficient to try just prime factors, which requires having a list of primes. You stop at the square root because the smallest factor of any composite number is no greater than its square root. However, this procedure is hopelessly inefficient when the numbers become large. For example, if the number is

1, 080, 813, 321, 843, 836, 712, 253

then its prime factorisation is

13, 929, 010, 429 × 77, 594, 408, 257

and you would have to try the first 624,401,249 primes in turn to find the smaller of the two factors. Of course, with a computer this is fairly easy, but if we start with a 100-digit number that happens to be the product of two 50-digit numbers, and employ a systematic search through successive primes, the universe will end before the computer finds the answer.

In fact, today’s computers can generally factorise 100-digit numbers. My computer takes less than a second to find the prime factors of $10^{99} + 1$, which looks like 1000 … 001 with 98 zeros. It is a product of 13 primes (one of them occurs twice), of which the smallest is 7 and the largest is

141, 122, 524, 877, 886, 182, 282, 233, 539, 317, 796, 144, 938, 305, 111, 168, 717

But if I tell the computer to factorise $10^{199} + 1$, with 200 digits, it churns away for ages and gets nowhere. Even so, the 100-digit calculation is impressive. What’s the secret? Find more efficient methods than trying all potential prime factors in turn.

We now know a lot more than Gauss did about the first of his problems (testing for primes) and a lot less than we’d like to about

the second (factorisation). The conventional wisdom is that primality testing is far simpler than factorisation. This generally comes as a surprise to non-mathematicians, who were taught at school to test whether a number is prime by the same method used for factorisation: try all possible divisors. It turns out that there are slick ways to prove that a number is prime without doing that. They also allow us to prove that a number is composite, without finding any of its factors. Just show that it fails a primality test.

The great grand-daddy of all modern primality tests is Fermat's theorem, not to be confused with the celebrated Fermat's last theorem, chapter 7. This theorem is based on modular arithmetic, sometimes known as 'clock arithmetic' because the numbers wrap round like those on a clock face. Pick a number – for a 12-hour analogue clock it is 12 – and call it the modulus. In any arithmetical calculation with whole numbers, you now allow yourself to replace any multiple of 12 by zero. For example, $5 \times 5 = 25$, but 24 is twice 12, so subtracting 24 we obtain $5 \times 5 = 1$ to the modulus 12. Modular arithmetic is very pretty, because nearly all of the usual rules of arithmetic still work. The main difference is that you can't always divide one number by another, even when it's not zero. Modular arithmetic is also useful, because it provides a tidy way to deal with questions about divisibility: which numbers are divisible by the chosen modulus, and what is the remainder when they're not? Gauss introduced modular arithmetic in the *Disquisitiones Arithmeticae*, and today it is widely used in computer science, physics, and engineering, as well as mathematics.

Fermat's theorem states that if we choose a prime modulus p, and take any number a that is not a multiple of p, then the $(p - 1)$ th power of a is equal to 1 in arithmetic to the modulus p. Suppose, for example, that $p = 17$ and $a = 3$. Then the theorem predicts that when we divide 3^{16} by 17, the remainder is 1. As a check,

$$3^{16} = 43,046,721 = 2,532,160 \times 17 + 1$$

No one in their right mind would want to do the sums that way for, say, 100-digit primes. Fortunately, there is a clever, quick way to carry out this kind of calculation. The point is that if the answer is not equal to 1 then the modulus we started with is composite. So Fermat's theorem forms the basis of an efficient test that provides a necessary condition for a number to be prime.

Unfortunately, the test is not sufficient. Many composite numbers, known as Carmichael numbers, pass the test. The smallest is 561,

and in 2003 Red Alford, Andrew Granville, and Carl Pomerance proved, to general amazement, that there are infinitely many. The amazement was because they found a proof; the actual result was less of a surprise. In fact, they showed that there are at least $x^{2/7}$ Carmichael numbers less than or equal to x if x is large enough.

