Calculus sin frontera

by Kathy Davis

Some things just pull so strong
Like the map of the sky is the map of your heart
- Ferron
This book was typeset using the \texttt{TUFTELATEX} package.

This book is copyrighted. You may use and distribute it freely, but it cannot be sold, incorporated in other texts, or modified without written permission of the author.

Unless required by applicable law or agreed to in writing, this book is distributed on an “\textit{as is}” \textit{basis}, \textit{without warranties or conditions of any kind}, either express or implied.

\textit{First Draft: Chapter 1, Sections 1 to 7, February 2019}
Table of Contents

Preface: Leaving the Sixties p5

Chapter 0: Why Calculus? p6
   Notes for Chapter 0 p8

Chapter 1: Numbers
   Section 1: Background & History p10
      Notes for Section 1 p17
   Section 2: Building the Numbers p25
      Notes for Section 2 p30
   Section 3: The Problem of Fractions p39
      Notes for Section 3 p44
   Section 4: The Problem of Irrationals p45
      Notes for Section 4 p51
   Section 5: The Real Numbers p54
      Notes for Section 5 p59

Section 6 Part 1 Decimals: Accuracy p62
   Section 6 Part 2 Decimals: Geometry and Decimals p67
      Notes for Section 6 p70

Section 7 Part 1: Units & Standards p72
   Notes for Section 7 Part 1 p76

Section 7 Part 2: Order of Magnitude p78
   Notes for Section 7 Part 2 p83

Section 8: Computer Numbers p86
   Notes for Section 8 p??

List of Figures p97

Index p101
Preface: Leaving the Sixties

I left for college in 1967: the sixties, when everything was changing. My dorm was co-ed, the first in the country. Vietnam war protests shut down the University and students smoked weed in the lounge. But also: an entering physics major, I picked up Feynman’s Lectures On Physics. Feynman wrote,

The special problem we tried to get at with these lectures was to maintain the interest of the very enthusiastic and rather smart students coming out of the high schools . . . . They have heard a lot about how interesting and exciting physics is – the theory of relativity, quantum mechanics, and other modern ideas. [Instead] they were made to study inclined planes, electrostatics, and so forth, and after two years it was quite stultifying.

Today, most students take an AP Calculus course in high school; they repeat the course in college, sometimes with the same book. They’ve heard about amazing discoveries in science, engineering, medicine and technology that change the world, but when they enter a calculus classroom, they’re magically transported back to 1967, as though the last fifty years of progress never happened.

This is not a calculus book; it’s a compendium of what gets left out of traditional books. Calculus touches on so much of the world: history, religion, philosophy, literature, psychology and especially the modern scientific and technical world.

Calculus is an amazing intellectual adventure. Let’s start.
Introduction: Why Calculus?

On Jody Foster’s first day at Yale she wrote a friend: ‘My calculus book is three inches thick. I can’t survive three inches of calculus.’ My students are more direct: ‘I’m going to be a doctor. Why do I need calculus?’ Or, ‘I’m gonna be an engineer: we use computers for everything. Why do I need this class?’ Instead of a direct answer, I’ll tell a story about a doctor, and an engineer, and the ECG, a graph of the electrical currents in a heartbeat, used to evaluate the health of a heart (for a quick summary, see p8).

The story begins in the 1790’s, when the Lucia Galvani, the wife of the Italian doctor Luigi, noticed that the muscles of some frog legs she’d hung out for dinner twitched when a spark of electricity touched them. Luigi followed up, experimenting for many years, concluding that the nervous system operated through electricity. But in 1790 there were no instruments to measure electricity; scientists detected electrical impulses in the body using frog legs. In the 1830’s, the neurophysiologist Carlo Matteucci laid a frog leg on a beating heart, and the leg twitched in rhythm with the heartbeats, showing that heart muscles also generate an electric current when they contract.

Opening the chest of a patient and laying a frog leg across the beating heart isn’t an easy or safe diagnostic technique, but the electrical impulses from the heart would be too weak to measure otherwise. It took another seventy years before the Dutch physiologist Willem Einthoven developed a device to measure those weak currents, and then did the extra engineering to record them.

Figure 2 shows an early ECG machine. The electric currents from the heart cause a very thin wire to move; on the right, a light beam shines on the wire, and on the left, a motorized strip of photograph paper records the movements. And . . . the world’s first human ECG recording, Figure 3.

Figure 2: Commercial ECG Machine
Bulkly and expensive, but Einthoven showed how the ECG can help diagnose diseases of the heart. He received the 1924 Nobel Prize in Medicine or Physiology for his work.

Figure 3: The First Human ECG
This ECG is blurred, but clearly shows the heartbeat has many different features. Einthoven was able to show how these features correspond to the contraction of muscles in different parts of the heart. He was able to show distortions in those features characterized some heart diseases.
Technology advances: Einthoven’s moving strip of photograph paper was replaced by a pen moving up and down, writing on a moving strip of paper. The paper has lines and squares to make reading the recordings accurate. And the graph is sharper, showing more detail (Figure 4).

Now the diagnosis: the patient is suffering from an erratic heart beat, and an MD can tell by measuring the distance between peaks. In Figure 4 each little box is .04 sec wide, so an old-fashioned MD would count the squares between the peaks, and compute how many beats per minute there were. In this (short) graph, an erratic heartbeat shows up as a varying number of beats per minute (for more details, see the note on p9).

That’s what a 1950’s MD would do; even in 2014, in a visit to a hospital, my ECG was on a piece of paper a lot like Figure 4. But it won’t help in an emergency room: there’s no time to count squares on a nice piece of paper. Fortunately, we’re not in the 1950’s: we have computers.

Technology advances: now it’s the engineer’s turn. To keep records of a patient’s heart condition, there’s no need to store dozens of sheets of paper in a file cabinet. Instead of paper, write the electrical signals directly into a computer.

Once the data is in a computer – well, instead of drawing the graph on paper and counting squares, a microchip draws the graph on a LCD screen . . . and it counts the squares, it computes the beats-per-minute.

Today you can buy a portable ECG unit (Figure 5). An MD can get a quick sense of whether a heart is behaving normally, and then, if needed, run a full ECG (which takes 10 minutes or so, requires a nurse, square counting, and $$. Technology is great.)

But before the chip can count squares, it has to somehow ‘see’ the peaks. Imagine an engineer programming the chip, working with an MD and neither took calculus. Calculus would have told them peaks happen at critical points, where a derivative is zero. Calculus tell us how to find critical points, but how does a computer do it? The software to find critical points, using a microchip, was written by Jiapu Pan in Shanghai and Willis Tompkins in Wisconsin, using a derivative-based algorithm. Figure 6 shows an example of the technique; it relies on successive mathematical operations to strip off the irrelevant parts of the beat.

This engineer and MD did need calculus, and their work is one of the most cited in biomedical engineering.
Notes

Notes for Chapter 0


An ECG records the electrical activity of the heart. Here’s a quick view of how.

The heart is a series of chambers to hold blood, muscles to pump blood, and veins and arteries to channel the blood. Muscle contractions are associated with electrical currents (we’ll go into more detail later in the book) and the current is measured by placing electrodes on the body; the record of the current flow is the ECG; see below.

On the left, the different chambers of the heart, and the major nerves that orchestrate the sequence of contractions of the chambers (for example, when one chamber contracts, the following chamber has to be open to receive the blood pumped; see fibrillation, below. Timing these events is crucial). On the right, a printout gives a visual record of how the different parts of the heart contract, the size of the contraction, and the timing between contractions; compare Figure 3. Well over a hundred years of study, matching symptoms with EKG’s allows an MD to diagnose heart problems using this information (you
can practice this yourself; see http://www.healio.com/cardiology/learn-the-heart/ecg-review/ecg-interpretation-tutorial/approach-to-ecg-interpretation).

The ECG serves as a kind of microscope into the heart, with the advantage that the doctor doesn’t have to do surgery to inspect the internal workings of the heart. It’s a simple, cheap way of monitoring basic health.

The erratic heartbeat in Figure 4 is called fibrillation. As we said, the heart is a series of chambers and pumps, and a working heart chamber pumps only after it gets blood from the preceding chamber. In fibrillation, the pumping happens at random, and the blood can slosh around the heart, never getting to the body. Depending upon the chamber that has the fibrillation, this can lead to death within minutes.
Chapter 1: Numbers

Section 1: Background & History

In the Introduction, we looked at an erratic heartbeat and wound up with some numbers and a graph. No-one even blinks, because numbers are already everywhere: time, speed, area, temperature, pressure, heart rate, GPA . . .

What are numbers, actually? And how did they spread all over our life?

Certainly they’re not things we can touch, like spoons. But – to take the simplest view – if ‘number’ is something in our minds, how did it get there?

This kind of question has troubled philosophers from all the way back, to now. For a bit of background, see p17. Although philosophy has been important in the development of modern science and mathematics, for now we’ll use modern science. Starting with a study of animal intelligence (see p19).

A scientist makes holes in a log and randomly places worms in those holes. A robin watches, and when the scientist leaves, the robin immediately flies to the log to munch on yummy worms (Figure 7). What’s surprising is the robin starts with the holes containing the most worms.

The study suggests that these birds can "count" at least to the extent of perceiving ‘more’ and ‘less’ and understanding how to exploit the perception. Other studies show this ability is cross-species: robins, rats, newly hatched chicks, new-born babies. They won’t pass an algebra test, but something in these tiny brains ‘gets’ number (this view is actually too simple, philosophically and scientifically; see p19 for a detailed discussion).

So: rats, robins and babies all count, but they don’t worry about being a size 6, affording $70K for a BMW, or counting change at the local Starbucks. Humans use numbers differently than animals do: how is this?

Part of the answer relates to information processing in our brain: we’re not very good at it. Studies show we can focus on only a very few items; we can keep only a few facts in immediate memory, and we can switch tasks only very slowly. Compared to our computers,
our hardware is slow, limited and obsolete (a good read on these limitations is David Eagleman’s *Incognito: The Secret Lives of the Brain*, Vintage 2012).

Number is the same: we can perceive the objects in Figure 8, but manipulating them in our mind (number) is hard. Some recent research suggests that the part of our brain which constructs number is limited and simply runs out of room (p21).

But: we can count. Counting is a form of external storage, whether using fingers or using number words or making marks on paper or sticks or bones (see Figure 9). Making marks on wood is in fact, cross-cultural, and persisted into the twentieth century in places where paper is rare.

In Figure 8, the mind retains the number eleven but loses the image of eleven green objects, their position, size and orientation. Retaining some but not all of the characteristics is a *symbolic representation* of our perception. Symbols add a new dimension: external counting makes numbers public, and even allows us to make stored records. The ability to take numbers (something in our brain) and translate them into physical, public form, is one difference between human and animal number use. Research suggest this connection connects two different areas of the brain, possibly pre-linguistic: see p21.

Fingers, words, and attention: we run out quickly. What humans do then is almost universal: anthropologists have found many societies in which people count higher than ten by referring to other parts of their bodies such as thighs, arms, etc (see p23).

Figure 10 shows a counting system used in markets in the Middle East. The joints of one hand mark one to twelve; the other hand counts how many groups of twelve we have. A written system needs markers for $1 \times 12$, for $2 \times 12 = 24$, up to $5 \times 12 = 60$.

Systems like this work well for personal transactions, but are no use for long distance trade – or for collecting taxes, or for ruling a country:

> Since the rules for collecting and manipulating numbers are widely shared, they can easily be transported across oceans and continents and used to co-ordinate activities or settle disputes. Perhaps most crucially, reliance on numbers and quantitative manipulation minimizes the need for intimate knowledge and personal trust. Quantification is well suited for communication that goes beyond the boundaries of locality and community.

Denise Schmandt-Besserat established the origin of written numbers in Mesopotamia and from this, the development of all writing. Figure 11 depicts clay objects used for record keeping, from Susa, Iran circa 3300 BCE (see p23). Schmandt-Besserat writes:

*The early city states still used tokens to control the levy of dues. When individuals could not pay, the tokens representing the amount of their debts were kept in a round clay envelope. In order to be able to verify the content of the envelope without breaking it, the tokens were impressed on the surface before enclosing them. A cone left a wedge-shaped mark and a disc a circular one. It was the invention of writing.*

With this innovation, number symbols become more than public: they were institutionalized in economic transactions and in government. A chain of innovations made the system efficient – to adapt it for use by large cities or states. Figure 12 shows an early Mesopotamian (modern Iran/Iraq) symbolic system, partially based on tens (see p23). Each round dot represents a ten; each vertical line a one. It’s as though we were to write 31 as 10,10,10,1,1. This notation has its advantages; we could just as easily write 1,10,1,10,1,10.

It’s harder to write something like 99 in this system. We could do like the Romans: introduce new symbols like V for five and L for fifty, then use what is called a subtractive system, writing IV to mean 5-1. But this gets complicated: you start to need lots of extra symbols (as in the Roman numerals I, V, X, L, C, M . . .).

Figure 13, from the Chinese Qin and Han dynasties, shows an improvement over the Mesopotamian notation. Scribes used vertical and horizontal lines for the numbers one through nine, they then arranged these using small boxes. The position of each box from the right determines 9 vs. 90 vs. 900. The change from Figure 12 to Figure 13 is the use of positional notation, which introduces standards as to which number goes where, making interpretation and use of the system more efficient. It also employs a further development, the use of empty boxes to denote zero.

This system, known as rod arithmetic, allowed the operations of addition, subtraction and multiplication to be performed by arranging numbers on a rod, aligning boxes, and performing the operations as we’d do today (Figure 14):

\[
\begin{array}{c}
5 & 4 \\
\_ & \\
2 & 3 \\
\_ & \\
3 & 1
\end{array}
\]

\[54 - 23 = 31\]
It’s a very fast method for doing arithmetic, and scholars in the Qin Dynasty needed it for massive engineering projects, like the Great Wall. So this is another factor in the spread of numbers into the many corners of life: the development of good notation, fast algorithms for performing computations.

Historically, though, much of the Chinese system remained in China. The writing at the top of Figure 16 shows how numbers were written in early Indian/Pakistani mathematics. The bottom shows the same numbers in Hindu-Arabic notation, the system used in modern Western science and technology.

The numbers are called ‘Hindu-Arabic’: by 662CE, use of the ‘Indian’ notational system had spread west to ancient Iran/Iraq. The ‘Arabic’ part comes from work at the Caliph’s court in Baghdad, a typical example is The Book of Addition and Subtraction according to the Indian Calculations, by the court mathematician Muhammad ibn Musa al-Khwarizmi (780 Ð 850), Figure 17. The Caliphate ruled a vast commercial and political empire: “Arabs ...had been using [Hindu numbers] for centuries to calculate interest, convert currencies, and solve other problems of trade” (see The Crest of the Peacock: Non-European Roots of Mathematics by George Gheverghese Joseph, Princeton University Press; Third edition, 2010.)

Work in Baghdad took the Hindu system and added the decimal point. It used decimals to write fractions, and developed notation for the zero. It provided algorithms needed to add, subtract, divide and multiply efficiently (the term ‘algorithm’ is a European mistranslation of al-Khwarizmi’s name). It provided everything needed to carry out Islamic commerce.

The Hindu-Arabic system came to Europe from contact with Mediterranean Islamic culture, particularly in Al-Andalus, modern Spain. The European scholar Gerbert of Aurillac (946-1003) studied in northern Spain; he used Hindu-Arabic numbers and an abacus much like the Heng/Zhong system of arranging columns of numbers. The quotation at the head of this Chapter is from a monk, emphasizing the importance of accurately computing the dates of church holidays.

Gerbert’s decimal numbers were an important contribution to church affairs (and he eventually became Pope!). An Italian, Leonardo of Pisa, was the son of an Italian merchant who traded in the Islamic world. Leonardo worked in customs, and learning the Hindu-Arabic algorithms was essential for his work. His The Book of the Abacus, introduced Arabic algorithms to European merchants and bankers.
Mathematics is more than commerce; its origins go very far back in the human story. Figure 18 is a sketch taken from an inscribed antler bone, with markings showing the phases of the moon: a lunar calendar (see p23). The bone may date to 32,000 BCE; however accurate the date, humans were clearly thinking about the cycles of the heavens for a very long time.

The constellations – Scorpio, Gemini, Capricorn, etc, come to us from Sumeria (modern southern Iraq), about 2000-3000 BCE. Similarly, the 360 degrees of a circle come from Babylonian base sixty counting systems.

These cultures also had very sophisticated computational techniques. Very early, numbers and writing were used for simple account keeping: amount received, amount taxed, number of cattle sold, etc. Fairly soon, though, these kinds of records became ‘hypothetical’, recording not, say, the amount of barley in storage, but the amount of grains required to make a specified quantity of beer; see Figure 19 and Hans J. Nissen and Peter Damerow, *Archaic Bookkeeping: Early Writing and Techniques of Economic Administration in the Ancient Near East*, University of Chicago Press, 1994.

This is a different kind of usage for numbers: not simply recording what is, but projecting future needs. Figure 20 shows a modern version: insurance estimates for repair of a house after fire damage, one of several hundred pages, but in 2015, generated by a computer.

Mesopotamian accounting was also sophisticated:

*The flood of documents […] contains in growing numbers not only such running accounts recording all the assets (primarily including arable land, raw materials and laborers) and liabilities (maintenance, labor costs, and so on) of the central administration, but also a standardized method of calculating the expected performance of laborers and of achieving comparable units of value of labor.*


They could predict what should be happening, and note any difference between that and actual performance: in other words, they had a system of accounting for errors and shortages.
Projecting into the future was not an invention of Mesopotamia; every sacrifice or prayer to the gods was an attempt to control an uncertain future. Professor Ulla Koch-Westenholz writes (see p23):

stars and planets were the celestial manifestations of gods, but also seem to have been gods in their own right . . . . Sometimes evil omens from a planet were seen as the expression of anger of the god whose celestial image the particular planet was (e.g. Jupiter = Marduk), so that particular god had to be appeased. In this way, messages could be sent directly from a god to the king, . . . auspicious Venus omens [are seen] as an expression of the love Ishtar holds for the king.