However, more sophisticated variants of Fermat's theorem can be turned into genuine tests for primality, such as one published in 1976 by Gary Miller. Unfortunately, the proof of the validity of Miller's test depends on an unsolved great problem, the generalised Riemann hypothesis, chapter 9. In 1980 Michael Rabin turned Miller's test into a probabilistic one, a test that might occasionally give the wrong answer. The exceptions, if they exist, are very rare, but they can't be ruled out altogether. The most efficient deterministic (that is, guaranteed correct) test to date is the Adleman-Pomerance-Rumely test, named for Leonard Adleman, Pomerance, and Robert Rumely. It uses ideas from number theory that are more sophisticated than Fermat's theorem, but in a similar spirit.

I still vividly recall a letter from one hopeful amateur, who proposed a variant of trial division. Try all possible divisors, but start at the square root and work *downwards*. This method sometimes gets the answer more quickly than doing things in the usual order, but as the numbers get bigger it runs into the same kind of trouble as the usual method. If you try it on my example above, the 22-digit number 1,080,813,321,843,836,712,253, then the square root is about 32,875,725,419. You have to try 794,582,971 prime divisors before you find one that works. This is *worse* than searching in the usual direction.

In 1956 The famous logician Kurt Gödel, writing to John von Neumann, echoed Gauss's plea. He asked whether trial division could be improved, and if so, by how much. Von Neumann didn't pursue the question, but over the years others answered Gödel by discovering practical methods for finding primes with up to 100 digits, sometimes more. These methods, of which the best known is called the quadratic sieve, have been known since about 1980. However, nearly all of them are either probabilistic, or they are inefficient in the following sense.

How does the running time of a computer algorithm grow as the input size increases? For primality testing, the input size is not the number concerned, but how many digits it has. The core distinction

cryptanalysis, the dark art of code-breaking. Many novel codes have been devised, and one of the most famous, invented by Ron Rivest, Adi Shamir, and Leonard Adleman in 1978, uses prime numbers. Big ones, about a hundred digits long. The Rivest-Shamir-Adleman system is employed in many computer operating systems, is built into the main protocols for secure Internet communication, and is widely used by governments, corporations, and universities. That doesn't mean that every new result about primes is significant for the security of your Internet bank account, but it adds a definite frisson of excitement to any discovery that relates primes to computation. The Agrawal-Kayal-Saxena test is a case in point. Mathematically, it is elegant and important, but it has no direct practical significance.

It does, however, cast the general issue of Rivest-Shamir-Adleman cryptography in a new and slightly disturbing light. There is still no class P algorithm to solve Gauss's second problem, factorisation. Most experts think nothing of the kind exists, but they're not quite as sure as they used to be. Since new discoveries like the Agrawal-Kayal-Saxena test can lurk unsuspected in the wings, based on such simple ideas as polynomial versions of Fermat's theorem, cryptosystems based on prime factorisation might not be quite as secure as we fondly imagine. Don't reveal your cat's name on the Internet just yet.

Even the basic mathematics of primes quickly leads to more advanced concepts. The mystery becomes even deeper when we ask subtler questions. Euclid proved that the primes go on for ever, so we can't just list them all and stop. Neither can we give a simple, useful algebraic formula for successive primes, in the way that x^2 specifies squares. (There do exist simple formulas, but they 'cheat' by building the primes into the formula in disguise, and don't tell us anything new.[11]) To grasp the nature of these elusive, erratic numbers, we can carry out experiments, look for hints of structure, and try to prove that these apparent patterns persist no matter how large the primes become. For instance, we can ask how the primes are distributed among all whole numbers. Tables of primes strongly suggest that they tend to thin out as they get bigger. Table 1 shows how many primes there are in various ranges of 1000 consecutive numbers.

The numbers in the second column mostly decrease as we move down the rows, though sometimes there are brief periods when they

go the other way: 114 is followed by 117, for instance. This is a symptom of the irregularity of the primes, but despite that, there is a clear general tendency for primes to become rarer as their size increases. The reason is not far to seek: the bigger a number becomes, the more potential factors there are. Primes have to avoid all of these factors. It's like fishing for non-primes with a net: the finer the net becomes, the fewer primes slip through.

range	number of primes
1–1000	168
1001–2000	135
2001–3000	127
3001–4000	119
4001–5000	118
5001–6000	114
6001–7000	117
7001–8000	106
8001–9000	110
9001–10,000	111

Table 1 The number of primes in successive intervals of 1000 numbers.