To gain the kings ear, astrologers needed to predict the appearance of constellations, or of eclipses, or any planetary sign: otherwise the best they could do was interpret events after they’d happened. But writing allowed them to compile heavenly events. Figure 21 shows a portion of the Enuma Anu Enlil compilation, named after the first words of the tablet, “the day when the gods Anu and Enlil...” (the ‘gods’ here referred to planets). In this example, the tablet records twenty-one years of the rising and setting of the planet Venus (modern astronomers can work backwards to date the tablet: one goes back to 1581 BCE; see p23).

Ancient Sumerian and Babylonian astronomers wrote a vast number of these texts; these relate positions of planets and constellations to earthly events. In other words, these ancient cultures practiced divination, the art of reading the future from omens. Omens could be patterns in the innards of sacrificed animals, or they could be the patterns of the heavens: astrology.

Beginning in the Babylonian period (about 1800 BCE), astronomers used the recorded position of planets over many years to develop systems for predicting future events in the heavens. Asger Aaboe (see p23) called these Systems A and B. Each divided the sky into several regions, and had the planets moving with different speeds through the different parts of the sky. The ratios of speeds was always that of numbers like 3:2, small consecutive integers (see p24).

Figure 22 gives an idea of how the predictions worked. The dots on the graph show the changes of position of Jupiter when it first becomes visible on the horizon (it’s a good place to make the observation, because ‘horizon’ is an easy measuring guide, and you can even measure the angle from the observer to the planet). The straight blue lines represent the formula used to predict planetary appearances; they follow the curve, and are the kinds of computations used for accounting.
Now this is something very strange. These astronomers knew the straight lines were not the true planetary positions; why did they use these lines? We can make lots of excuses:

i) It’s astrology. Does anyone expect it to be accurate?

ii) They used the techniques they had. What else could they do?

iii) Sure there were errors, but the errors were small.

iv) Although the predictions had errors, the numbers they really wanted turned out OK.

The very odd thing is – we are going to see these excuses continue to be used for at least another two thousand years (except for i).

We’ll follow the history, and the philosophy, of making these ‘wrong but good enough’ computations. We’ll see that this kind of issue is fundamental to the way mathematics is used in making predictions about the world.

As a side remark, Babylonian astronomy may even have given us the idea of ‘music of the spheres’. Greek theories of musical harmony also relied on ratios of small consecutive integers, just as above. They knew Babylonian tables (see p24) and may have seen the analogy.

In any case, we can see an historical answer to the question at the beginning of the chapter. By 2000 BCE, numbers were an essential part of trade, government, scientific astrology, and even religion.

Alas, this explanation only works if mathematics is useful. The deeper question was raised by Nobel laureate Eugene Wigner, in The Unreasonable Effectiveness of Mathematics in the Natural Sciences. How is it that mathematics tells us about – well, everything? Does math have some special relationship to the way the universe works?

For contemporary thinking about Wigner’s article see R. W. Hamming, at http://www.calvin.edu/~scofield/courses/m161/materials/readings/Hamming.pdf

For a bit more background on the philosophical issues raised by Wigner, see p18.
Notes for Chapter 1 Section 1: Introduction


The Christian holy days, such as Easter, were determined by the appearance of a new moon, hence Easter as a lunar holiday. This doesn’t translate easily into calendars based on cycles around the sun, or solar calendars. Computing the projected day of Easter is a kind of prediction of the future. To an early Christian, the date of the Resurrection of Christ clearly had cosmic significance – hence, without numbers, all would be lost.

The religious importance of calendars goes very far back in human history. As one example, a Jewish sect called the Essenes or Yahad, living in the desert from about 200 BCE to 70 CE, developed an alternative to the traditional Jewish lunar calendar (Figure 24). They divided the year into 364 days, so that each of the four seasons, and all religious holidays, would fall on the same calendar day every year. This contrasts with a lunar calendar, where the holidays have to be determined by the appearance of the moon. Since this has to be determined by a human, the Yahad considered it less perfect, less reflective of the divine.

p10 Here’s a typical expression of what constitutes mathematics, in the view of Newton and his contemporaries:

*Mathematics is the science of number, extension, and measure in abstraction from material things.*


We’ve suggested that ideas like extension, measure and number derive from perception. But – if we draw a circle, we can recognize it looks circular, even though nothing we draw can ever be a perfect circle. So, how do we ‘abstract’? And, if real circles are all imperfect, how can we know facts about abstract circles, like \( A = \pi r^2 \)?

These are old questions in philosophy: ‘How do we know?’ ‘What can we know?’ ‘What does it mean to understand?’ These questions are part of the study of epistemology, the philosophy of how we can know. Here, we’re asking: ‘If an abstraction isn’t a physical object, how can we perceive it or understand it?’

Plato conjectured that there was something like an idea of Circle (a *form* in his words). The form would be unchanging (eternal), unlike
the various pictures of circles we could draw. As to how it gets into our minds, Plato tried to demonstrate that the only way we could come to know eternal ideas is for our our souls to be eternal; in the passage between one death and rebirth, the soul experiences Circle directly. In life, individual drawings can only ‘partake’ of the form Circle, and our living senses can only grasp a small part of the reality. That grasping can only be done by ignoring the world of the senses, and the errors that world presents. Only by the use of reasoning and logic can a person come to apprehend something of eternal truths. This makes true understanding very much like a spiritual discipline.

Aristotle took a different view of abstraction. One example of his approach applied to Greek theories of proportion. To give a modern version: we know that if $\frac{a}{b} = \frac{c}{d}$, then it’s also true that $\frac{a}{c} = \frac{b}{d}$. Aristotle remarked that this is first established for lengths, then for areas and times. He suggests, then, that the abstraction is assuming what is, in our experience, true for these numbers must be true for numbers in general. Abstraction, then, is generalization from experience. Explanations like Aristotle’s appear again, in the idea that what is true for rational numbers is also true for real and complex numbers; this generalization was called permanence of form.

For Plato’s theory see Francis MacDonald Cornford’s Plato’s Cosmology: The Timaeus of Plato, Hackett Publishing Company, Inc, 1997. Cornford presents Plato’s dialogue, with section by section comments to guide the reader. Later Platonic dialogues discuss some of the difficulties with his theory, for example: what does it mean ‘to partake of Number’? See the Parmenides dialogue at, say, Wikipedia.


p16 Wigner’s question, how is it that mathematics can tell us about the universe, has a long history. The earliest we can reasonably go back is to Pythagoras (about 500 BCE) and the Pythagorean school or religion or sect, depending. This has been called the "most controversial subject in all Greek philosophy" because everything we know comes from sources after his death (for a close examination of the evidence, see W. K. C. Guthrie, A History of Greek Philosophy: Volume 1, The Earlier Presocratics and the Pythagoreans, Cambridge University Press; Revised ed. edition February 28, 1979).

Pythagoras believed that people have souls, which migrate across lifetimes; to him, everything in the world was sacred, ordered, and beautiful. The world seems chaotic and often brutal, are tied down by our bodies, but if we could transcend bodily experience, we could
see this order. We would no longer require a body, but exist as soul alone.

For both Plato and the Pythagoreans, the practice of reasoning and logic, and that of rejecting sense information, can only be accomplished by highly trained individuals. It seems analogous to spiritual disciplines for knowing God or achieving Nirvana. In fact, Gautama Buddha and Pythagoras were roughly contemporary. See Karl Jaspers’ idea of the Axial Age; See Wikipedia, or Karen Armstrong’s The Great Transformation: The Beginning of Our Religious Traditions, Anchor Press 2007.

How is this relevant to Wigner? Pythagoras believed that the study of the properties of number is the study of the universe.

"His most striking discovery, and the one which is said to have exercised the widest influence over this thought and to have been the foundation of his mathematical philosophy, was in the field of music. [...] The octave is produced by the ration 2:1, the fifth 3:2, and the fourth 4:3. [...] The discovery lay in the existence of an inherent order, a numerical organization within the nature of sound itself, and it appeared as a kind of revelation concerning the nature of the Universe."

W. K. C. Guthrie, The Greek Philosophers From Thales to to Aristotle, Methuen & Company Ltd, 1950

This discovery is often referred to as the idea of the kosmos: a universe of beauty and order. And, again, finding this order was linked with the goal of merging with the divine: “For Pythagoras then the purification and salvation of the soul depended [...] on philosophia; and this word, then as now, meant using the powers of reason and observation in order to gain understanding.” (Guthrie, above). This has echoes in Plato, but also in Aristotle (see p30) and Newton.

For a modern view on Wigner’s question, see Steven Weinberg, To Explain the World: The Discovery of Modern Science, Harper Perennial; Reprint edition 2016.

p10 Figure 7 and information on the robin experiment are from Simon Hunt, Jason Low and K. C. Burns, Adaptive numerical competency in a food-hoarding songbird, Proc. R. Soc. B (2008) 275, 2373.


p10 We said that the explanation ‘robins understand number’ is simplistic. What would a ‘not-simplistic’ discussion of the science and the philosophy be?
The robin experiment isn’t decisive: the robin might not even notice worms. Maybe the robin has been conditioned: when humans put things into logs, it’s edible. After all, thousands of experiments with dogs, rats, etc, lead to the idea of \textit{conditioned behavior}. Give a rat a bit of food for pressing a large button, and an electric shock for pressing a small one and it’ll quickly learn to press that large button.

This critique is from a theory of psychology known as \textit{behaviorism}. The idea is that all we can observe is what animals do, and perhaps there’s nothing more to it than associations like ‘food:humans’. You don’t need to say the robin can count, or even have any mental ideas like ‘number’. Maybe there’s nothing but conditioning.

Noam Chomsky famously pointed out this is certainly is too simple a story for, say, our use of numbers. Presented with three chickens, we’re supposed to be conditioned to say the number: ‘three’. We’re far more likely to say: ‘Whoa! You cooking a big dinner, there?’ Or ‘So are chickens on sale this week?’ And neither of those are, in any reasonable sense, a conditioned reaction.


Again: is whether the robin is ‘really’ counting? Perhaps the robin doesn’t count at all; maybe it notices the human stays longer at one hole. Or returns to one hole. And that means there’s more food in that hole. To be serious, you’d have to consider alternative ideas about what the robin is doing, then run an entire collection of experiments, varying the way the worms were presented – for example, staying near a hole for a long time but not placing any worms there.

For chimps, and human babies (Figure 25), that large number of experiments has been done: for a non-technical summary, and references to the research literature, see Gwen Dewar’s article, \textit{What babies know about numbers} at https://www.parentingscience.com/what-babies-know-about-numbers.html. These experiments rule out a large number of alternatives to the explanation ‘babies can count.’

To go more deeply than ruling out alternatives: If number is perceived by the brain, where exactly is that done in the brain? And, if we can ‘localize’ where number is perceived, what happens if that area is damaged by trauma or a stroke? The general idea is this:

\textit{Humans and many other animal species have evolved a capacity to represent approximate number. This ‘number sense’ is at the heart of the preverbal ability to perceive and discriminate large numerosities and relates to the intraparietal sulcus} [Figure 26], a brain area which contains...
neurons tuned to approximate number . . . and which is functionally active already at 3 months of age in humans. Children discriminate numerosity long before language acquisition and formal education, as early as at 3 hours after birth.

Manuela Piazza et. al., Cognition 116(2010) p33


So – does damage to this area of the brain affect the perception of number? Brian Butterworth describes an extreme case, in an individual code-named ‘Signora Gaddi’. Shown a piece of paper with three marks on it, most people could immediately say ‘three’. Signora Gaddi cannot do this, though she can count the items: ‘one, two, three’. When paying for groceries, she simply opens her purse and asks the clerk to take the right amount of money. Yet in all other areas of life, she appears to be an educated lady who leads a normal life. See Butterworth, *What Counts: How Every Brain is Hardwired for Math*, Free Press 1999. Apparently, the number deficit is separate from other mental functions – consistent with the theory that brains recognize number as such, and that it is not a consequence of other mental systems.

The research on number recognition is in Harvey, B.M., it et. al, *Topographic representation of numerosity in human parietal cortex*, Science, 341, p. 1123- (2013). Figure 27 shows how the localization is arranged. The higher numbers seem to have less and less area of the brain devoted to their recognition, suggesting this is why recognition fails at high numbers.

We’ve just seen some of the argument for where the brain perceives/processes/recognizes small numbers. The next question is how that processing rises to consciousness, or, to (verbal or physical) symbolic numbers. That transition is likely to be something learned, rather than something present at birth:

[...] the acquisition by children of the first number words and their matching to numerosities appears to be a long and hesitant process which does not seem to lead on naturally from the preverbal skills that are already in place in infants.


And again, can this transition, or linking, be localized in the brain?
One of these non-numerical circuits is in the left frontal lobe, which is associated with linguistic representations, in this case, representations of exact numerical values. The other is found bilaterally in the parietal lobes, a part of the brain associated with visuospatial functions in general, and, by Dehaene et al. in particular, with representations of approximate quantities in the form of a number line. These findings are compelling and provocative and they provide further support for the view that humans have at least two means of representing and processing quantity. One is the ability to make perceptually based judgments and comparisons. In this, degree of accuracy varies with set size. The other allows precise quantification through the use of symbols, concepts and rules.


And in more recent research,

Converging behavioural, brain-imaging, and neurophysiological results suggest that knowledge of number is an evolved competence of the animal and human brain, with a cortical basis in bilateral intraparietal cortex. The number sense hypothesis postulates that this cerebral system is available early on during development, possibly during infancy, and guides the learning of numerals and arithmetic in childhood. Indeed, an association of number processing tasks with intraparietal areas has been demonstrated in 4- and 5-y-old children.

Veronique Izard *et. al*, *Distinct Cerebral Pathways for Object Identity and Number in Human Infants*, Plos Biology, February 2008, Volume 6, Issue 2

The distinction between the internal processing of number, and the public naming of numbers, is at the core of the philosopher Ludwig Wittgenstein’s *private language* argument: is it possible for a human to have a language that no-one else has? Or does the idea of language necessarily involve rules, rules which only make sense if they are public? See http://plato.stanford.edu/entries/private-language/ or the Wikipedia article on private language: https://en.wikipedia.org/wiki/Private_language_argument

Wittgenstein’s critiques are definitely ‘not-simple’ philosophy, but there’s an preliminary case that needs to be understood. We mentioned the behaviorist view, that numbering might be nothing more than a conditioned reaction: you see three objects, hear the word ‘three’ and that’s all that’s going on in the brain. But of course, it isn’t all; to begin with, there’s the whole issue of ‘seeing three objects’.
What in our brain knows that those objects are – well, distinct objects, as opposed to blending into one large blur of landscape? How does our brain know ‘objects’? We’ll return to this in First Expedition: Vision.

p11 The hypothesis of one ancestor for all languages is called the ProtoWorld hypothesis; see the Wikipedia article https://en.wikipedia.org/wiki/Proto-Human_language. Reconstructed works are preceded by an asterisk, as in *tik. *tik means a finger; also related to indicate, to point, and also to digit, and to the number one.


The links between basic number words and body parts may have a neurological explanation. Fayol and Seron, supra, remark: "It could indeed be that the linkage between preverbal number knowledge and language is in fact mediated by the relations children establish between number concepts and the use of their fingers and hands. As rightly noted by Butterworth [*The Mathematical Brain*, Macmillan, 1999] in all human cultures, children use their fingers to count before they are systematically taught arithmetic in school."

p12 Figure 12 is from T. Cuyler Young, Jr., of the Royal Ontario Museum, Toronto, Canada, who excavated Godin Tepe in the 1960’s. It is provided from Denise Schmandt-Besserat of the University of Texas at Austin, in *How Writing Came About*, University of Texas Press; Abridged edition 1997.

p12 The Schmandt-Besserat quote is in *From Accounting To Writing*, https://sites.utexas.edu/dsb/tokens/from-accounting-to-writing. She showed written numbers evolved from ancient accounting circa 8000 BCE; government uses came later.

p14 This interpretation is due to Alexander Marshack; see *The Roots of Civilization: The Cognitive Beginnings of Man’s First Art, Symbol and Notation*, Moyer Bell Ltd, December 1991.


p15 Systems A and B for predicting planetary positions are in Asger


Chapter 1: Numbers

Section 2: Building The Numbers

We’ve argued that numbers gained their place in human life because of their public nature. Professional mathematicians form a separate community who use, and think about the uses of, number. If number were a branch of government, or a religion, they’d be the lawyers and priests. Like lawyers, they have their own understandings, which are not necessarily the everyday ways most people use numbers. One of the issues we’ll have to deal with is whether the professionals add any value to everyday understandings.

We’ll start with the numbers \{1, 2, \ldots\}, called the counting, or natural numbers. These came so early in human history that, like \(1 + 1 = 2\) they seem as unshakeable as the foundations of the earth. For mathematicians, it’s more complicated: it can take a semester to prove \(1 + 1 = 2\). First, you have to define ‘1’.

Why would mathematicians bother? It’s part of the attempt to discover true knowledge about the world, and is related to the same approach as taken by Pythagoras and Plato (p18) – the use of a strict discipline of reason alone, to avoid being confused by sense experience. What is this discipline?

Mathematicians use a system based on two foundations: axioms, and deductive reasoning. Historically,

"The word axiom comes from the Greek word αξιωµα (axioma), meaning "considered worthy".

In turn, this is from αξιος (axios), meaning "being in balance". Axios itself comes from Proto-Indo-European *ag-ty-o- "weighty"; the image here is an object weighed on a balance; checked for accuracy, thus ‘worthy’ of trust.