The 'net' even has a name: the sieve of Eratosthenes. Eratosthenes of Cyrene was an ancient Greek mathematician who lived around 250 BC. He was also an athlete with interests in poetry, geography, astronomy, and music. He made the first reasonable estimate of the size of the Earth by observing the position of the Sun at noon in two different locations, Alexandria and Syene – present-day Aswan. At noon, the Sun was directly overhead at Syene, but about 7 degrees from the vertical at Alexandria. Since this angle is one fiftieth of a circle, the Earth's circumference must be 50 times the distance from Alexandria to Syene. Eratosthenes couldn't measure that distance directly, so he asked traders how long it took to make the journey by camel, and estimated how far a camel typically went in a day. He gave an explicit figure in a unit known as a *stadium*, but we don't know how long that unit was. Historians

generally think that Eratosthenes's estimate was reasonably accurate.

His sieve is an algorithm to find all primes by successively eliminating all multiples of numbers already known to be prime. Figure 2 illustrates the method on the numbers up to 102, arranged to make the elimination process easy to follow. To see what's going on, I suggest you construct the diagram for yourself. Start with just the grid, omitting the lines that cross numbers out. Then you can add those lines one by one. Omit 1 because it's a unit. The next number is 2, so that's prime. Cross out all multiples of 2: these lie on the horizontal lines starting from 4, 6, and 8. The next number not crossed out is 3, so that's prime. Cross out all multiples of 3: these lie on the horizontal lines starting from 6, already crossed out, and 9. The next number not crossed out is 5, so that's prime. Cross out all multiples of 5: these lie on the diagonal lines sloping up and to the right, starting at 10. The next number not crossed out is 7, so that's prime. Cross out all multiples of 7: these lie on the diagonal lines sloping down and to the right, starting at 14. The next number not crossed out is 11, so that's prime. The first multiple of 11 that has not already been crossed out because it has a smaller divisor is 121, which is outside the picture, so stop. The remaining numbers, shaded, are the primes.

1	7	13	19	25	31	37	43	49	55	61	67	73	79	85	91	97
2	8	14	20	26	32	38	44	50	56	62	68	74	80	86	92	98
3	9	15	21	27	33	39	45	51	57	63	69	75	81	87	93	99
4	10	16	22	28	34	40	46	52	58	64	70	76	82	88	94	100
5	11	17	23	29	35	41	47	53	59	65	71	77	83	89	95	101
6	12	18	24	30	36	42	48	54	60	66	72	78	84	90	96	102

Fig 2 The sieve of Eratosthenes.

The sieve of Eratosthenes is not just a historical curiosity; it is still one of the most efficient methods known for making extensive lists of primes. And related methods have led to significant progress on what is probably the most famous unsolved great problem about primes: the Goldbach conjecture. The German amateur

mathematician Christian Goldbach corresponded with many of the famous figures of his time. In 1742 he stated a number of curious conjectures about primes in a letter to Leonhard Euler. Historians later noticed that René Descartes had said much the same a few years before. The first of Goldbach's statements was: 'Every integer which can be written as the sum of two primes, can also be written as the sum of as many primes as one wishes, until all terms are units.' The second, added in the margin of his letter, was: 'Every integer greater than 2 can be written as the sum of three primes.' With today's definition of 'prime' there are obvious exceptions to these statements. For example, 4 is not the sum of three primes, because the smallest prime is 2, so the sum of three primes must be at least 6. But in Goldbach's day, the number 1 was considered to be prime. It is straightforward to rephrase his conjectures using the modern convention.

In his reply, Euler recalled a previous conversation with Goldbach, when Goldbach had pointed out that his first conjecture followed from a simpler one, his third conjecture: 'Every even integer is the sum of two primes.' With the prevailing convention that 1 is prime, this statement also implies the second conjecture, because any number can be written as either

$n + 1$ or $n + 2$ where n is even. If n is the sum of two primes, the original number is the sum of three primes. Euler's opinion of the third conjecture was unequivocal: 'I regard this as a completely certain theorem, although I cannot prove it.' That pretty much sums up its status today.