In Greek philosophy, an axiom was a claim which could be seen to be true without further checking. This raises issues: be seen by whom? See p30. There were suggestions, all the way from Pythagoras to Plato to Aristotle, that the status of axioms was something to be evaluated by only the most pure, or most enlightened, or most knowledgeable individuals: in short, an elite who have had special training. These same kinds of issues also arose in the development of science; we’ll discuss that later.
The 'deductive reasoning' part (from *duka-, to draw out or lead) was codified by Aristotle in the text *Prior Analytics*, where he listed the forms of correct reasoning (for 'forms' see p31). One form is the syllogism; the classic example is:

All men are mortal
Socrates is a man
therefore Socrates is mortal

(In a syllogism, 'therefore' means 'it is correct to reason')

One of the first books to follow the axiom/deductive method was Euclid’s *Elements*, a book of geometry and number theory. Euclid set definitions, axioms, postulates, common notions, etc. He then gave deductive proofs for the mathematical results.

Examples from Euclid:
Definition: "A straight line is a line which lies evenly with the points on itself."
Postulate: "[One can] draw a straight line from any point to any point."
Common Notion: "If equals be added to equals, the wholes are equal."

And the historically important 'parallel postulate'
"That, if a straight line falling on two straight lines make the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which are the angles less than the two right angles."

Proposition: "A straight line falling on parallel straight lines makes the alternate angles equal to one another, the exterior angle equal to the interior and opposite angle, and the interior angles on the same side equal to two right angles."


Why all this ... machinery and technical terminology, and where did it all come from? It’s a rather long story; see p30.

Mathematicians followed this approach in developing axioms for natural numbers. First, a bit of notation. The natural numbers are denoted by $\mathbb{N}$, and intuitively given as a list:

$$\mathbb{N} = \{1, 2, 3, \ldots\}$$

If we want to include the negatives and zero, we have the integers,

$$\mathbb{Z} = \{0, \pm1, \pm2, \pm3, \ldots\}$$

The $\mathbb{Z}$ here is from the German word *Zählen*, meaning, d’oh, ‘num-
bers’. None of this, though, tells us what numbers are, and none of this tells us what’s actually on the infinite list \{1, 2, 3, \ldots\}.

Gottlob Frege gave a definition based on the meaning of ‘counting’. Frege (a German mathematician working in the mid nineteenth century) thought back to a child pointing at toys one at a time and saying the numbers ‘one, two, three’: the process of counting is that of matching one collection with another, like toys to fingers, or number words. If we match the two collections \{\Sigma, \Delta, \Gamma\} and \{\infty, +, \int\} we’d say they have the same number of items. Matching up different collections partitions the world into chunks which have the same number, and we could define three to be the collection of all collections that have the same number as \{\Sigma, \Delta, \Gamma\}; see p37.

A different approach follows an idea older than Aristotle, common in early Greek mathematics. The number ‘1’ was considered the father of all numbers, because all numbers could be generated from him. And we wouldn’t consider ‘1’ as a number at all, as a father is not one of his own children.

A thousand years after Aristotle, the mathematician al-Khwarizmi put it this way:

Because one is the root of all numbers, number is nothing but a collection of ones.

Intuitively, \(\mathbb{N} = \{1, 1+1, 1+1+1, \ldots\}\). The Italian mathematician Giuseppe Peano used this approach to define the natural numbers through four axioms \(P_1, \ldots, P_4\).

\(P_1\): If \(n\) is a natural number, then \(n\) has a successor. Peano denoted it as \(n'\); we’ll write it and think of it as \(n+1\).

\(P_2\): If \(n+1 = m+1\), then \(n = m\).

\(P_3\): There’s a unique natural number \(n\) which is not the successor of any other natural number. We denote this as the number ‘1’.

Call \(P_1, \ldots, P_3\) the ‘plus one’ property. These axioms aren’t enough. A collection like \{\ldots – \frac{1}{2}, \frac{1}{2}, 1, \frac{3}{2}, 2, \ldots\} has the plus one property, it’s easy to make more, but they all have extra stuff. The way to eliminate the extras is by defining the natural numbers as the smallest collection with the ‘plus one’ property. Then the trick to define ‘smallest’ is to require that there are no smaller collections with the ‘plus one’ property:

\(P_4\): Assume \(S\) is a collection of natural numbers that contains 1. Assume it also has the ‘plus one’ property: whenever \(n\) is in \(S\), \(n+1\) is also in \(S\). Then \(S\) is actually all of \(\mathbb{N}\).

Axiom \(P_4\) has another name, the Axiom of Induction. But it still isn’t
enough: we don’t know if there are any actual collections where all four axioms are true.

Back to al-Khwarizmi: \( \mathbb{N} = \{1, 1+1, 1+1+1, \ldots\} \). Here, it’s the ‘…’ that trips us up (see p35). Of course, ‘…’ means, ‘continue doing the same thing forever,’ but we’d have to understand ‘forever’ – that is, we’d need a grasp on infinity. That’s the fifth axiom:

P5: An infinite set exists.

This is called the Axiom of Infinity; see p37 for more on ‘infinite’.

Now we know what the counting numbers are, what do we really know? From Frege’s construction, we’d have the number 2 as the super-collection of all “two” sets: \( \{1, 3\}, \{\pi, e\}, \{\Gamma, \sigma\} \ldots \). Yeah, that really helps. Or, we have it as \( \{1, 1+1, 1+1+1, \ldots\} \). Again, not helping.

Here’s the issue: if I cash a check for $432, the cashier could give me 432 one dollar bills, which is a major pain to use, or even to count. I’m more likely to get four hundreds, three tens and two singles. This, I can easily work with. The idea is very old; see Figure 29. For us, instead of repeating three tens, we use multiplication to write repeated addition as a product: \( 30 = 10 + 10 + 10 = 3 \cdot 10 \). Thirty three is then \( 33 = 3 \cdot 10 + 3 \). We also use exponents to make it easier to write multiplications: ten tens is \( 10 \cdot 10 = 10^2 \). This series of notations is the decimal system; written as a decimal, our cashed check becomes \( 4 \cdot 100 + 3 \cdot 10 + 2 = 4 \cdot 10^2 + 3 \cdot 10^1 + 2 \cdot 10^0 \).

With the Peano axioms, we can prove that every natural number is a decimal, but it’s tricky. The Axiom of Induction suggests a way: show that the decimals obey the plus one axioms. First, we’d show the number ‘1’ is a decimal (easy: \( 1 = 1 \cdot 10^0 \)). We’d also show that if you can write any number \( n \) as a decimal, you can also write \( n + 1 \) as a decimal. That’s easy for numbers like 43 and 44: if we know 43 is a decimal, then \( 43 = 4 \cdot 10 + 3 \), and so \( 44 = 43 + 1 = (4 \cdot 10 + 3) + 1 = 4 \cdot 10 + 4 \) is also a decimal.

This doesn’t work so nicely when we have to carry. It’s hard to write a general carrying formula, because numbers like 99, 999, 3989 etc, are all different; there’s no general pattern. So it’s all a little harder: the right technique here is called strong induction; see p38.

Decimals group numbers by size: in daily life a burger is less than ten dollars, local concerts less than a hundred, books for a semester less than a thousand, and so on. These gradations are called cognitive reference points; we actually use these to think about complicated numbers. A car repair bill of $326 becomes, in my mind, ‘around three hundred’ – see p38.
Translating that into algebra, grouping by size looks like this: I take \( n = 326 \), and note \( 100 \leq n \), but also that \( n < 1000 \). We’d write that as \( 10^2 \leq n < 10^3 \), and in general:

**The Archimedean Principle:** If \( n \) is a natural number, then there is always a power of ten, \( 10^m \), with \( m \geq 0 \) and \( 10^m \leq n < 10^{m+1} \)

The powers of ten measure all the natural numbers.
Notes for Chapter 1 Section 2: Building The Numbers

p26  We'll start with a brief discussion of the origins of the axiomatic/deductive way of mathematics, though much is likely "lost in the mists of time." The Danish historian Jens Herup (In Measure, Number, and Weight: Studies in Mathematics and Culture (SUNY series in Science, Technology, and Society) SUNY Press 1994) describes a ‘pre-scientific’ mathematics: problems and ideas shared across cultures, spread by trade. Figure 30 shows an example. We know ancient arithmetic could be carried out by arranging stones, then made into patterns such as squares and triangles. Building the patterns would suggest relationships we’d call algebraic, but were treated by Euclid as geometric.

Once you see these patterns, it’s natural to develop ideas such as even/odd numbers, Pythagorean triples \((a, b, c \text{ with } a^2 + b^2 = c^2)\). Even/odd and appear in the proof that \(\sqrt{2}\) is irrational (see p51).

But, why? Why does adding an odd number to a square number give another square number? Aristotle believed that deductive proofs were explanations, and that an explanation is only good if it is of a certain kind. To explain a complicated idea, you’d want to refer back to simpler ideas that are already understood – hence axioms, common notions.

Where do axioms and common notions come from? This is the problem of knowledge – epistemology; see the discussion on p17. As a pupil of Plato, he would have been familiar with Plato’s theory of knowledge (memories of direct contact with the world of the eternal Forms. See p18). His own view was a reaction to Plato:

"Aristotle’s account of knowledge of the indemonstrable first premises of sciences is found in Posterior Analytics II.19 […] what he says there is that it is another cognitive state, \(nous\) (translated variously as "insight", "intuition", "intelligence"), which knows them. […] What he is presenting, then, is not a method of discovery but a process of becoming wise. He says that in order for knowledge of immediate premises to be possible, we must have a kind of knowledge of them without having learned it."


As we mentioned on p26, knowing the axioms requires some special kind of – wisdom, in short. Compare the research on number recognition, p21: partly built in, partly learned.
Euclid was trained in the school of Plato, in addition to that work, he must have known Aristotle’s ideas about proof: he certainly used those. In addition, Euclid had many predecessors; several wrote texts also called Elements. A great deal of modern research has shown how Euclid incorporated and built on and to some extent systematized earlier work. For the earlier work, see Morris Kline, *Mathematical Thought from Ancient to Modern Times Volume 1* Oxford University Press 1990, Volume One. For the influence on Euclid, see for example David Fowler’s *The Mathematics of Plato’s Academy: A New Reconstruction*, Clarendon Press 1987.

Hørup remarked that once deductive mathematics began, some sort of organizational system would have been essential, to avoid proofs that depended on each other (‘circular reasoning’). He gives an example, the proof that the sum of the interior angles of a triangle is 180° degrees.

Figures 31-32-33 provide a ‘picture-proof’, which seems clear enough. But pictures can have problems:

i) Figure 32 involves drawing a parallel line; how do we know a parallel line exists? How do we know there’s only one parallel?

ii) Figure 33 uses the result that complementary angles are equal. Aristotle, in the Prior Analytics, remarks "those persons who […] are drawing parallel lines […] do not realize they are making assumptions which cannot be proved unless parallel lines exist."

The point of preliminaries like axioms, common notions, etc) is to eliminate all these concerns.

The next few pages are a discussion of the influence of formal reasoning on Western culture; to skip this, head to p37.

p26 The syllogism is called a form of reasoning. Aristotle used the noun *eidos* (εἰδος), from the Proto Indo-European verb "weyd, to see. Eidos was then originally a thing seen, it’s shape.

"[…] Aristotle focuses primarily on living things, characterized by hierarchical orderings of complex arrangements of material. So the horse […] is a complex arrangement of tissues and organs with emergent properties that make it a horse rather than a cow, for example."


For Aristotle, then, form implied a structural arrangement. In the classic "All men are mortal. Socrates is a man. Therefore Socrates is mortal", we can see the structure If we write the syllogism in symbols (as Aristotle did): "All A are B. x is A. Therefore x is B".
Forms of reasoning have had a deep effect on European intellectuals. The idea of formal reasoning, and the deductive structure and certainty of Euclid, has appealed to many philosophers and scientists, among them Galileo, Descartes, Hooke, Newton, Leibnitz, and Hobbes (see Figure 34). It even influenced Thomas Jefferson’s Declaration of Independence.

**Axioms:** We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness. That to secure these rights, Governments are instituted among Men, deriving their just powers from the consent of the governed.

**Theorem:** That whenever any Form of Government becomes destructive of these ends, it is the Right of the People to alter or to abolish it, and to institute new Government.

Aristotle wrote down many (but not all) of the forms of reasoning, hence, *formal logic*. When you couple that with Plato, who set standards for education, you get an approach to reasoning that influenced the next two thousand years of Western thought.

For example, in the European middle ages, the *trivium* (the ‘place where three roads meet’), grammar, rhetoric, and logic, would give the student the equivalent of a modern B.A.. The *quadrivium* (‘place where four roads meet’), arithmetic, geometry, music and astronomy, would be comparable to graduate study. Geometry, of course, meant Euclid’s geometry. This educational system guaranteed that all educated European men knew Euclid’s work.

The mathematician Gottfried Leibnitz was also influenced by Euclid; Leibnitz believed that “The only way to rectify our reasonings is to make them as tangible as those of the Mathematicians, so that we can find our error at a glance, and when there are disputes among persons, we can simply say: Let us calculate [calculemus], without further ado, to see who is right.” To accomplish this translation of all thought into symbols that could be manipulated formally, Leibnitz proposed that ideas are formed from a small number of simple ideas; when we find those, we can express more complicated ideas as algebraic-like combinations of the simpler ones.

The British mathematician George Boole took a significant step towards making mathematics into a Leibnitz-like language. Logic uses abstract symbols; Boole showed it can be translated into algebra, *Boolean Algebra*. One can then use algebra to express more complex forms of reasoning like “Socrates or Homer is human. All human are
mortal. Therefore Socrates or Homer is mortal.’
For the history of logic, see The Development of Logic, Mary and William Kneale, Oxford University Press 1962.

Boole titled his book The Laws of Thought, and for many intellectuals, formal logic seemed to offer a standard for all thought. It sounds like a happy ending, a resolution of all the problems of reasoning.

Does it work?

Even if we limit it to reasoning in mathematics, we can still ask: Was this the right way to think? Any approach to all of mathematics should do three things:

i) All known math can be fit into this approach.

ii) Any statement in this approach can be either proved or disproved (this is called completeness).

iii) This approach shouldn’t allow us to prove things that are false (this is called consistency).

As it happens, it is not possible to prove all these things, though mathematicians put in a very good try. The German mathematician David Hilbert working at the end of the 1800’s began by showing that Euclid’s geometry could be reduced to algebra/arithmetic, raising the issue of consistency/completeness for arithmetic. This was partly motivated by work on Euclid’s parallel postulate; two mathematicians had developed geometries that denied the postulate, raising the question whether their geometries were consistent. Hilbert’s work showed that all were equally consistent, by reducing the question to the consistency of arithmetic.

The British mathematicians-philosophers Bertrand Russell and Alfred North Whitehead began an ambitious project to show that all mathematics could be reduced to logic alone. Their project failed; Russell was able to show that their approach led to a statement much like ‘This statement is false’, which if true is false, and if false is true. Thus it’s either unprovable, so that the system is incomplete; or it is provable, making the system inconsistent. Russel and Whitehead were forced to add extra, non-logical axioms.

The Austro-Hungarian mathematician Kurt Gödel, working in the early 1900’s, was able to show that Boolean logic was both complete and consistent; this left open Hilbert’s reduction of geometry to arithmetic. In 1931, Gödel showed that arithmetic was either incomplete or inconsistent. Gödel’s proof realized some of Leibnitz’s dream, translating ideas into a system one could compute with. Gödel could assign numbers to mathematical statements (which is where arithmetic comes in). Then the construction was, like Russel’s, something like: "#2106: Statement #2106 is false."

This didn’t end mathematics; most mathematicians believe their own work isn’t going to run into these kinds of problems; applications of math are justified by how well they actually work, and, in any case, most mathematicians might argue that deductive proofs are the best tool we have to establish the accuracy of reasoning.

The British mathematician Alan Turing wrote of a simple (hypothetical) machine that could perform the operations of Boolean logic and arithmetic, and therefore could do all formal (or syntactic) reasoning. There are called Turing machines. In his 1936 paper, On Computable Numbers, with an Application to the Entscheidungsproblem he showed that Gōdel’s results hold for these machines as well.

That’s not the end of the issue, though: as with Leibnitz (p32), who believed that algebraic/deductive logic is the best way to ensure the quality of all human decisions, a group of people argued that this is the way humans actually think.

Some of this was influenced by the work of the American Claude Shannon who showed in 1937 that existing machines of that era could be wired to perform all the operations of Boolean algebra and arithmetic. The work of Turing and Shannon demonstrated that anything that can be computed can be computed by electronic circuits (Figure 35). It’s not much of an intellectual stretch, then, to argue that brains are actually performing some kind of neural computation, and therefore we can understand thought through these computer analogies. For some of the history, see Howard Gardner, The Mind’s New Science: A History of the Cognitive Revolution, Basic Books, 1987.

To quote the cognitive scientist Jerry Fodor,

"The key idea [...] is that cognitive processes are computational; and the notion of computation thus [borrows] heavily from the foundational work of Alan Turing. A computation, according to this understanding, is a formal operation on syntactically structured representations. Accordingly, a mental process, qua computation, is a formal operation on syntactically structured mental representations."


What does this mean? A sense impression or a thought would need
to be some kind of numerical-logical complex or structure (note the vagueness here) that computation could work on. Moreover, that complex would have to be purely formal (syntactic); the computation would have to work without looking at the meaning of the sense impression. Such complexes are the representations of the sense impression or thought to which Fodor refers.

Not all scientists believe all of this; for a discussion of how far these ideas may take us, and may not be able to take us, see Jerry Fodor, Reply to Steven Pinker ‘So How Does The Mind Work?’, Mind & Language February 1, 2005.