The modern convention, in which 1 is not prime, splits Goldbach's conjectures into two different ones. The even Goldbach conjecture states:

Every even integer greater than 2 is the sum of two primes.

The odd Goldbach conjecture is:

Every odd integer greater than 5 is the sum of three primes.

The even conjecture implies the odd one, but not conversely.[12] It is useful to consider both conjectures separately because we still don't know whether either of them is true. The odd conjecture seems to be slightly easier than the even one, in the sense that more progress has been made.

Some quick calculations verify the even Goldbach conjecture for small numbers:

$$4 = 2 + 2$$
$$6 = 3 + 3$$
$$8 = 5 + 3$$
$$10 = 7 + 3 = 5 + 5$$
$$12 = 7 + 5$$
$$14 = 11 + 3 = 7 + 7$$
$$16 = 13 + 3 = 11 + 5$$
$$18 = 13 + 5 = 11 + 7$$
$$20 = 17 + 3 = 13 + 7$$

It is easy to continue by hand up to, say, 1000 or so – more if you 're persistent. For example 1000 = 3 + 997, and 1,000,000 = 17 + 999,993. In 1938 Nils Pipping verified the even Goldbach conjecture for all even numbers up to 100,000.

It also became apparent that as the number concerned gets bigger, there tend to be more and more ways to write it as a sum of primes. This makes sense. If you take a big even number, and keep subtracting primes in turn, how likely is it that *all* of the results will be composite? It takes just one prime to turn up among the resulting list of differences and the conjecture is verified for that number. Using statistical features of primes, we can assess the probability of such an outcome. The analysts Godfrey Harold Hardy and John Littlewood performed such a calculation in 1923, and derived a plausible but non-rigorous formula for the number of different ways to express a given even number n as a sum of two primes: approximately $n/[2(\log n)^2]$. This number increases as n becomes larger, and it also agrees with numerical evidence. But even if this calculation could be made rigorous, there might just be an occasional rare exception, so it doesn't greatly help.

The main obstacle to a proof of Goldbach's conjecture is that it combines two very different properties. Primes are defined in terms of multiplication, but the conjectures are about addition. So it is extraordinarily difficult to relate the desired conclusion to any reasonable features of primes. There seems to be nowhere to insert

several times in the relevant quarter of the table. Why? Because 20 sums have to fit into a set with only 13 members. So on average each boldface number appears about 1.5 times. (The actual number of sums is 27, so a better estimate shows that each boldface number appears twice.) If any even numbers are missing, the overlap must be bigger still.

We can play the same game with a larger upper limit – say 1 million. A formula called the prime number theorem, chapter 9, provides a simple estimate for the number of primes up to any given size x. The formula is $x/\log x$. Here, the estimate is about 72,380. (The exact figure is 78,497.) The corresponding shaded region occupies about one quarter of the table, so it provides about $n^2/4 = 250$ billion boldface numbers: sums of two primes in this range. This is vastly larger than the number of even numbers in the range, which is half a million. Now the amount of overlap has to be gigantic, with each sum occurring on average 500,000 times. So the chance of any particular even number escaping is greatly reduced.

With more effort, we can turn this approach into an estimate of the probability that some even number in a given range is not the sum of two primes, assuming that the primes are distributed at random and with frequencies given by the prime number theorem – that is, about $x/\log x$ primes less than any given x. This is what Hardy and Littlewood did. They knew that their approach wasn't rigorous, because primes are defined by a specific process and they're not actually random. Nevertheless, it's sensible to expect the actual results to be consistent with this probabilistic model, because the defining property of primes seems to have very little connection with what happens when we add two of them together.

Several standard methods in this area adopt a similar point of view, but taking extra care to make the argument rigorous. Sieve methods, which build on the sieve of Eratosthenes, are examples. General theorems about the density of numbers in sums of two sets – the proportion of numbers that occur, as the sets become very large – provide other useful tools.

When a mathematical conjecture eventually turns out to be correct, its history often follows a standard pattern. Over a period of time, various people prove the conjecture to be true provided special restrictions apply. Each such result improves on the previous one by relaxing some restrictions, but eventually this process runs out of steam. Finally, a new and much cleverer idea completes the proof.

For example, a conjecture in number theory may state that every positive integer can be represented in some manner using, say, six special numbers (prime, square, cube, whatever). Here the key features are *every* positive integer and *six* special numbers. Initial advances lead to much weaker results, but successive stages in the process slowly improve them.

The first step is often a proof along these lines: every positive integer that is not divisible by 3 or 11, except for some finite number of them, can be represented in terms of some gigantic number of special numbers – say 10^{666}. The theorem typically does not specify how many exceptions there are, so the result cannot be applied directly to any specific integer. The next step is to make the bound effective: that is, to prove that every integer greater than $10^{10^{42}}$ can be so represented. Then the restriction on divisibility by 3 is eliminated, followed by a similar advance for 11. After that, successive authors reduce one of the numbers 10^{666} or $10^{10^{42}}$, often both. A typical improvement might be that every integer greater than 5.8×10^{17} can be represented using at most 4298 special numbers, for instance.

Meanwhile, other researchers are working upwards from small numbers, often with computer assistance, proving that, say, every number less than or equal to 10^{12} can be represented using at most six special numbers. Within a year, 10^{12} has been improved in five stages, by different researchers or groups, to 11.0337×10^{29}. These improvements are neither routine nor easy, but the way they are achieved involves intricate special methods that provide no hint of a more general approach, and each successive contribution is more complicated and longer. After a few years of this kind of incremental improvement, applying the same general ideas but with more powerful computers and new tweaks, this number has risen to 10^{43}. But now the method grinds to a halt, and everyone agrees that however much tweaking is done, it will never lead to the full conjecture.

At that point the conjecture disappears from view, because no one is working on it any more. Sometimes, progress pretty much stops. Sometimes, twenty years pass with nothing new … and then, apparently from nowhere, Cheesberger and Fries announce that by reformulating the conjecture in terms of complex meta-ergodic quasiheaps and applying byzantine quisling theory, they have

obtained a complete proof. After several years arguing about fine points of logic, and plugging a few gaps, the mathematical community accepts that the proof is correct, and immediately asks if there's a better way to achieve the same result, or to push it further.

You will see this pattern work itself out many times in later chapters. Because such accounts become tedious, no matter how proud Buggins and Krumm are of their latest improvement of the exponent in the Jekyll-Hyde conjecture from 1.773 to 1.771 + ε for any positive ε, I will describe a few representative contributions and leave out the rest. This is not to deny the importance of the work of Buggins and Krumm. It may even have paved the way to the great Cheesberger-Fries breakthrough. But only experts, following the developing story, are likely to await the next tiny improvement with bated breath.

In future I'll provide less detail, but let's see how it goes for Goldbach.

Theorems that go some way towards establishing Goldbach's conjecture have been proved. The first big breakthrough came in 1923, when Hardy and Littlewood used their analytic techniques to prove the odd Goldbach conjecture for all sufficiently large odd numbers. However, their proof relied on another big conjecture, the generalised Riemann hypothesis, which we discuss in chapter 9. This problem is still open, so their approach had a significant gap. In 1930 Lev Schnirelmann bridged the gap using a fancy version of their reasoning, based on sieve methods. He proved that a nonzero proportion of all numbers can be represented as a sum of two primes. By combining this result with some generalities about adding sequences together, he proved that there is some number C such that every integer greater than 1 is a sum of at most C prime numbers. This number became known as Schnirelmann's constant. Ivan Matveyevich Vinogradov obtained similar results in 1937, but his method also did not specify how big 'significantly large' is. In 1939 K. Borozdin proved that it is no greater than $3^{14,\,348,\,907}$. By 2002 Liu Ming-Chit and Wang Tian-Ze had reduced this 'upper bound' to e^{3100}, which is about 2×10^{1346}. This is a lot smaller, but it is still too big for the intermediate numbers to be checked by computer.