The Austrian philosopher Ludwig Wittgenstein (Figure 36) questioned the role of logic in mathematics. Note that rules – such as syllogisms – are rules about how to make correct conclusions in spoken or written language. Logic thus sets standards for linguistic behavior. His critique of logic was part of a general analysis of ‘following a rule’, or ‘going on in the same way’.

One of the issues most associated with the later Wittgenstein is that of rule-following. […] Wittgenstein [introduces] an example: ” […] we get [a] pupil to continue a series (say +2) beyond 1000 – and he writes 1000, 1004, 1008, 1012”. What do we do, and what does it mean, when the student, upon being corrected, answers ‘But I did go on in the same way’? Wittgenstein proceeds [to ask] “How do we learn rules? How do we follow them? […] Are they in the mind, along with a mental representation of the rule? Do we appeal to intuition in their application? Are they socially and publicly taught and enforced?”

Wittgenstein thus offers the suggestion that perhaps following these linguistic rules is simply a form of training, like learning the correct pronunciation of ‘Dumbledore.’ See Wittgenstein’s early work, Remarks on the Foundations of Mathematics, MIT Press 1983, or his later Philosophical Investigations, Pearson, 1973.

If logic really is mere verbal training, then it seems unlikely that it’s behind the way actual brains think. After all, much of what brains do is perceive the environment, move muscles, regulate hormones … none of this is linguistic. On the other hand, we believe that most animals have some kinds of built-in logic. If we saw a snake, we’d probably try to avoid it, as do other animals. We believe the process is something like this: “That thing over there is a snake. All snakes are dangerous. I’d better climb/run/crawl/fly/hide.” That’s the kind of inference you’d expect from Boolean logic.

But is that what’s really happening? Perhaps seeing a snake causes a
fear response, which releases the hormone cortisol, which activates all kinds of other chemical reactions, resulting in our trying to avoid the snake. All this business of ‘snakes are dangerous’ might be some verbal story we make up to explain a reaction we actually have little control over.

A paramecium (Figure 37) is a clearer and less controversial example. The paramecium, swimming happily along, might bump into something. Roughly, the bump stretches the cell membrane, causing certain pores (ion channels) to open, allowing K\(^+\) ions to leave the cell. The change in K\(^+\) concentration allows an electrical signal to spread through the cell, causing the cilia to beat. A bump in the front causes the paramecium to move backwards; and a bump in the back causes a movement forwards. There’s no "Uh oh, I bumped into something. That could be a predator. Predators could eat me; I better get moving." None of that; the only logic present is evolution.

Others have questioned whether human minds regularly use formal thinking – even in verbal thinking or talking. Here’s one kind of experiment: if our brain is built on logic, then two sentences that are logically equivalent should be processed equally well. Human experiments show this doesn’t always happen. Here’s one example: look at the following three sentences

i) "If that jacket is over a hundred, I won’t buy it."
ii) "Hey, it’s only $90. So you’re buying it then?"
iii) "Yo! I see you bought that jacket. So you got it for less than a hundred, then."

Many people have a hard time seeing which of the three are the same. See Women, Fire and Dangerous Things: What Categories Reveal About the Mind, University Of Chicago Press 1990.

The same kinds of concerns arise in the study of ethics. One view of ethics is that we have moral rules, like ‘don’t harm other humans’, and that these rules act as axioms for determining moral behavior. Contemporary research suggests that we seem to have two kinds of ethical systems. One system is based on ideas about being treated fairly, respecting elders and people in authority, keeping ourselves clean (especially cleansing before worship), and adhering to the social norms of our group. We make decisions about being treated fairly very quickly: a fast reaction is to get angry when someone cuts in front of us in line. Only after we make that judgement do we rationalize it. For this research, see Daniel Kahneman Thinking, Fast and Slow, Farrar, Straus and Giroux, 2013, as well as Jonathan Haidt, The Happiness Hypothesis, Basic Books 2006.
Whether or not human brains use it, Boolean logic is easy to implement using electric circuits. Since mathematics is an axiomatic-deductive system, perhaps computers could be programmed to do mathematics. What has happened is that mathematicians have used computers to check complicated proofs. While it doesn’t sound very creative, there is an outstanding application of this idea: the solution of the four-color problem (Figure 38).

To draw a map, you color it so that countries with common borders have different colors. The theorem proves that you need at most four different colors. It was proved in 1976 by Kenneth Appel and Wolfgang Haken of the University of Illinois; for their proof, see Appel, Kenneth; Haken, Wolfgang (October 1977), "Solution of the Four Color Map Problem", Scientific American, 237 (4).

The idea was to find a minimal collection of maps that could disprove the conjecture (about 1,900 such) and then use a computer to color them with only four colors, so you’ve eliminated all the examples where four colors might not work. With the computers of the time, it took over a thousand hours.

We now return to notes on the main text.

p27 Frege’s construction is based on the idea of equivalence of sets (that is, he introduced ways to find whether two collections had the same number). Sets \( A \) and \( B \) are equivalent if there exists a mapping \( f : A \to B \) which is one to one and onto; that is, for each element \( b \in B \), there is one element \( a \in A \) with \( f(a) = b \) (onto), and there is only one such \( a \) (one to one). The collection of all sets equivalent to \( A \) is called the equivalence class of \( A \), denoted \( [A] \) (again, the idea being that \( [A] \) is the collection of all collections of things that have the same number of objects). The collection of all equivalence classes defines numbers.

Frege wrote Die Grundlagen der Arithmetik (The Foundations of Arithmetic); soon after, the British philosopher Bertrand Russell noted that phrases such as ‘the collection of all collections’ leads to paradoxes (see Russell’s Paradox on Wikipedia https://en.wikipedia.org/wiki/Russell%27s_paradox). These paradoxes make Frege’s approach untenable. A modification would see that ‘the collection of all’ is too large; something needs to be done to restrict how many objects are in the collection.


p28 The axiom that infinite sets exist is phrased a bit oddly. It states
that there exists a set $S$ and a map $f : S \rightarrow S$ such that $f$ is one to one but not onto. Here’s the idea: let the set be our collection of natural numbers, $\mathbb{N} = \{1, 2, 3, \ldots\}$, and define $f$ as $f(n) = n + 1$. By the Peano axiom $P_1$, every $n$ has a successor $n + 1$, so $f$ is defined for every $n$. To show $f$ is one to one, let $f(n) = f(m)$; then $n + 1 = m + 1$. By Peano $P_2$, $n = m$. To show $f$ is not onto, assume there’s an $n$ with $f(n) = 1$. But this means $n + 1 = 1$, or, in the terms of the Peano axioms, ‘1’ is the successor of $n$. But Peano $P_3$ says ‘1’ is not the successor of any $n$.

The Axiom of Infinity was introduced by the Hungarian mathematician John von Neumann, as part of a very concrete construction of the natural numbers; he used set theory to construct standard meanings for 1, 2, etc. If $\phi$ denotes the empty set, then $\phi$ is a good candidate to be zero, as $\phi$ has zero elements inside. Then $\{\phi\}$ is a candidate for the number one, as it has one element in it, and:

$$1 = \{0\} = \{\phi\}; \quad 2 = \{0, 1\} = \{\phi, \{\phi\}\}; \quad 3 = \{0, 1, 2\} = \{\phi, \{\phi\}, \{\phi, \{\phi\}\}\}$$

Now we see how to continue, and the Axiom of Infinity allows us to ‘keep on going’ to yield the infinite set $\mathbb{N}$. In the study of infinities, $\mathbb{N}$ can be proved to be the ‘smallest’ infinite list. It’s a bit tricky to define how one infinity can be smaller than another, the idea is to exploit again one-to-one and onto maps. These matters were addressed by the German mathematician Richard Dedekind, who wrote the influential book Was sind und was sollen die Zahlen? (roughly, "What are numbers and what should they be?"). The individual who did most to clear up issues about infinities of different sizes was the mathematician Georg Cantor, who developed the theory of transfinite numbers in the late nineteenth century. For a popular, non-technical presentation, see Rudy Rucker’s Infinity and the Mind, Princeton University Press 2004. A basic technical introduction is in Michael J. Schramm’s Introduction to Real Analysis, Dover Publications 2008.

p28 In cognitive science, numbers like like 10, 100, 100 . . . are called cognitive reference points. Eleanor Rosch did a series of experiments consistent with the idea that these numbers were preferentially referred to when subjects thought about a collection of random numbers such as 102, 173. See Eleanor Rosch, Cognitive Reference Points, Cognitive Psychology 7, 1974 p532.

Chapter 1: Numbers

Section 3: The Problem of Fractions

It may help to think of fractions as a necessary evil.

Much of what we think of as modern mathematics was developed in civilizations, which used bureaucracies as an organizational tool. In modern terms, these governed India, China, the Middle East and Egypt. All of these developed a class of administrators – individuals who could read, write, and perform mathematical computations. The mathematics was important; the bureaucracies regulated land use, paid or forced labor, wages, livestock, agriculture, military affairs, construction, trade . . . . Often, these professional administrators would be concerned with what we think of as science – for example, astronomy.

All of this involved a great deal of counting, measuring, and computing; administrators developed techniques for doing computations efficiently (and techniques for training the next generation). These are the computational techniques we’ll examine.

We’ll start with Egypt, where our literate administrators were scribes. Workers (or forced laborers) were paid in standard-sized units of grain, or bread, or beer. The difficulty was in dividing, say three standard loaves among four workers. Each worker gets 3/4 of a loaf, but how? Do you cut each loaf in half, then half again, to get fourths? Then each worker gets three little slices? Let’s not even try jugs of beer. The scribes needed more than the ability to write fractions like 2/3; they needed to compute with fractions, and relate those computations to other kinds of numbers. We have much the same problem, for example, in food rationing during wars or other crises. Let’s say everyone gets two and a third ounces of cooking oil per week. Multiply that by 23 million people and you need .... well, you need an efficient way to do that kind of computation.

Figure 39 is an Egyptian papyrus showing Egyptian techniques: fractions were rewritten as sums of unit fractions. Unit fractions are fractions with one as the numerator, so, instead of writing 3/4, the Egyptian system was to write 3/4 = 1/2 + 1/4. This method also makes the division of the loaves of bread more practical: everyone gets a half loaf, then a quarter loaf.

Written numbers, written language, and accounting seem to have originated in the Middle East (historical Mesopotamia – see p44). Their scribal techniques developed over thousands of years, involving mixed systems (like our 9\frac{2}{3}, 2\frac{1}{4} and 2.25), as well as mixed units (pints,
quarts, liters). Mesopotamian scribes also used base sixty 'decimals'; they'd write \( \frac{1}{72} \) as \( \frac{1}{72} = \frac{5}{60} \), and \( \frac{4}{35} = \frac{1}{72} + \frac{1}{720} = \frac{5}{60} + \frac{20}{60^2} \). Or, in a modern decimal-like notation 5, 20.

Base sixty is very convenient for dealing with unit fractions: \( \frac{1}{2} = \frac{30}{60} \); \( \frac{1}{3} = \frac{20}{60} \) (try that with decimals!). The outcast here is \( \frac{1}{7} \); in modern notation, \( \frac{1}{7} = 0.142857142857142857 \) and so on. The "and so on" doesn't have an acceptable mathematical definition, but it can be thought of as a shorthand for the phrase "if you continue to divide, you will continue to get blocks of 142857".

Historically, Mesopotamian scribes wrote something like \( \frac{1}{7} = 0.142857 \) and then warned "approximation given since 7 does not divide". This leaves quite a bit out: for example, if \( \frac{1}{7} \) is the amount of tax on a piece of land, and you're a government, you want the largest number you can get away with (rounding up). If you're the one paying that tax, you want the smallest (rounding down). Writing \( \frac{1}{7} = 0.142857 \) doesn't say where you are.

You could argue that this is a basic deficiency of decimal notation, and for this reason, fractions simply aren't the same as decimals.

Mesopotamians took a different view: at some point, an unknown scribe wrote

\[
8, 34, 16, 59 < \frac{1}{7} < 8, 34, 18
\]

The decimal version is

\[
0.14285640 < \frac{1}{7} < 0.1428611
\]

Writing \( 0.14285640 < \frac{1}{7} < 0.1428611 \) tells you the largest and smallest value you could take, but now the problem is, it doesn't leave you with just one number. Our practical scribes had a solution: take the average, 0.14285785. Now we have a single number to use, and we know the largest and smallest variations.

Moderns think about this differently: we'd say 0.14285785 is an approximation to the real value of \( \frac{1}{7} \), but that it isn't the real value. The way to talk about approximate versus real value is to introduce the idea of error: error = real value - approximation. Here, the error is \( \frac{1}{7} - 0.14285785 \). This doesn't seem to help, because we don't know the real value of \( \frac{1}{7} \). But we do know how large and how small \( \frac{1}{7} \) could be:

\[
0.14285640... < \frac{1}{7} < 0.1428611
\]

Now subtract the average from all three sides:

\[
0.14285640... - 0.14285785 < \frac{1}{7} - 0.14285785 < 0.1428611 - 0.14285785
\]
And, rewriting,

\[-0.00000235 < \text{error} < 0.00000235\]

or as we’d write it today,

\[|\text{error}| < 0.00000235\]

Actually, we’d write \( |\text{error}| < 2.35 \times 10^{-6} \). This might look familiar; it hints at the modern idea of a limit. What’s missing is the epsilon ‘\( \epsilon \)’, which controls how small the error gets. We’ll deal with that in Section 3, on real numbers.

How does this this help our scribe? Imagine some lowly scribe presenting the taxes to his boss. The boss remarks, ”You have taken the seventh part; there is an error. Perhaps the tax is too small?” But now our scribe can bow low and say, ”Oh Shining One, the tax on this land is ten bushels of rice, and the error is but a part of one grain of rice”. I’d probably hate being a scribe.

There’s point to this silly story: when we talk about whether fractions are really decimals, or whether infinite decimals are limits, we’re exporting our own twenty-first century beliefs back thousands of years, to a place they don’t belong. The historian Eleanor Robson makes this point explicit:

On the constructivist historical view, the emphasis is on difference, localism, and choice: why did societies and individuals choose to describe and understand a particular mathematical idea or technique one particular way as opposed to any other? How did the social and material world in which they lived affect their mathematical ideas and praxis?


Robson takes a specific example from a Mesopotamian ”problem set”: A square is \( \frac{1}{3} \) cubit and \( \frac{1}{2} \) finger on each side; what is its area? (the answer should be \( 9\frac{1}{2} \) grains).

Here we have to deal with conversion of units and mixed decimal/fraction notation. The scribe first converts the numbers to base sixty notation, squares the number (using the same kinds of techniques we’d use to do a multiplication), and then converts the answer back to decimal/fraction notation – in different units. The actual answer is not \( 9\frac{1}{2} \); the scribe has converted the true area to one that’s simpler to write in mixed notation. Base sixty, to the scribe, is just a computational technique to make certain conversions and computations easy. There’s no issue of whether fractions are ”really” decimals.
But, historical anachronisms aside, we are after a modern understanding of fractions. Let’s start. First, notation: fractions are quotients of integers, \( \frac{p}{q} \), so we write the collection of all fractions as \( \mathbb{Q} \) (Quotients). Getting these into decimals takes work.

Let’s take a simple fraction, \( \frac{23}{10} \) and convert it to 2.3. To start, the fraction \( \frac{23}{10} \) has a piece, 2, to the left of the decimal point. What’s left over is the fractional part, the \( \frac{3}{10} \). You can access the fractional part by taking away the first part, to get 2.3 – 2 = .3. Now multiply by 10: \( 10 \cdot \frac{3}{10} = 3 \), and you have the part to the right of the decimal point.

How do I know to not multiply by 100 to get 30? Why does multiplying by 10 seem to be just right? If we’d had \( \frac{230}{100} \), multiplying by 100 = 10^2 would have been ‘just right’, and multiplying by 10 would be ‘not enough’. Let’s translate this into mathematics: the part to the left of the decimal point is in \( \mathbb{N} \); the part to the right of the decimal point is the fraction \( f \), where \( 0 \leq f < 1 \).

The mathematical way to say 10 is ‘just right’ for the fractional part \( f = .3 \) is that \( \frac{1}{10} \leq .3 < \frac{1}{10^2} \). In contrast, for \( f = .07 \), the fact that 100 = 10^2 is ‘just right’ and 10 = 10^1 is ‘too small’, gets rewritten as \( \frac{1}{10^2} \leq .07 < \frac{1}{10^3} \). The general idea is:

**The Archimedean Principle for Rationals:** If \( r = \frac{p}{q} > 1 \) is a rational number, then there is always a power of ten, \( 10^m \), with \( m \geq 0 \) and \( 10^m \leq n < 10^{m+1} \). If instead \( 0 < \frac{p}{q} < 1 \), there’s a negative power of ten, \( 10^{-m} \), with \( m \geq 1 \) and \( 10^{-m} \leq \frac{p}{q} < 10^{-m+1} \).

To check, multiply both sides of the inequality by \( 10^m \), and then you get \( 1 \leq 10^m \cdot \left( \frac{p}{q} \right) < 10 \). You’ve now got a number between one and ten; that’s the next decimal place of \( \frac{p}{q} \) (see p44 for a sketch of the proof). For a finite decimal, repeating the process will bring out all the decimal places, one by one.

Repeating decimals don’t fit this scheme very well – there’s always a fractional part left over. And again we have to ask: does this mean that decimals are the wrong idea to understand fractions?