In 1969 N.I. Klimov obtained the first specific estimate for Schnirelmann's constant: it is at most 6 billion. Other mathematicians reduced that number considerably, and by 1982

Hans Riesel and Robert Vaughan had brought it down to 19. Although 19 is a lot better than 6 billion, the evidence pointed to Schnirelmann's constant being a mere 3. In 1995 Leszek Kaniecki reduced the upper bound to 6, with five primes for any odd number, but he had to assume the truth of the Riemann hypothesis. His results, combined with J. Richstein's numerical verification of the Riemann hypothesis up to 4×10^{14}, would prove that Schnirelmann's constant is at most 4, again assuming the Riemann hypothesis. In 1997 Jean-Marc Deshouillers, Gove Effinger, Herman te Riele, and Dmitrii Zinoviev showed that the generalised Riemann hypothesis (chapter 9) implies the odd Goldbach conjecture. That is, every odd number except 1, 3, and 5 is the sum of three primes.

Since the Riemann hypothesis is currently not proved, it is worth trying to remove this assumption. In 1995 the French mathematician Olivier Ramaré reduced the upper estimate for representing odd numbers to 7, without using the Riemann hypothesis. In fact, he proved something stronger: every even number is a sum of at most six primes. (To deal with odd numbers, subtract 3: the result is even, so it is a sum of six or fewer primes. The original number is this sum plus the prime 3, requiring seven or fewer primes.) The main breakthrough was to improve existing estimates for the proportion of numbers, in some specified range, that are the sum of two primes. Ramaré's key result is that for any number n greater than e^{67} (about 1.25×10^{29}), at least one fifth of the numbers between n and $2n$ are the sum of two primes. Using sieve methods, in conjunction with a theorem of Hans-Heinrich Ostmann about sums of sequences, refined by Deshouillers, this leads to a proof that every even number greater than 10^{30} is a sum of at most six primes.

The remaining obstacle is to deal with the gap between 4×10^{14}, where Jörg Richstein had checked the theorem by computer, and 10^{30}. As is common, the numbers are too big for a direct computer search, so Ramaré proved a series of specialised theorems about the number of primes in small intervals. These theorems depend on the truth of the Riemann hypothesis up to specific limits, which can be verified by computer. So the proof consists mainly of conceptual pencil-and-paper deductions, with computer assistance in this particular respect. Ramaré ended his paper by pointing out that in principle a similar approach could reduce the number of primes from 7 to 5. However, there were huge practical obstacles, and he wrote that such a proof 'can not be reached by today's computers'.

In 2012 Terence Tao overcame those difficulties with some new and very different ideas. He posted a paper on the Internet, which as I write is under review for publication. Its main theorem is: every odd number is a sum of at most five primes. This reduces Schnirelmann's constant to 6. Tao is renowned for his ability to solve difficult problems in many areas of mathematics. His proof throws several powerful techniques at the problem, and requires computer assistance. If the number 5 in Tao's theorem could be reduced to 3, the odd Goldbach conjecture would be proved, and the bound on Schnirelmann's constant reduced to 4. Tao suspects that it should be possible to do this, although further new ideas will be needed.

The even Goldbach conjecture seems harder still. In 1998 Deshouillers, Saouter, and te Riele verified it for all even numbers up to 10^{14}. By 2007, Tomás Oliveira e Silva had improved that to 10^{18}, and his computations continue. We know that every even integer is the sum of at most six primes – proved by Ramaré in 1995. In 1973 Chen Jing-Run proved that every sufficiently large even integer is the sum of a prime and a semiprime (either a prime or a product of two primes). Close, but no cigar. Tao has stated that the even Goldbach conjecture is beyond the reach of his methods. Adding three primes together creates far more overlap in the resulting numbers – in the sense discussed in connection with Figure 3 – than the two primes needed for the even Goldbach conjecture, and Tao's and Ramaré's methods exploit this feature repeatedly.

In a few years' time, then, we may have a complete proof of the odd Goldbach conjecture, in particular implying that every even number is the sum of at most four primes. But the even Goldbach conjecture will probably still be just as baffling as it was for Euler and Goldbach.

In the 2300 years since Euclid proved several basic theorems about primes, we have learned a great deal more about these elusive, yet vitally important, numbers. But what we now know puts into stark perspective the long list of what we don't know.