The Chinese scholar Lui Hui (Figure 40) expressed similar ideas when calculating the value of \( \pi \): he used the approximation \( \pi \approx 3 \) and warned that this was not the true value, but was good enough for most practical purposes (see p44).
Lui Hui also said how he’d estimated $\pi$: he computed the area of a 96-agon inscribed inside a circle (a 96-agon is a 96-sided figure; for comparison, a triangle is a 3-agon). He also gave a formula for going from one approximation to a better one: see Figure 41, where he goes from the area of a 6-agon to that of a 12-agon. He then goes to a 24-agon, a 48-agon, and finally a 96-agon.

Something new happened: for $\frac{1}{7}$, we got a decimal approximation. But Lui Hui generates not just one number, but a whole collection of numbers. Technically, the collection of numbers Lui Hui gave for $\pi$ is called a sequence (see p44). The sequence is a collection of better and better approximations. Again, we think ‘limit’ and ‘numbers like $\frac{1}{7}$ are limits of actual finite decimals.’ This is teleology: we know how all these issues turned out, so we’re imagining that the path mathematics took was somehow preordained. Limits, etc, may not be at all what Lui was thinking. And perhaps at some point, it might have been argued that decimals really were a poor choice, because for fractions like $\frac{1}{7}$, ‘7 does not divide.’
Notes for Chapter 1 Section 3: The Problem of Fractions


p42 The Archimedean Principle for rational numbers begins with the division algorithm (not surprising; $\frac{p}{q}$ is a division!). Roughly, a number like $\frac{80}{9}$ can be written as $8 + \frac{8}{9}$; the second term is a fraction less than one. In Archimedean terms, this shows that if we have $10^0 < 8 < 10^1$, it’s still true that $10^0 < 8 + \frac{8}{9} < 10^1$.

The second half, dealing with $0 < r < 1$, follows by applying the the Archimedean Theorem to $\frac{1}{r} > 1$ and then inverting the inequalities.

p42 The work of the mathematician Lui Hui appeared in commentaries and solutions to the Chinese text The Nine Chapters on the Mathematical Art, written in 263 CE.

p43 Technically, a sequence is more than just a collection of numbers; there’s also a sense of one number following another (think of the cognate word sequel for the movie that follows the original; similarly the word second is also a cognate: it’s the number following the first). To say that approximations get better and better, we need a sense of the direction to go so we can get get better; the sequence provides that direction. So, a sequence comes with a first number, a second number, etc. For ‘first number’ we write $a_1$, the second would be $a_2$, and so on. The sequence is then $(a_1, a_2, \ldots)$.
Chapter 1: Numbers

Section 4: The Problem of Irrationals

We’ve been talking about mathematics as an extension of our mental perception of number, but number is just one quality we perceive. We also know qualities like size, position, length, area, angle, volume, weight . . . . Very early on, numbers were linked with these other kinds of qualities. That link created an association leading to entirely new kinds of number: those connected to geometry.

How exactly qualities like length and area came to be understood as number is a complex story, one which is still being worked out. We’ll discuss the neurophysiology in detail, in Section 9, Time and Space. See also Geometry and Decimals, p67, and Units & Standards, p72.

Geometric qualities: length, distance, area, volume . . . Why? Because bureaucracies always tax, and how do you compute tax? If you just take part of the harvest from a farmer, the farmer can hide the harvest before the tax collector visits. Instead, compute the area of the land planted, then compute how much produce the land should yield, then take part of that (actual Egyptian practices were much more sophisticated than this; see p51). A simple scheme, but ancient inheritance involved subdividing land amongst many children, so taxable lands had complicated shapes. Figure 44 shows an Egyptian computation for the area of a complicated figure.

The area of a triangle is half the base times the height, so geometers had to understand the connection between numbers and lengths. This unites counting with measuring; quantity and geometry. We use yardsticks all the time, but the link was less obvious in ancient times: Figure 45 implies tools for surveying are gifts from the Gods. People took their lengths seriously – and state taxation depended on public trust in those tools (public standards, yet).

Figure 46 shows a problem in geometry from Mesopotamia, about 1700 BCE. It’s a right triangle, with two equal sides, each side has length one. The Pythagorean theorem tells us the square of the hypotenuse is $1^2 + 1^2 = 2$, but what’s the length of the hypotenuse? On the tablet, in base 60, it’s written $1, 24, 51, 10$; in decimal notation, $1.41421$. We’d call it $\sqrt{2}$. The Mesopotamian answer in Figure 46 looks like the kind of approximations we used for $\frac{1}{2}$: it seems there’s nothing new or surprising here.

How did the Mesopotamian scribes get their approximations? No one knows, but here’s the best idea historians came up with; it gives the same answers the Mesopotamian scribe wrote.
Start with a first guess for $\sqrt{2}$: say, $\frac{3}{2}$ or 1.5. That’s too big, but $2/\frac{3}{2} = \frac{4}{3} = 1.33\ldots$ is too small. Take the average; that’ll be in-between, so it will be closer to the true value than either guess. So, if $g_1$ is the first guess, then you get a second, better guess with

$$g_2 = \frac{1}{2} \left( g_1 + \frac{2}{g_1} \right)$$

With $g_1 = \frac{3}{2}$, then

$$g_2 = \frac{1}{2} \left( \frac{3}{2} + \frac{2}{3} \right) = \frac{1}{2} \left( \frac{3}{2} + \frac{4}{3} \right) = \frac{17}{12}$$

And again:

$$g_3 = \frac{1}{2} \left( \frac{17}{12} + \frac{2}{17} \right) = \frac{1}{2} \left( \frac{17}{12} + \frac{24}{17} \right) = \frac{577}{408}$$

How good are these approximations? Square them, and compare with 2:

$$g_1^2 = \frac{9}{4} = 2.25; \quad g_2^2 = \frac{289}{144} = 2.00694; \quad g_3^2 = \frac{332929}{166464} = 2.0000060073\ldots$$

We’re back in the Lui Hui situation: not just an approximation, but a way to get better and better approximations. And as before, you carry out as many decimal places as you need, get on with your job, report to the Chief Scribe, get your beer ration and go home.

But Greek mathematicians discovered $\sqrt{2}$ is not like $\frac{1}{7}$: it cannot be written as a quotient $p/q$, and it is not a repeating decimal (for the proof, see p51). $\sqrt{2}$ isn’t rational; it’s *ir*-rational (‘*ir-*’ means ‘not’ as in ‘*ir*-relevant’). For some history of irrationals, see see p52.

So: what is $\sqrt{2}$ ? We could still say it’s like $\frac{1}{7}$, because you can get more and more digits of $\sqrt{2}$. While you can’t get them by long division, at least you can get the numbers $g_1, g_2, \ldots$, whose squares approximate 2.

But this is a kind of fraud. When you do the long division for $\frac{1}{7}$, you can see exactly where the decimal starts to repeat and why. With this, you can find the error in any one approximation. With the $g_1, g_2, \ldots$ you don’t know what the $g_1, g_2, \ldots$ are going to do, or why. Could $g_4$ be $\frac{17}{12}$ again? Then $g_5$ would be $\frac{577}{408}$ again, and you wouldn’t get better approximations. How can you rule that out? And how can you show it does get ‘better and better’?

The final answer came something like 2000 years later, so these are hard questions. Historically, the most significant response was to avoid the question, which involved rethinking the relation between geometry and number.
The Greek mathematician Eudoxus of Cnidos (408 - 355 BCE) undid the link between quantity and geometry by developing a consistent theory of magnitude. Eudoxus used magnitude as an undefined term; one could think of magnitudes as being the lengths of lines, or areas and volumes of figures; he showed how to manipulate magnitudes as ratios, analogous to manipulation of ratios of numbers. For example, we can define \( \frac{a}{b} = \frac{c}{d} \) to mean \( am = bn \). For certain line segments, such as the diagonal of a square, you can no longer think of the magnitude as being the length (as it’s irrational); you have to think that the diagonal itself is the magnitude. You can then answer questions such as, if you double the magnitudes of the sides of a square, do you double the magnitude of the diagonal (yes).

Ratios were enough to do the geometry Eudoxus and most Greek mathematicians wanted, for example, the construction of figures using a ruler and a compass (see p52).

The Eudoxian theory was influential for centuries; even Newton, in his *Arithmetica Universalis* of 1707 defined numbers as ratios of line segments. The prevailing opinion was stated by the German mathematician Michael Stifel (1486-67), who was critical of using approximations to define an irrational:

> . . . considerations compel us to deny that irrational numbers are numbers at all. To wit, when we seek to subject them to [decimal representation] . . . we find they flee away perpetually, so that not one of them can be apprehended precisely . . . Now that cannot be called a true number which which is of such a nature that it lacks precision . . . so an irrational number . . . is hidden in a kind of cloud of infinity.

Newton’s refusal to accept irrationals may seem inconsistent with his discovery of limits and calculus. However, Newton had a very classical training beginning with Euclid, and tended to think geometrically. The idea of expressing an irrational number as a limit of rational numbers might have made no sense to him.

In Europe there were no alternatives to Eudoxus for over a thousand years, which didn’t prevent (some) mathematicians from using algebra to deal with “numbers” like \( \sqrt{2} \). The mathematician Leonardo of Pisa, who wrote under the name ‘Fibonacci’, was aware of Arabic work on algebra; in 1225 he published the solution to a problem mentioned by Omar Khayyam, in his book *Al-jabr*: solve the equation \( x^3 + 2x^2 + 10x = 20 \) (see Figure 47).

Fibonacci showed there were no integer solutions, no rational solutions, and that the solution could not be constructed by ruler and compass. So the number was irrational, but of some unknown kind.

---

*Figure 47: Solving The Cubic*

Five hundred years after Omar Khayyam, the Italian mathematician Girolamo Cardano found a formula for finding roots of cubics. Here, the formula is applied to Khayyam’s equation by the Wolfram Alpha computer program.

For the method, see David W. Henderson, *Geometric Solutions of Quadratic and Cubic Equations*, www.math.cornell.edu/
The Mesopotamian scribe on p44 would probably shrug his shoulders: what did it matter, as long as he could compute three or four digits of these numbers, and keep the Chief Scribe happy?

For us it’s more difficult: am I going to run into new kinds of irrationals each time I solve a new equation?

In an attempt to create some kind of order, mathematicians began to rethink their irrationals. Fractions like $\frac{1}{2}$, and irrationals like $\sqrt{2}$, $\frac{1+\sqrt{5}}{2}$ . . . and even Fibonacci’s irrational, are all solutions of equations:

$$7x - 1 = 0, \ x^2 - 2 = 0, \ x^2 - x - 1 = 0, \ x^3 + 2x^2 + 10x - 20 = 0$$

Solutions were designated algebraic numbers, since they could be obtained by solving algebra equations with integer coefficients (note the shift away from Eudoxean geometric methods, to ideas that come from algebra).

The numbers $\pi$ and $e$ didn’t seem to be similar to algebraic numbers; the mathematician Euler remarked (1744) that these two seemed to go beyond the the techniques of algebra. As the Latin for ‘go beyond’ is ‘transcend’, Euler suggested that these two were transcendental numbers. In 1878 the number $e$, and in 1882, $\pi$, were each shown to be transcendental: that is, they were not solutions of algebraic equations with integer coefficients.

Of course $\pi$ is a solution to the equation $\cos(\theta) = -1$, and $e$ to the equation $\ln(x) = 1$; the functions $\cos(\theta)$ and $\ln(x)$ are transcendental functions. Again, new equations, new irrationalals. By this time, European mathematicians knew many new kinds of functions, for example the Bessel function $J_0(x)$, which describes the pattern of rings when light is diffracted through a small hole (see Figure 48). All light through microscopes and telescopes gets diffracted, and the presence of the first dark ring determines how close two objects have to be to blur into one, through the lens. In short, $J_0(x)$ determines the resolution of the lens. The dark rings occur when the intensity is zero, that is, $J_0(x) = 0$. Is this going to involve totally new kinds of numbers?

Just how many kinds of irrational are there?

There’s another issue: we know we can compute more and more decimal places of accuracy for $\frac{1}{7}$, and we believe we can do that for $\sqrt{2}$ and $\pi$, but what about these new numbers? Do we even know they’re decimals?
Answering this question is an invitation to a new world: chains of numbers with no decimal expansion, ascending and descending into the infinitely large and infinitely small.

To enter this world, all you have to do is imagine .9999... doesn’t equal 1. Intuitively, the decimal ‘never gets there.’ So there’s a gap: then $\gamma = 1 - .9999... > 0$ measures the size of the gap: what kind of things are inside that gap? We’re going to discover many kinds of numbers inside there – and none of those numbers are decimals. This is the problem: the world of numbers might be strange, almost beyond imagining.

Let’s look a bit at $\gamma$. How big is it? What’s the decimal expansion?

If the decimal expansion of $\gamma$ starts with something like .276..., then that first decimal place makes $\gamma > .1$. But $\gamma = 1 - .9999... = 1 - .9 - (.09999...) < 1 - .9 = .1$ We can’t have $.1 < \gamma < .1$, so $\gamma$ has to start with something like $\gamma = .0276...$. The problem is that $\gamma = 1 - .9999... < 1 - .99 = .01$ and again, we can’t have $.02 < \gamma < .01$, so $\gamma$ has to start with something like $\gamma = .00276...$

The problem is, actually, that this never stops: $0 < \gamma < \frac{1}{10^n}$ for all $n$. $\gamma$ has the decimal expansion $\gamma = 0.00000...$ but still $\gamma$ isn’t zero. So we have what we feared: a number that doesn’t have a decimal expansion at all. There goes the beer ration for our scribe.

We could say $\gamma$ is an infinitely small number, but not zero. Then $\gamma > \frac{1}{2} > \frac{1}{3} > \frac{1}{4} > \ldots$ are also infinitely small numbers with no decimal expansion. And so are $\gamma > \gamma^2 > \gamma^3 \ldots$.

So we don’t have just one infinitely small number, we have whole chains of them, getting smaller and smaller. We’ll see it gets worse – infinitely worse.

Since $\gamma < \frac{1}{10^n}$ for all $n$, then $\omega = \frac{1}{7} > 10^n$ for all $n$. Again, this means $\omega$ can’t start with $\omega = 1.374\ldots$, and it can’t start with $\omega = 1384.732\ldots$ because $1384.732\ldots < 10,000 = 10^4$ but $\omega > 10^4$. So $\omega$ can’t start with any numbers before the decimal point: $\omega$ is infinitely large. And so are $\omega^2 < \omega^3 \ldots$.

And as if those aren’t large enough, $\mu = \omega^\omega$ is larger than all of them. Now we start again and we get $\nu = \mu^\mu \ldots$ and we get chains of larger and larger infinities.

Now let’s get more infinitely small numbers:

$$\gamma > \gamma^2 > \gamma^3 \ldots > \gamma^\omega \ldots > \gamma^\mu \ldots$$

The gap contains a nightmare of infinities!
There’s supposed to be a way out, if you know limits. We’re supposed to know \(0.999\ldots = 1\) because the collection \(\{0.9, 0.99, 0.999\ldots\}\) has 1 as a limit. Let’s try that. First, a little notation:

\[
\begin{align*}
0.9 &= \frac{9}{10}; \\
0.99 &= \frac{9}{10} + \frac{9}{10^2}; \\
0.999 &= \frac{9}{10} + \frac{9}{10^2} + \frac{9}{10^3} \ldots
\end{align*}
\]

Now we can talk about the limit: to say the limit of the \(\{0.9, 0.99, 0.999\ldots\}\) is 1, is to say that for every \(\epsilon > 0\) there’s a point after which the sequence is at least \(\epsilon\) close to 1:

\[
|1 - \left(\frac{9}{10} + \frac{9}{10^2} + \frac{9}{10^3} \ldots + \frac{9}{10^k}\right)| < \epsilon
\]

But

\[
1 - \left(\frac{9}{10} + \frac{9}{10^2} + \frac{9}{10^3} \ldots + \frac{9}{10^k}\right) = \frac{1}{10^k}
\]

So we’re saying there’s a point after which \(\frac{1}{10^k} < \epsilon\). But, this is the whole thing about numbers like \(\gamma\): if \(\epsilon\) is one of our infinitely small numbers, it’s the other way around: \(\frac{1}{10^k} < \epsilon < \frac{1}{10^k}\).

Not only do numbers like \(\gamma\) mess up ideas about decimals, they mess up the whole theory of limits.

Maybe our logic is bad? We can find a mistake in one of those computations? To tell us no such \(\gamma\) could ever exist?

No.

The British mathematician John Conway, Figure 50, invented a number system called the surreal numbers, \(S\), which has infinitely small and infinitely large numbers (though his construction is nothing at all like our simple intuition for \(\gamma\)). Surreal numbers have a kind of not-really-a-decimal expansion; very very roughly, an infinitely large number might be something like

\[
\cdots + 3\gamma^2 - 5\gamma + \cdots + \frac{3}{10^2} + 7 + 2 \cdot 10^5 + \cdots + 3\omega - 5\omega^3 + \cdots + 3\mu + 2\mu^2 \cdots + 5\nu + \cdots
\]

But they’re not decimals, or any finite kind of thing; it even requires rethinking what it means to take infinite sums. See the references on p53.

There is no simple way to eliminate the surreals; you have to define real numbers in a way that will exclude them right from the start. We are making a choice; we are constructing real numbers to be the way we think they should be. The next section shows how it was done.
Notes for Chapter 1 Section 4: The Problem of Irrationals

p45 The amount of tax depends not just the area of a field, but also on the quality: fields were graded by their size and position as well as the presence of canals, trees and wells, and possible damage inflicted by floods. A second assessment was made before the harvest, followed by a final weighing of the harvested and threshed grain. See Corrina Rossi, *Mixing, Building and Feeding: Mathematics and Technology in Ancient Egypt*, in Eleanor Robson and Jacqueline Stedall, *The Oxford Handbook of the History of Mathematics*, Oxford University Press 2011.

p45 A photograph of the Yale Babylonian Collection’s Tablet YBC 7289 (c. 1800 to 1600 BCE), showing a Babylonian approximation to the square root of 2 (1, 24, 51, 10 base 60). Babylonian mathematicians knew Pythagoras’ Theorem relating the sides of a right triangle. The photo is by Yale professor Bill Casselman; see http://www.math.ubc.ca/ cass/Euclid/ybc/ybc.html).

p46 The result that $\sqrt{2}$ is not a rational is contained in the works of the Greek mathematician Euclid. The original proof is ‘lost in the mists of time’, but there’s reason to believe it was more of a plausible picture than a proof. The rough sketch of a proof that we give is a modernized version of the one he gave (and is an algebraic version of that more ancient geometric picture-proof).

The first result we need is that every integer is either even (has a factor of 2: 2, 4, 6, 8, . . .) or odd (has no factor of 2: 1, 3, 5, 7, . . .). This is clear enough from the list, and is easy to prove using induction. A little algebra gives that the squares of even numbers are even (4, 16, 36, 64, . . .), and the squares of odd numbers are odd (1, 9, 25, 49, . . .). Easy, but important, because it allows us to do things where we’d want to use square-root: normally, if $n$ had a factor of 2, all we could say that $n = \sqrt{n^2}$ has a factor of $\sqrt{2}$. But, using even/odd, we can do more: if $n^2$ has a factor of 2, then $n$ has a factor of 2, instead of just a factor of $\sqrt{2}$.

The next result we need is that you can simplify fractions by canceling out common factors: $60/36 = 30/18 = 15/9 = 5/3$ and now there’s no longer any common factors. This is a bit harder to prove; it needs our strong induction.

Now we can start: say you can find $\sqrt{2}$ as a fraction $p/q$. You might as well cancel out common factors. Then

$$\frac{p}{q} = \sqrt{2} \quad \text{so} \quad \frac{p^2}{q^2} = 2 \quad \text{so} \quad p^2 = 2q^2$$

This shows $p^2$ has a factor of 2, so $p$ also has a factor of 2; write it
\( p = 2r. \) Then \( p^2 = 2q^2 \) becomes \((2r)^2 = 2q^2\) or \(4r^2 = 2q^2\). Divide by 2 to get \(2r^2 = q^2\); now \(q^2\) has a factor of 2, so \(q\) also has a factor of 2. But we already cancelled all the 2’s, and so, no such \(p\) and \(q\) can exist.

The discovery of irrational numbers is attributed to the philosopher Hippasus of Metapontum; the irrational in question was likely \(\frac{1 + \sqrt{5}}{2}\), derived from a pentagram, but much of this is lost to history. Hippasus lived in the late fifth century BC (that is, from 500 to 401 BCE, closer to 401), and was a member of the Pythagoreans, an ascetic and mystic sect who believed that numbers represented qualities in the world: maleness, justice, etc. Ratios of integers explained musical harmony and the ‘music of the spheres’. Pythagoreans explained all the world by integers, so the discovery of irrationals was a serious challenge to their beliefs.

A note on ‘rational’ and ‘irrational’. For much of Greek mathematics, fractions were thought of as ratios. Hence, our quotients \(\frac{p}{q}\) were, to the Greeks, the rational numbers, because they were ratios. Historically, ‘rational’ is from the word ‘ratio,’ from the Latin word ‘to compute’. This in turn is from Proto Indo-European, *reh-, to ‘put in order’. None of this has meanings we associate with ‘being rational’ or ‘thinking rationally’.

New irrationals appeared in Book X of Euclid’s Geometry; these came from ‘ruler and compass’ constructions. For example, a right triangle with side lengths one has hypotenuse \(\sqrt{2}\). Now start using \(\sqrt{2}\) as a base of a right triangle; continuing on, you can get \(\sqrt{1 + \sqrt{2}}, \sqrt{\sqrt{5} + \sqrt{3}}, \sqrt{\sqrt{5} \pm \sqrt{3}}\). Euclid showed that the theory of magnitudes could generate these; his irrationals all came from geometry.

Both Archimedes and Hero of Alexandria worked on finding the value of \(\pi\), the ratio of the circumference of a circle to its diameter, or the area to the square of the radius. They approximated the circle by \(n\)-agons, dissected these into triangles, as in Figure 51, and computed areas or lengths of their sides; either led to irrationals. But they needed numbers; both used variations on the estimates

\[
p + \frac{r}{2p + 1} \leq \sqrt{p^2 + r} \leq p + \frac{r}{2p}
\]

While the estimate is reminiscent of the Mesopotamian computations for \(\frac{1}{7}\), the computation was motivated, once more, by geometry.

Figure 52 gives a geometric proof that \((p + q)^2 = p^2 + 2pq + q^2\): it shows that if you start with a square of area \(p^2\), you can get a square of larger area \((p + q)^2\) by adding two rectangles of area \(pq\), and a
Here’s how this led to an approximation for $\sqrt{2}$: $\sqrt{(p+q)^2} = p + q$, so $\sqrt{p^2 + 2pq + q^2} = p + q$. If $q$ is a small number (think $q = .01$) then $q^2$ is even smaller (think $q^2 = .0001$). In our approximation, we can ignore it; then $\sqrt{p^2 + 2pq}$ is about the same as $\sqrt{p^2 + 2pq + q^2} = p + q$. Now let $r = 2pq$, then $q = \frac{r}{2p}$ and we get Archimedes’ approximation, $\sqrt{p^2 + r}$ is about $p + \frac{r}{2p}$. This is also the result we’d get from tangent line approximations later in the book.

Chapter 1: Numbers

Section 5: The Real Numbers

In the mid-1800’s, many European mathematicians were working on irrationals (see p59). There were two issues; we’ve discussed the first: are all irrationals infinite decimals? Answering this question was part of a movement: the arithmetisation of geometry, eliminating intuitive ideas from geometry, and replacing them with better defined notions of arithmetic and algebra. The second issue is the problem of limits.

Limits, by this time, were important to all kinds of mathematics. For example, we can compute \( \sqrt{2} \). Take \( \{g_1, g_2, \ldots\} \) where \( g_2 = \frac{1}{2} \left( g_1 + \frac{2}{g_1} \right) \), etc, and if we believe that \( \{g_1, g_2, \ldots\} \) has a limit, \( g \), then the limit has to be \( g = \frac{1}{2} \left( g + \frac{2}{g} \right) \). Solving this gives \( g^2 = 2 \).

We suspected the \( \{g_1, g_2, \ldots\} \) became progressively better approximations to \( \sqrt{2} \); with a theory of limits, we can say the limit of the approximations is \( \sqrt{2} \).

A bit after Newton, calculus was differentiation, integration, approximation by tangent lines, and infinite series. Especially infinite series. These could be treated as three separate techniques, until the work of Cauchy, Figure 54, who in 1821 had published the first calculus book: Cours d’analyse. Cauchy gave careful proofs of the main results of calculus (see p59), but he also showed how the three topics of traditional calculus could be done carefully by reducing them to problems about limits. No understanding of limits: no calculus. Cauchy was aware of techniques for approximating numbers like \( \pi, \sqrt{2} \), and also of how one could talk about the accuracy of approximations, using the same ideas we discussed for computing \( \frac{1}{7} \) on p43: the error in approximating the true value \( t \) by an approximation \( a \) is given by \( |t - a| \), and we can check how small the error is by writing inequalities like \( |t - a| < 2.35 \times 10^{-6} \), as we did with \( \frac{1}{7} \) (p42).

But, leftover from Newton and Stifel, p47, there was still no idea of what \( \sqrt{2} \) is as a number rather than a ‘magnitude’. We might prove \( \{g_1, g_2, \ldots\} \) converges by showing \( |\sqrt{2} - g_n| \) is small – but to do this, we first need to know \( \sqrt{2} \) is a number. Standoff.

The mathematician Georg Cantor resolved these issues, giving a construction of the real numbers. Why construction? Because it wasn’t as though everyone suddenly hit on the one idea that was out there waiting to be discovered. Nor did everyone say ‘Why of course, that’s what I was thinking all along’. In reality, several other constructions competed for acceptance, and others besides Cantor’s are still used.

The real numbers were built, not handed down.

Figure 53: Georg Cantor
Cantor (1845-1918) solved the riddle of the irrationals, and opened up the theory of the infinite.

Figure 54: Augustin-Jean Cauchy
The French mathematician Cauchy (1789-1857) and a few of his many mathematical accomplishments. He developed the modern theory of limits, and published the first modern calculus book. He showed how all the results of calculus came from limit theorems; this is why all modern textbooks start with limits.
We’re going to do an overview of Cantor’s construction; a large part of modern mathematics depends on this. I first saw the construction in the senior mathematics courses I took at Cornell in 1970; it hasn’t gotten easier, though fifty years of thinking on it has helped.

Cantor took important ideas from from Cauchy; what he needed was Cauchy’s idea of how to talk about convergence without mentioning the limit. That idea starts with a list of numbers (a sequence) \((a_1, a_2, a_3, \ldots)\) abbreviated as \((a_k)\); it’s understood that \(k\) goes through all the natural numbers, one after another (sequentially).

If a sequence \((a_k)\) converges, then after a while all the \(a_k\) have to be close to the limit, so they have to be close to each other. The phrase ‘after a while’ translates to the existence of a number \(N\) specifying exactly when that closeness begins to happen; the word ‘all’ translates to ‘all numbers bigger than \(N\).’ In symbols, \(j \geq N, k \geq N\).

’Close to each other’ means that the distance between them is small; that translates into saying that \(|a_j - a_k|\) is small. How small? Well, to be a limit, it has to get smaller and smaller. Put another way, if you tell me how small you want it, I can make it that small. ‘How small you want it’ translates to ‘for any \(\epsilon > 0\), and ‘I can make it that small’ now translates to \(|a_j - a_k| < \epsilon\). When you put it all together, you have a definition of convergence that doesn’t mention the actual limit:

**Cauchy Condition for Convergence:** If the sequence \((a_k)\) converges, then for every \(\epsilon > 0\), there is a number \(N\) such that, whenever both \(j \geq N, k \geq N\), then \(|a_j - a_k| < \epsilon\).

Cantor introduced a second idea: we believe that the sequence \((3, 3.1, 3.14, 3.145, \ldots)\) converges to \(\pi\), but we don’t know what \(\pi\) is. Why not define \(\pi\) to be the sequence \((3, 3.1, 3.14, 3.145, \ldots)\)? In general, a real number is defined to be any sequence? Well, not any sequence: the sequence \((1, 0, 4, 0, 9, 0, 16, 0, \ldots)\) doesn’t converge, so we don’t want that; we want only the convergent sequences. That is, Cauchy sequences! Let’s check:

\[
\begin{align*}
|a_1 - a_2| &= |3 - 3.1| = .1 = \frac{1}{10} \\
|a_2 - a_3| &= |3.1 - 3.14| = .04 = \frac{4}{100} \\
|a_3 - a_4| &= |3.14 - 3.145| = .005 = \frac{5}{1000}
\end{align*}
\]

These are getting smaller, so it’s a nice start on the road to showing the sequence is Cauchy, but it isn’t enough: we need all \(j \geq N, k \geq N\), and here we only have \(j \geq N, k = j + 1\). See p59 for a full proof.

The third idea is this: go back to \((3, 3.1, 3.14, 3.145, \ldots)\). All of these are finite decimals; they’re rational numbers. So: a real number is a Cauchy sequence of rational numbers.
There’s a fourth idea, and it’s that pesky $\epsilon$ being infinitely small issue, again. Cantor avoided it by restating the Cauchy condition: instead of saying ‘for every $\epsilon > 0’$, he says ‘for every rational number $\epsilon > 0’$. The Archimedean Principle for Rationals (p42) now eliminates any worries about infinitely small epsilons.

Cantor calls the collection of Cauchy sequences the real numbers, $\mathbb{R}$. There are lots of problems:

**Question:** So what are rational numbers in this new definition?

**Answer:** A rational number, say $\frac{1}{2}$, would be the sequence $\{\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \ldots\}$.

**Question:** How am I supposed to add, subtract, etc, these things?

**Answer:** This turns out to be amazingly easy: take two real numbers $R = (r_1, r_2, \ldots)$ and $S = (s_1, s_2, \ldots)$ then define $R + S$ to be $R + S = (r_1 + s_1, r_2 + s_2, \ldots)$. Same for multiplication, etc. These are sequences of rational numbers, and it’s easy to show they’re Cauchy.

**Question:** What happens to the whole theory of limits?

**Answer:** We’d like to have the usual definitions, but the tricky part is defining statements like $|R_k - R| < \epsilon$. That’s just $R - \epsilon < R_k < R + \epsilon$, so what we really need to understand is inequalities, like $R < S$.

That’s the same as asking for the meaning of $0 < S - R$, so in the end, we just need to know what it means for a real number $R = (r_1, r_2, \ldots)$ to be positive.

Here’s where Cantor’s sneaky fourth idea comes in: using rational numbers to define inequalities. We’d like to just say that $R$ is positive if the rational numbers making up $R$ are positive, but that doesn’t work: take $R = (1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots)$. Intuitively, $R$ is the ‘limit’, and that ‘limit’ is zero, not positive. So Cantor defined $R > 0$ as follows: $R = (r_1, r_2, \ldots) > 0$ means there is a rational number $r > 0$ and an $N$ such that if $k \geq N$, then $r_k \geq r$: the rational $r > 0$ keeps the ‘limit’ away from zero.

We also put on the extra words ‘if $k \geq N’$. Again, it’s about limits: the sequence $(-1, -0.5, 0, 0.5, 0.59, 0.599, 0.5999 \ldots)$ has limit .6, which is positive: what the first three terms do is irrelevant. That what $N$ does: if I take $N = 4$, then I’m saying the first three are irrelevant. Now with $r = .5$, and $N = 4$ for all $k \geq N, r_k \geq r$.

**Question:** Many sequences of rationals converge to $\sqrt{2}$, so which am I supposed to use?

**Answer:** We can use any sequence we like. All you need to do is check: sequences $(a_k), (b_k)$ define the same real number if, intuitively, they have the same limit; see p59 for the technical details.

**Question:** What if I take a limit of sequences of reals? Do I get some weird new kind of number?
**Answer:** Actually, no. If you took a sequence of real numbers \( \{R_1, R_2, \ldots\} \) with limit \( R \), then you can also find rational numbers \( \{r_1, r_2, \ldots\} \) with \( R = \{r_1, r_2, \ldots\} \). Very roughly, the idea is this: if

\[
\begin{align*}
R_1 &= (a_1, a_2, a_3, \ldots) \\
R_2 &= (b_1, b_2, b_3, \ldots) \\
R_3 &= (c_1, c_2, c_3, \ldots)
\end{align*}
\]

Then

\[
R = (a_1, b_2, c_3, \ldots),
\]

which will work.

With this out of the way, Cantor proves three easy results:

One: Every Cauchy sequence of real numbers converges to a real number. This is called the completeness of the real numbers.

Two: Every real number is a limit of rational numbers; in fact, if \( R = (r_1, r_2, \ldots) \), then, with sloppy notation, \( \lim r_k = R \). This is called the density of the rational numbers.

Three: All the known irrationals were limits of rational numbers, so all the known irrationals are now official real numbers.

Finally, Cantor’s definition of \( R > 0 \), using rational numbers, allows us to show the real numbers \( \mathbb{R} \) don’t contain infinitely small numbers. Here’s why:

Take a supposed infinitely small real number \( \gamma = \{r_1, r_2, \ldots\} > 0 \). \( \gamma > 0 \) means there’s a rational number \( r > 0 \) and an \( N \) such that for all \( k \geq N \), \( r_k \geq r \).

Now we have our rational number \( r = \frac{p}{q} \leq \gamma < \frac{1}{10^k} \) for all \( k \). But this contradicts the Archimedean Principle for Rationals (p42), that for a rational \( r, 0 < r < 1 \), there’s an \( m \) with \( \frac{1}{10^m} < r \). No infinitely small rational numbers, no infinitely small real numbers. And, BTW, \(.9999\ldots \) actually is equal to 1.

We used decimal expansions to get an idea of what is in the infinite collection \( \mathbb{N} \); we’d like something similar for \( \mathbb{R} \). We know each real number is a limit of rational numbers, and each rational number is a repeating decimal, but it might be that the limit is some very weird kind of object. For example, we’d like \( \pi = (3, 3.1, 3.14, \ldots) \), but it might instead be more like \( (7, 2.1, 6.3, 2.4 \ldots) \). We’ll look at it in the next section, but for now, it’s worth noting, that \( (3, 3.1, 3.14, \ldots) \) isn’t necessarily the best way to approximate and to compute \( \pi \); see p60.

But, all we have right now is that the real numbers are all limits of rationals, and amongst those, the irrationals are the non-repeating decimals. We also know that every infinite decimal gives rise to a Cauchy sequence like \( (3, 3.1, 3.14, \ldots) \), whose limit is the real number representing that infinite decimal.
Cantor’s theory raised questions still troubling. For example: we can actually compute the decimal approximations for $\sqrt{2}$ and $\pi$ with the methods given by the Mesopotamians and by Lui Hui. We can even write computer programs to do the computations. Call numbers like $\pi$, $\sqrt{2}$ computable numbers. Are all real numbers computable? Cantor showed they are not: most real numbers are not computable (see p59).

If there are so many (most, actually) real numbers that aren’t even computable, how real is $\mathbb{R}$? Put another way, we have more numbers than we can understand – did we go too far? It certainly could seem so. After all, we have a very concrete understanding of what a number ought to be; recall the discussion of animal counting from p10. Numbers are likely built into the way our brain sees the world. Cantor’s construction is nothing like our built-in view of a number.

On the other hand, Eudoxus’ theory of magnitude, p47, views $\sqrt{2}$ as an undefined ‘magnitude,’ whatever that is. It may well be that the basic construct of number that evolution has given us for picking berries is not adequate for understanding a wider and wider universe.