We know, for instance, that there are infinitely many primes of the form $4k + 1$ and $4k + 3$; more generally, that any arithmetic sequence[13] $ak + b$ for fixed a and b contains infinitely many primes provided a and b have no common factor. For instance, suppose that $a = 18$. Then $b = 1, 5, 7, 11, 13$, or 17. Therefore there exist

3
The puzzle of pi Squaring the Circle

PRIMES ARE AN OLD IDEA, but circles are even older. Circles led to a great problem that took more than 2000 years to solve. It is one of several related geometric problems that have come down to us from antiquity. The central character in the story is the number π (Greek 'pi') which we meet at school in connection with circles and spheres. Numerically it is 3.14159 and a bit; often the approximation 22/7 is used. The digits of π never stop, and they never repeat the same sequence over and over again. The current record for calculating digits of π is 10 trillion digits, by Alexander Yee and Chigeru Kondo in October 2011.[14] Computations like this are significant as ways to test fast computers, or to inspire and test new methods to calculate π, but very little hinges on the numerical results. The reason for being interested in π is not to calculate the circumference of a circle. The same strange number appears all over mathematics, not just in formulas related to circles and spheres, and it leads into very deep waters indeed. The school formulas are important, even so, and they reflect π's origins in Greek geometry.

There, one of the great problems was the unsolved task of squaring the circle. This phrase is often employed colloquially to indicate a wrong-headed approach to something, rather like trying to fit a square peg into a round hole. Like many common phrases extracted from science, this one's meaning has changed over the centuries.[15] In Greek times, trying to square the circle was a perfectly reasonable idea. The difference in the two shapes – straight or curved – is totally irrelevant: similar problems have valid solutions.[16] However, it eventually turned out that this particular problem cannot be solved using the specified methods. The proof is ingenious and technical, but its general nature is comprehensible.

In mathematics, squaring the circle means constructing a square whose *area* is the same as that of a given circle, using the traditional methods of Euclid. Greek geometry actually permitted other methods, so one aspect of the problem is to pin down which methods are to be used. The impossibility of solving the problem is then a statement about the limitations of those methods; it doesn't imply that we can't work out the area of a circle. We just have to find

another approach. The impossibility proof explains why the Greek geometers and their successors failed to find a construction of the required kind: there isn't one. In retrospect, that explains why they had to introduce more esoteric methods. So the solution, despite being negative, clears up what would otherwise be a big historical puzzle. It also stops people wasting time in a continuing search for a construction that doesn't exist – except for a few hardy souls who regrettably seem unable to get the message, no matter how carefully it is explained.[17]

In Euclid's *Elements* the traditional methods for constructing geometric figures are idealised versions of two mathematical instruments: the ruler and the compass. To be pedantic, compass *es*, for the same reason that you cut paper with scissor *s*, not with a scissor – but I will follow common parlance and avoid the plural. These instruments are used to 'draw' diagrams on a notional sheet of paper, the Euclidean plane.

Their form determines what they can draw. A compass comprises two rigid rods, hinged together. One has a sharp point, the other holds a sharp pencil. The instrument is used to draw a circle, or part of one, with a specific centre and a specific radius. A ruler is simpler: it has a straight edge, and is used to draw a straight line. Unlike the rulers you buy in stationery shops, Euclid's rulers have no marks on them, and this is an important restriction for the mathematical analysis of what they can create.

The sense in which the geometer's ruler and compass are idealisations is straightforward: they are assumed to draw infinitely thin lines. Moreover, the straight lines are exactly straight and the circles are perfectly round. The paper is perfectly flat and even. The other key ingredient of Euclid's geometry is the notion of a point, another ideal. A point is a dot on the paper, but it is a physical impossibility: it has no size. 'A point', said Euclid, in the first sentence of the *Elements*, 'is that which has no part.' This sounds a bit like an atom, or if you're clued into modern physics, a subatomic particle, but compared to a geometric point, those are gigantic. From an everyday human perspective, however, Euclid's ideal point, an atom, and a pencil dot on a sheet of paper, are similar enough for the purposes of geometry.