Another answer is this: $\mathbb{R}$ is the smallest collection of numbers that contains all the limits of numbers from $\mathbb{Q}$ (and in which the ordinary rules of arithmetic hold). In technical terms, $\mathbb{R}$ is the smallest complete ordered field containing $\mathbb{Q}$.

And here’s a third answer: we talked about algebraic numbers being solutions to polynomial equations (with integer co-efficients; see p48), but numbers like $\pi$, $e$ are solutions to equations with trig or logs. From this standpoint, $\mathbb{R}$ is, in a way we’ll make precise, just what we need to get solutions to all equations. In technical terms, if $f(x)$ is a continuous function, and the equation $f(x) = 0$ has a solution, then that solution is a real number. Of course, as with $x^2 + 1 = 0$, the solution could be a complex number. But the complex numbers $z$ are all of the form $z = x + iy$ where $x$ and $y$ are real numbers, and we’re back to real numbers.

So – like the story of the three bears – $\mathbb{R}$ is just right.
Notes for Chapter 1 Section 5: The Real numbers


For a list of contemporary, alternative constructions of the real numbers, see The Real Numbers – A Survey of Constructions by Ittaty Weiss, Rocky Mountain J. Math. Volume 45, Number 3 (2015), 737-762.

p54 Cauchy’s work on convergence wasn’t motivated by the status of the irrationals. He was trying to give the methods of calculus a logical justification by supplying rigorous proofs for all the results. This type of work is ‘working on the foundations’ of the calculus; the analogy being that without good foundations, buildings collapse. Would mathematics collapse without Cauchy?

After Newton and Leibnitz developed calculus, mathematicians worked to extend their ideas and give applications to physics, mechanics and engineering; meaning and proof were low priority. This began to change in the mid-eighteenth hundreds when it became apparent that some of these applications were actually wrong. Cauchy was one of several mathematicians working to separate out the true results from the false; we’ll meet some of these later on.

p56 Two Cauchy sequences in Cantor’s construction are equivalent if they have the same limit. Since Cantor was trying to develop a theory of limits, he used a Cauchy sequence idea: Cauchy sequences \((a_k), (b_k)\) are equivalent if for every \(\epsilon > 0\), there is a number \(N\) such that, whenever \(k \geq N\), then \(|a_k - b_k| < \epsilon\). The collection of all sequences equivalent to \((a_k)\) is denoted \([a_k]\), and the real number corresponding to \((a_k)\) is actually the equivalence class \([a_k]\).

p55 For the Cauchy condition it isn’t enough to show \(|a_3 - a_4| < \epsilon\); you also have to show \(|a_3 - a_5| < \epsilon\) and \(|a_3 - a_6| < \epsilon\) etc. But for some sequences there’s a cheat: if \(|a_n - a_{n+1}| < \frac{C}{10^n}\) for all \(n\), then the sequence is Cauchy. We certainly have that with 3, 3.1, 3.14, 3.145, ... ; in fact it’s easy to show any infinite decimal gives a Cauchy sequence. Say you have .8769539 ... ; make a sequence out of it as we did with \(\pi\): (.8, .87, .876, ...). Then you get \(|a_n - a_{n+1}| < \frac{9}{10^n}\), because decimal digits can’t be greater than 9.

p58 What real numbers are ‘computable?’ To understand ‘computable’, look at our basic example, the computation of \(\sqrt{2}\). We start
with a guess $g_1$, then set $g_2 = \frac{1}{2} \left( g_1 + \frac{2}{g_1} \right)$. The sidenote shows how to make a computer program for this. We could even say that the computable numbers are the numbers that you can get from computer programs. It’s a practical definition of computing!

This means that the collection of computable numbers will be found in the collection of all computer programs (along with a lot of junk, like ‘hello world’). But the collection of all programs is contained in the collection of all finitely long collections of words. We’ll see the collection of all real numbers is the same as the collection of infinite decimals, which has the same size as the collection of all infinite lists. Canto showed that the collection of all infinite lists is much larger than the collection of finite lists, so the computable numbers are a smaller collection than the collection of all real numbers.

Put another way, if you randomly picked a number from all the reals, the probability that the number is computable is zero. This argument is part of Cantor’s theory of transfinite numbers. For a popular, non-technical presentation, see Rudy Rucker’s *Infinity and the Mind*, Princeton University Press 2004. A basic technical introduction is in Michael J. Schramm’s *Introduction to Real Analysis*, Dover Publications 2008.

On a more hopeful side, we do know that all the algebraic irrationals are computable, and of course this includes all the irrationals known to Euclid.

p57 We have a method to approximate numbers like $\sqrt{2}$ by decimals: the method gives the successive approximations 1, 1.4, 1.41, 1.414, …. If you square these numbers, you get 1, 1.96, 1.9881, 1.999396, …, but there are other ways to approximate $\sqrt{2}$.

There’s an old technique (going back to the Greeks), called *continued fractions*, that gives a very general way to approximate a large class of irrationals. Here’s the idea: $\sqrt{2}$ satisfies the quadratic $x^2 - 2 = 0$. I can’t factor that, but I can factor $x^2 - 1 = 1$ as $(x - 1)(x + 1) = 1$, hence

$$x - 1 = \frac{1}{1 + x}$$

or

$$x = 1 + \frac{1}{1 + x}$$

Oops, there are $x$’s on both sides. That’s okay, I already solved for $x$; let’s plug that into the right side:

$$x = 1 + \frac{1}{1 + x} = 1 + \frac{1}{1 + \left( 1 + \frac{1}{1 + x} \right)} = 1 + \frac{1}{2 + \frac{1}{1 + x}}$$

If we continue, we get

$$x = 1 + \frac{1}{2 + \frac{1}{2 + \frac{1}{2 + \ldots}}}$$
Now let’s compute $\sqrt{2}$. If we start with $x = 1$, we get $x = \frac{3}{2}$, then
\[ \frac{7}{5}, \frac{17}{12}, \frac{41}{29}, \ldots \]
The squares are 2.25, 1.96, 2.006944, ..., 1.99888109, ...

For a more serious overview, see http://www.maths.surrey.ac.uk/hosted-sites/R.Knott/Fibonacci/cfINTRO.html
Chapter 1: Numbers

Section 6 Part 1: Accuracy

We saw that every real number is a limit of rational numbers, and therefore a limit of decimals. So $\pi$ could be represented as the limit of $(4, 3,3, 5, 2,9, 2.95, 3.28, \ldots)$. We’d rather have it simpler, for example, $(3, 3.1, 3.14, \ldots)$. That is, we’d like to start by picking out the 3 in $\pi$, then get the .1, and so on.

The closest example of something like this was back on p29: we showed all natural numbers could be written with decimals. This followed from the Archimedean Principle:

The Archimedean Principle:
If $x$ is a positive integer, then there is always a power of ten, $10^m$, with $m \geq 0$ and $10^m \leq x < 10^{m+1}$

We showed a similar principle for rationals. Because Cantor showed each real is limit of rationals, we can prove:

The Archimedean Principle for Reals: If $x > 1$ is a real number, then there’s an $m \geq 0$, with $10^m \leq x < 10^{m+1}$.

What we’ll try to do next is use this to find the ‘decimal expansion’ for any real number. We’ll make one up: $r = 632.0141596\ldots$ With $m = 2$, $100 \leq 632.141596\ldots < 1000$. Then $1 \leq \frac{632.0141596}{100} < 10$, and therefore $r/100$ has a leading decimal digit (that is, one of the numbers {0, 1, 2, \ldots, 9}), plus a fractional part less than 1.

Now subtract the leading decimal digit ($6 \cdot 100$) from $r$, and repeat the above, eliminating the 3 and the 2, until you’re left with just the fraction (.0141596\ldots). We can also prove an Archimedean principle for fractions:

For positive fractions $y$ less than one, $\frac{1}{y} > 1$, so $10^m \leq \frac{1}{y} < 10^{m+1}$, and flipping the inequalities: there is always a power of ten, $10^m$, with $m \geq 0$ and $\frac{1}{10^{m+1}} < y \leq \frac{1}{10^m}$.

In our case, $m = 2$ and $0.01 < 0.0141596\ldots \leq 0.1$. Again, divide by .01: then $1 < 1.4159\ldots \leq 10$ and once again we have a leading decimal digit plus a fraction. Subtract the 1, and apply the same procedure to .41596\ldots.
What we’ve done is to generate a sequence, \( \{600, 630, 632, 632.01, 632.014, \ldots \} \).

After the second term, we get:
\[
0 \leq (632.0141596 \ldots) - (632.0) = 0.0141596 \ldots \\
0 \leq (632.0141596 \ldots) - (632.01) = 0.0041596 \ldots \\
0 \leq (632.0141596 \ldots) - (632.014) = 0.0001596 \ldots
\]

These inequalities are enough to show that 632.0141596 \ldots is the limit of \( \{600, 630, 632, 632.01, 632.014, \ldots \} \): all real numbers are limits of very simple finite decimals. We can call this the decimal expansion of real numbers (understanding of course, that .1 and .099999 \ldots are different expansions of the same number).

We now want to move on to a different topic: many (most) real numbers are not finite decimals, and to use them, we have to try to understand them in finite terms. This brings us back to the problem of \( \frac{1}{7} \) on p. 40. We talked about approximations and errors: \( \frac{1}{7} \) isn’t a finite decimal, but a number like 0.1428575 is a good approximation. Good because know the error: \( \text{error} = \frac{1}{7} - 0.1428575 \). On p. 40, we saw \(-0.00000235 < \text{error} < 0.00000235 \), or, \( |\text{error}| < 2.35 \times 10^{-6} \)

We’ll re-interpret the list of inequalities at the top of this page as errors. We’ll write the real number \( r = 632.0141596 \ldots \), and use \( a_1 = 632, a_2 = 632.01, a_3 = 632.014, a_4 = 632.0141, a_5 = 632.01415 \), as approximations by finite decimals.

Then
\[
0 \leq |r - a_1| < .0141596 \ldots \\
0 \leq |r - a_2| < .0041596 \ldots \\
0 \leq |r - a_3| < .0001596 \ldots
\]

We wrote these inequalities because they give a special kind of error estimate: they tell how many decimal places of accuracy we have. In particular, \( a_1 \) has got the first place to the right of the decimal correctly, \( a_2 \) has the second, \( a_3 \) the third \ldots. So we we get more and more accurate decimal places. More than that: once we get, say, the second decimal place right, it stays right.

This is a new way of thinking about error: we could say that \( \pi \) and its approximation 3.145 agree to three decimal places, and this would tell us what finite decimals we could trust. But this, in use, doesn’t work very well. Here’s why.

Let’s take the exact number \( r = 1 \), and compare it to approximations \( a = .999 \) and \( b = 1.001 \). Since \( r = 1.000 \ldots \), \( r \) and \( b \) agree to two decimal places, while \( r \) and \( a \) agree to no decimal places at all.

This suggests we want to rethink ‘number of decimal places’, and
think about errors instead. The error in approximating $r$ by $b$ is
\[ |r - b| = .001, \]
and the error for approximating $r$ by $a$ is
\[ |r - a| = .001. \]

So ‘number of decimal places’ and ‘error’ give different standards for accuracy. It gets even trickier: this time take numbers $a = 3.1451$ and $b = 3.1458$. Each agrees with $\pi$ to three decimal places. But if we rounded to three decimal places, we’d be ignoring the fact that $b$ is much closer to 3.146 than to 3.145. So, approximations to three decimal places are not rounding properly.

What we want is an idea that brings together all three ideas: error, decimal places, and rounding. The intuitive idea is to say that $r$ and $a$ agree to $n$ decimal places if they become equal when you round correctly to $n$ places. Since the cut-off for rounding up or down is 5, we get:

**Definition: Number of Decimal Places of Accuracy**

Numbers $a$ and $b$ agree to $n$ decimal places if
\[ |a - b| < .5 \times 10^{-n} \]

Repeating the approximations on the previous page,
\[
0 \leq r - a_1 < .0141596 \ldots < .005 = .5 \times 10^{-1} \\
0 \leq r - a_2 < .0041596 \ldots < .005 = .5 \times 10^{-2} \\
0 \leq r - a_3 < .0001596 \ldots < .0005 = .5 \times 10^{-3}
\]

These work out just the way we want: $r$ and $a_1$ agree to one decimal place, $r$ and $a_2$ to two decimal places. And for the example where $r = 1.000 \ldots$, $a = .999$, $b = 1.001$, both $|r - b| = .001$, and $|r - a| = .001$. As $.001 < .005 = .5 \times 10^{-2}$, $r$, $a$, $b$ each agrees with $r$ to two decimal places.

One important note: the idea ‘round to the same number’ is only an intuition. If we were to take $a = .1234$ and $b = .1236$, then we would have $|a - b| = .0002 < .5 \times 10^{-2}$, so they agree to three decimal places. If you round $a$ to three decimal places, you get .123, but $b$ rounds to .124: not the same. The point to this example is that ‘round to the same number’ is the intuition, but the real meaning is given by the definition. It’s what we’ve discussed before: these definitions set standards that everyone can implement in the same way; intuitions don’t enjoy this universality.
Decimals work well with geometric and analytic intuitions, but can be inconvenient for modern science. Two examples:

i) The absorption of light by a photosynthetic molecule occurs in .000000000000015 seconds. This is also the amount of time that electronic stock market transactions take place.

ii) The amount of CO$_2$ in the atmosphere is increasing, but temperatures on earth haven’t increased as much. A new study shows that from 1865 to 1997, the worlds oceans have absorbed about 150000000000000000000000 Joules of energy: see Figure 56. (a Joule is about the amount of energy released by dropping a tomato from three feet).

Scientific notation helps deal with zeroes. It counts the number of decimal places to the right of the leading decimal, and puts that into a power of ten. It also requires a single (non-zero) digit to the left of the decimal point. Thus, $100 = 1 \times 10^2$, $21060 = 2.106 \times 10^4$; 150000000000000000000000 Joules is $1.5 \times 10^{23}$ Joules.

For a fractional numbers like .0231, scientific notation counts the number of decimal places to the right of the non-zero leading decimal, so $.0231 = 2.31 \times 10^{-2}$ and .000000000000015 seconds is $1.5 \times 10^{-16}$ seconds. Thinking slightly differently, $.0231 = \frac{231}{10^4} = 2.31 \times 10^{-2}$. In either case, the effect is still to have only a single (non-zero) digit to the left of the decimal point.

Powers of ten are reference points for daily life, but powers of a thousand are more often used in science, engineering, and technology: we have a meter (10$^0$ meters), a millimeter (10$^{-3}$ meters), a micrometer (10$^{-6}$ meters), a nanometer (10$^{-9}$) meters, etc. 10$^{-15}$ units is a femto-unit, so the photosynthetic reaction above takes place in 15 femtoseconds. Similarly, 10$^{21}$ units is a zetta-unit, so the amount of heat energy released into the atmosphere is 150 zettajoules.

A system using powers of 10$^3$ or 10$^{-3}$ is called engineering notation. In this notation, 21060 would be written as 21.060 $\times$ 10$^3$, while .0231 would be 23.1 $\times$ 10$^{-3}$

Computer engineering is an application of engineering notation. Consumers look at processor speed: ‘wow that’s a 4 ghz processor’. Scientists prefer to measure how many computations per second a computer can do. Additions or multiplications of natural numbers are called integer operations; those can be very fast. But most scientific computations are with decimals; these are called floating point operations. Scientists and engineers need to know how many floating point operations a computer can perform per second; this is referred to as FLOPS (the FLOP measurement is not easy to calcu-
late; it’s often a measure of an ideal, rather than what a computer can consistently do on real computations; see p70).

To look at some examples of computer speeds, we need terminology. A mega- is $10^6$ (a million); a giga- is $10^9$ (a billion); a tera- is $10^{12}$ (a trillion); a peta- is $10^{15}$ (a quadrillion).

The first commercial supercomputer was the CRAY-1 (Figure 57) introduced to Los Alamos in 1976; it computed at 100 megaflops. In comparison, the first Macintosh from 1984 was about 10,000 times slower. Apple’s fastest computer in 2016, the MacPro, runs at a maximum of 436 gigaflops.

And the 2016 world speed champion is the Chinese Sunway Taihu Light, (Figure 58) running at 93 petaflops, though America, Switzerland, Japan and Saudi Arabia also have (slower) petaflop machines.

Extremely fast computers are used to do aerodynamic computations to reduce air resistance in 18-wheelers; to compute Bitcoins; to forecast the weather a week in advance; to design nuclear weapons; to model the motions of molecules as they engage in chemical reactions (for example, to design more effective drugs or to find genetic components of disease).

And, of course, computer gaming, Figure 59. As the author understands these things, the faster the graphics processor, the better a computer game looks. Major game manufacturers are happy to oblige, and some of the most sophisticated computer design and engineering now goes into game processors. Figure 60 shows Microsoft’s 2017 processor, running an astounding 6 petaflops. Personally, the author wasted an hour trying to get Lara out of the first puzzle, and has permanently sworn off games.

We mentioned floating point operations on a computer. We’ll look at computer numbers in Section 8; most modern computers use a different notation, IEEE 754 normalized floating point notation. A number like 3.145 would be written $3145 \times 10^{-3}$; the number .0023 is $23 \times 10^{-4}$.