These ideals are not attainable in the real world, however carefully you make the instruments and sharpen the pencil, and however smooth you make the paper. But idealism can be a virtue, because

these requirements make the mathematics much simpler. For instance, two pencil lines cross in a small fuzzy region shaped like a parallelogram, but mathematical lines meet at a single point. Insights gained from ideal circles and lines can often be transferred to real, imperfect ones. This is how mathematics works its magic.

Two points determine a (straight) line, the unique line that passes through them. To construct the line, place your ideal ruler so that it passes through the two points, and run your ideal pencil along it. Two points also determine a circle: choose one as the centre, and place the compass point there; then adjust it so that the tip of the pencil lies on the other point. Now swing the pencil round in an arc, keeping the central point fixed. Two lines determine a unique point, where they cross, unless they are parallel, in which case they don't cross, but a Pandora's box of logical issues yawns wide. A line and a circle determine two points, if they cross; one point, if the line cuts the circle at a tangent; nothing at all if the circle is too small to meet the line. Similarly two circles either meet in two points, one, or none.

Distance is a fundamental concept in the modern treatment of Euclidean geometry. The distance between any two points is measured along the line that joins them. Euclid managed to get his geometry working without an explicit concept of distance, by finding a way to say that two line segments have the *same* length without defining length itself. In fact, this is easy: just stretch a compass between the ends of one segment, transfer it to the second, and see if the ends fit. If they do, the lengths are equal; if they don't, they're not. At no stage do you measure an actual length.

From these basic ingredients, geometers can build up more interesting shapes and configurations. Three points determine a triangle unless they all lie on the same line. When two lines cross, they form an angle. A right angle is especially significant; a straight line corresponds to two right angles joined together. And so on, and so on, and so on. Euclid's *Elements* consists of 13 books, delving ever deeper into the consequences of these simple beginnings.

The bulk of the *Elements* consists of theorems – valid features of geometry. But Euclid also explains how to solve geometric problems, using constructions' based on ruler and compass. Given two points joined by a segment of a line, construct their midpoint. Or trisect the segment: construct a point exactly one third of the way along it. Given an angle, construct one that bisects it – is half the size. But some simple constructions proved elusive. Given an angle, construct one that trisects it – is one third the size. You can do that

for line segments, but no one could find a method for angles. Approximations, as close as you wish, yes. Exact constructions using only an unmarked ruler and a compass: no. However, no one really needs to trisect angles exactly anyway, so this particular issue didn't cause much trouble.

More embarrassing was a construction that could not be ignored: given a circle, construct a square that has the same area. This is the problem of squaring the circle. From the Greek point of view, if you couldn't solve that, you weren't entitled to claim that a circle *had* an area. Even though it visibly encloses a well-defined space, and intuitively the area is *how much* space. Euclid and his successors, notably Archimedes, settled for a pragmatic solution: assume circles have areas, but don't expect to be able to construct squares with the same area. You can still say a lot; for instance, you can prove, in full logical rigour, that the area of a circle is proportional to the square of its diameter. What you can't do, without squaring the circle, is to construct a line whose length is the constant of proportionality.

The Greeks couldn't square the circle using ruler and compass, so they settled for other methods. One used a curve called a quadratrix.[18] The importance they attached to using only ruler and compass was exaggerated by some later commentators, and it's not even clear that the Greeks considered squaring the circle to be a vital issue. By the nineteenth century, however, the problem was becoming a major nuisance. Mathematics that was unable to answer such a straightforward question was like a cordon bleu cook who didn't know how to boil an egg.

Squaring the circle sounds like a problem in geometry. That's because it is a problem in geometry. But its solution turned out to lie not in geometry at all, but in algebra. Making unexpected connections between apparently unrelated areas of mathematics often lies at the heart of solving a great problem. Here, the connection was not entirely unprecedented, but its link to squaring the circle was not at first appreciated. Even when it was, there was a technical difficulty, and dealing with that required yet another area of mathematics: analysis, the rigorous version of calculus. Ironically, the first breakthrough came from a fourth area: number theory. And it solved a geometric problem that the Greeks would never in their wildest dreams have believed to possess a solution, and as far as we can tell never thought about: how to construct, with ruler and compass, a regular polygon with 17 sides.