Each of scientific, engineering, and IEEE754 put the decimal point in a different position. It might be nice if there were only one correct way to write real numbers, but — in real life, it all depends on the application.
Chapter 1: Numbers

Section 6 Part 2: Geometry and Decimals

The word ‘geometry’ comes to us from ancient Greek, *geo*, earth, and *metric*, to measure. To the Greeks, geometric lengths like the diagonal of a square clearly existed; you could construct them with ruler and compass. Aristotle drew a sharp distinction between number and magnitude (length, area, etc):

Aristotle views the division between number and magnitude to map neatly onto an exclusive division between the discrete and the continuous. Number is said to be discrete, because its parts do not share any boundary, whereas sharing a boundary is one of Aristotle’s criteria for the continuity of one item with another. Lines are continuous, in contrast, because each point may be regarded as a boundary shared between segments.


See p70

As we saw on p47, irrational numbers allowed that distinction to remain up until the time of Newton. With the arithmetisation of number,, we can reconnect number and magnitude, by showing that every point on a line corresponds to exactly one real number.

Intuitively, the connection will be like taking a ruler to measure lengths: rulers associate numbers to positions; this association is called the real line. Figure 61 shows an early ruler; humans have been using the idea for millennia.

To make a ruler, take a line, pick a starting point to correspond to the number zero, and choose some fixed length to represent the natural number 1. Mark off the line in units of 1, moving to the right. By the Peano Axioms, p27, this gives us all the natural numbers, arranged along the right side of zero (we can put the negative numbers on the left side of zero; the ideas will be the same).

To account for fractions, a ruler in English units would start subdividing the space between integers: first, cut the space in half, then each piece is cut in half again, etc. This marks off the ruler in halves, eighths, sixteenths, and so on. Again, as we move right we go from \(\frac{1}{2}\) to \(\frac{1}{4}\) to \(\frac{1}{8}\) and so on. Older yardsticks are like this; in the metric system, we divide the space between units into ten pieces, then each of those into ten again, etc. This yields a yardstick marked off in
centimeters, micrometers, millimeters and so on, are like this.

So far what we have are just ruler markings on the line; the markings correspond to finite decimals like \( \frac{1}{10} + \frac{3}{10} + \frac{4}{10^2} \): we don’t even have \( \frac{1}{7} \) yet. For that we’ll need a little more.

Let’s take \( \pi \) as an example. Since it’s a real number, it’s a limit of known decimals: \( \pi = 3.1459 \ldots \). To locate this with a ruler, the initial digit tells us it’s between 3 and 4. The second digit tells us it’s between the markings for 3.1 and 3.2, and so on: each decimal digit gives us a range of markings on the ruler, and \( \pi \) is between all of those. Does this locate \( \pi \) as a single point on the line? Let’s try it.

Imagine there was a second number, \( \phi \) sharing the space with \( \pi \). Then \( \pi, \phi \) would each be in the interval between 3 and 4, so they’d have to be no further apart than the length of that interval:

\[
-1 \leq \pi - \phi \leq 1
\]

And, as \( 3.1 \leq \pi, \phi \leq 3.2 \), again they share that interval, and can’t be further apart than the length of that interval

\[
-.1 \leq \pi - \phi \leq .1
\]

And, as \( 3.14 \leq \pi \leq 3.15 \), we have

\[
-.01 \leq \pi - \phi \leq .01
\]

We’ll cut this short: \( \pi, \phi \) share the same location, so for every \( m \),

\[
-\frac{1}{10^m} \leq \pi - \phi \leq \frac{1}{10^m}
\]

By the Achimedean Principle for reals, neither \( \pi - \phi \) nor \( \phi - \pi \) can be a positive number; hence \( \pi - \phi = 0 \). There’s only one number at these successively accurate ruler markings.

If you run this process in reverse, any point on the line sits between markings on the ruler, and these give the decimal expansion.

This construction lets us use geometric and decimal intuitions together, in a consistent way:

i) As we move right on the line, we go from 2 to 32 to 632, and we’d write \( 2 < 32 < 632 \): so we can associate rightwards motion with growing larger.

ii) Inequalities like those in i) are easy: we just check 632 has more decimal places to the left of the decimal point than 32.

iii) Decimals work well with our notions of limits to infinity: as we move right along the line, we move towards infinity, and, correspondingly, as we add more zeroes immediately to the left of the decimal points, numbers go to infinity (see p68): 98, 870, 7600, 65000 . . ..
For decimal fractions, moving right goes from .125 to .25 to .5; this is easier to see as moving from .125 to .250 to .500, and again the leading decimal place determines the size.

Again, decimals work well with the idea of a limit to zero. As we move left to zero, we add more and more zeroes immediately to the right of the decimal point: .1, .02, .003, ... (see p68).

The relation between infinite decimals and straight lines relies on intuitions about geometric lines; for example, that we can subdivide the line into smaller pieces, indefinitely. No physical system is like this: if we had a metal ruler, we’d soon come down to the individual crystals composing the metal, Figure 63. No surprise here; we understand atoms, and the atomic theory is quite ancient (the Indian school of philosophy, Vaieka argued for the existence of atoms, too small to be seen, as early as 500-600 BCE: see https://plato.stanford.edu/entries/early-modern-india/#VaiAt). So, if you believe in a world where only practical issues are important, then subdivisions and rulers stop here. At this point, then, real numbers also stop, and the supposed correspondence between infinite decimals and lines is just philosophy.

But, what about distances inside atoms? How would we measure that? One answer is that quantum theory limits our ability to simultaneously measure position and velocity; does this mean we really can’t measure very small distances?

Scientists studying black holes and quantum theories of gravity are facing exactly these issues. Some theories suggest that space itself is quantized, rather than being infinitely divisible; that there may be a smallest possible length. That length would be comparable to the Planck length: approximately $1.6 \times 10^{-35}$ meters.

There: we don’t need infinite decimals. Just decimals 36 places long. It’s not clear this is a vast improvement.

For a discussion of quantum gravity and Planck length, see the blog BackReaction (http://backreaction.blogspot.com/2006/05/) on The Minimal Length Scale.
Notes for Chapter 1 Section 6: Decimals

p66 For some of the difficulties thinking about FLOPS computations, see https://devtalk.nvidia.com/default/topic/745504/comparing-cpu-and-gpu-theoretical-gflops/

p67 Aristotle’s idea was this: we say the line is continuous; this means that if you take a single point away from the line, it falls into two pieces. The German mathematician Richard Dedekind used this idea to construct the reals; we’ll sketch his idea.

To begin, we can’t talk about ‘the line’, since that’s exactly what we’re trying to construct. We can talk about the rationals, Q, either as quotients of integers \( \frac{p}{q} \), or as finite and repeating decimals. For the pieces, we have left and right pieces, L and R. To be the left and right pieces, all L has to be left of all R: if \( l \) is in L, \( r \) is in R, then \( l < r \).

To make sure these really are the pieces, the two together have to make up all of Q. Finally, and this is picky but important, you have to actually take away a point: neither L nor R can be all of Q. We’ll call the pair \((L, R)\) a cut, or just a real number.

So a number like \( \sqrt{2} \) would have L be, intuitively, something like \( L = \{ \ldots 1, 1.4, 1.41, 1.414, \ldots \} \) and the right piece would be like \( R = \{ \ldots 2, 1.5, 1.42, \ldots \} \).

Of course, we’d have to define addition, multiplication, and so on, but there’s a very important definition we’ve left out: What does \((L_1, R_1) \leq (L_2, R_2)\) mean? An easy way to think is with a picture:

![Diagram of cuts](image)

If we think of the real number located where the arrow tips meet, then the top is smaller, because \( L_1 \) is smaller than \( L_2 \) – or you could say because \( R_2 \) is smaller than \( R_1 \).

With that settled, we could talk about limits, but Dedekind was after something different: he wanted to think of \( \sqrt{2} \) as being the next number after \( L = \{ \ldots 1, 1.4, 1.41, 1.414, \ldots \} \), called the least upper bound of L. Or, you could just as well use R, and say \( \sqrt{2} \) is the first number before \( R = \{ \ldots 2, 1.5, 1.42, \ldots \} \) – the greatest lower bound of R. In this way, we can make sense of both ‘the next after’ and ‘the first before’,
For example, if we had a sequence like \{1, 1.4, 1.41, 1.414, \ldots\}, the ‘next number after’ would be \(\sqrt{2}\), and this would be the limit of the sequence. Working along these lines would show that Cantor’s construction and Dedekind’s result in the same real numbers.

Dedekind added the ‘picky but important’ requirement that neither \(L\) nor \(R\) is allowed to be \(Q\). What if \(L = Q\)? Then we get the cut \(\omega = (Q, \phi)\). But, every \(L_1\) is smaller than \(Q\). If we look at our definition of ‘\(\leq\)’, every number every \((L_1, R_1)\) is smaller than \(\omega\). So \((Q, \phi)\) is, basically, \(\infty\).

Dedekind’s picky condition prevents us from constructing infinitely large numbers like \(\omega\).

More decimal places to the left of the decimal makes numbers larger. A decimal-based definition of limit would go something like this:

**Definition** We say \(x \to \infty\) if \(x\) contains more and more decimal places to the left of the decimal point.

We’ll see later how this matches with standard definitions of \(x \to \infty\).

We can use the same ideas of size in decimal numbers to see how to define \(x \to 0\); the numbers .5, .05, .005 get smaller and smaller. So,

**Definition** We say \(x \to 0\) if \(x < 1\) and \(x\) contains more and more leading zeros immediately to the right of the decimal point.

Each of these definitions can be made to work because of the Archimedean Property: the powers of ten measure the size of real numbers, and powers of ten are expressed as a 1 followed or preceded by zeroes.
Chapter 1: Numbers

Section 7 Part 1: Units & Standards

Woe to those who give short weight!

-The Holy Quran

This section is about standard units for measurements – meters, degrees Celsius, Joules... We start learning the standards of our culture shortly after we’re born: the words, the ways to pronounce them, how to behave towards elders, how to dress and act in public. But also, the names of numbers, letters or characters or kanji.

While many of these are universal – all cultures have words for white and red, for example (see p76) but meaning differs amongst cultures. Americans associate white with purity and red with danger, while Chinese associate red with good fortune (Figure 64) and white with death. Researchers now believe that basic color perceptions are built into our brains, but that culture tells us what to do with those colors.

In the same way, both our society and the way we interact with the world tell us what to do with numbers. The Nootka, a fishing culture on Vancouver Island, use month names like Eneecoresamilth, salmon fishing moon (see p76). The name reflects how Nootka culture interacts with the world.

Religions tell a different part of the story. Chinese New Year, Passover, Easter, Ramadan and Holi are all lunar celebrations: all are associated with the appearance of a full or new moon, and all occur on a different days or months as the years change. Even though the date of these holidays changes from year to year, lunar holidays are a sophisticated solution to scheduling public holidays before instantaneous communications. They work because everyone can look and see the moon (see p76).

So, standards have to be accessible: Figure 65 shows British standards for lengths, set in a public square outside Greenwich Observatory. Along these lines, a sixteenth-century German town defined the ‘foot’ as follows:

Stand at the door of a church on a Sunday and bid 16 men to stop, tall ones and small ones, as they happen to pass out when the service is finished; then make them put their left feet one behind the other, and the length thus obtained shall be a right and lawful rood to measure and survey the land with, and the 16th part of it shall be the right and lawful foot (see p760).
Polynesian men working together to make a boat measure off distances using the length of their finger joints. In pre-revolutionary France, over 700 local French units existed, with some 250,000 different measures (see p76).

Each of these ways of measuring works in a context: in surveying one town, a local measure of the foot is simple and useable. Boats, held together with lashed reeds, are flexible enough to allow small variations in lengths. We saw this in Babylonian astronomy, p24: to chart the position of the planets, astronomers used an easily accessible reference, the first appearance of a planet on the horizon. Again, in pre-revolutionary France, plots of land might be measured by their productivity or the difficulty in working the land. But these kinds of standards work less well when ships are made of metal plates. They also don’t travel very well: as French villages became connected into an empire, traders needed long tables of equivalents between measures, and the empire had difficulty assessing taxes.

But enforcing standard measure across an empire is difficult: Figure 66 shows the Babylonian Sun God Shamash, holding a standard measuring rod and a coiled rope, both used in surveying land. In the Hebrew Holy Bible, Leviticus 19: 36, we read:

‘You are to have honest balances, honest weights, an honest dry measure, and an honest liquid measure; I am Yahweh your God, who brought you out of the land of Egypt.’

The suggestion in these quotations, and in the header to this Section, is that standards are a basic form of honesty and necessary for public order. What better way to enforce standards than to connect them with the law of God?

Even today, standards are a serious issue. In a hospital,

Clinical error and negligence are responsible for disabling injuries in about 1 in 25 hospital admissions. Most of these injuries are caused by adverse drug events [...] Converting among ratios, percentages, international units, mols, micrograms, and milligrams causes substantial difficulty (see p76).

Here’s an example anyone might come across (see Figure 67): the illegal psychoactive drug MDMA. It has an average half-life of 8 hours in the body, and drug tests can detect it in blood concentrations as low as 500 nanograms per milliliter. An average amount of blood is 5.9 liters, and average dose is 120 milligrams. How long before MMDA is undetectable?

You could easily spend ten minutes converting these different measures into only one standard unit.
Modern scientific units are based on the metric system, first introduced in post-revolutionary France in the late 1700’s. The standardization of measures was taken up by the (formerly Royal) Academy of Science; revolutionary principles and Enlightenment ideals suggested standards should be based on nature itself, and would then be universal. Eventually, four major principles were set:

i) The unit of length should be based on a fraction of the circumference of the earth – measured on a circumference that passed through France (of course).

ii) The standards should be linked together into one coherent system: for example, length is measured in meters, so the units for area have to be square meters.

iii) Numbers should be expressed in base ten.

iv) Fractions and multiples of basic units should be named systematically, using Greek prefixes: milli-meter, mega-watt, etc. These are to denote units in ranks of 10, 10^2, \ldots

Little was natural or universal about these principles (see p77); many scientists argued for using a pendulum to measure lengths: the time for a pendulum to go a full cycle is

\[
t = \sqrt{\frac{L}{g}}
\]

Here \( g \) is the acceleration of gravity, and \( L \) is the length of the pendulum. Now set the meter to that length which lets the pendulum go through a full cycle in exactly one second. Unfortunately, \( g \) varies with one’s position on the earth, though this wasn’t well understood at the time (see Figure 68).

Other objections were that a base eight system (instead of a decimal system) would allow shopkeepers to easily compute half, then half again, and half again. A base twelve allows halves, thirds, sixths. And, why Greek names? (“These names are novel and unintelligible to the large majority of our citizens, are not necessary for the maintenance of the Republic.”)

All of these early standards were based on measures of ‘natural’ objects; the kilogram was based on the weight of water at a given temperature, and this standard was converted into an equivalent weight of a platinum-iridium bar kept in a bank vault in Paris. What happens next is told in Rachel Courtland’s article *The Kilogram, Reinvented* (see p77):

*Once every few decades, a scientist plucks the cylinder from its perch with chamois-leather-padded pincers, rubs its surface with a cloth soaked in*
alcohol and ether, and steam-cleans it. Then he puts the prototype in a precise balance that compares it to the bureau’s official copies, which are in turn compared to copies kept by member countries. And thus the prototype mass trickles down to set the standard for the rest of the world.

The system has been far from seamless. When the cylinder was last removed from the vault in 1988, the bureau’s metrologists were disappointed to discover that its mass and those of its official copies had drifted apart by as much as 70 micrograms since 1889.

The metric system was adopted on 23 September 1801 (or, speaking of standards, the Revolutionary government gave the date as 1 vendémiaire an X); many businesses covertly kept the old measures. Government documents, as well as legal, military and engineering documents were required to be submitted in metric units, though for some time dual systems were used in the government. And, of course, public standards needed to be publicly available; the Agency of Weights and Measures printed seventy thousand conversion tables. Mass production of meter sticks was more difficult, and eventually the Committee for Public Safety turned the task over to the Atelier de perfectionnement – an armory specializing in mass production of interchangeable parts for rifles (see p77). In France, use of the old units only died out in the early 1900’s.

Internationally, the United States still uses British units. In 1999, a joint American-British space probe crashed into the surface of Mars because flight controllers had no real idea where the probe was. Software designers in Britain used meters; those in the United States used feet, and the ‘orbiter’ entered too deeply into the Martian atmosphere and broke up. The NASA report (see p77) noted that metric units were specified in the contract; that NASA had a history of reusing old, undocumented code, and that two navigators had reported problems but were ignored.

Standards alone accomplish little; only people can make them work.
Notes for Chapter 1 Section 7 Part 1: Units & Standards

The universality of certain colors, called focal colors, contradicts theories of cultural and physical relativism. Brent Berlin and Paul Kay looked at 20 oral world languages, and

"Berlin and Kay found out that people focus certain points in the color continuum as a kind of orientation. Such reference points or ‘best examples’ were called ‘foci’. Focal colors had not only been detected in English but also in the remaining 19 languages (Berlin & Kay, Basic Color Terms: Their Universality and Evolution, Berkeley: University of California Press 1969).

Eleanor Rosch went further, using experiments to determine the physical and psychological role of focal colors:

Rosch was able to find out that focal colors are more perceptually salient than non-focal ones (Rosch, Eleanor (1973), Natural Categories, Cognitive Psychology 4: 328-350). This cognitive salience is probably not anchored in language but reflects certain physiological aspects of [...] perspective mechanisms. Later, she coined the term ‘prototype’ instead of ‘focal’.

Quotations are from http://www.glottopedia.org/index.php/Focal_Colors

Nootka numbers from William J. Folan

Even in a lunar calendar, deciding the exact day the moon is full or new has to be standardized. In Islam, this is the duty of the Imam. In a small village in Africa, it would be the duty of the Chief Priest. Chinua Achebe, the Nobel prize-winning author, describing the process in his novel Arrow of God:

‘The moon he saw that day was as thin as an orphan fed grudgingly by a cruel foster-mother. He peered more closely to make sure he was not deceived by a feather of cloud.’

The quotation defining the foot is from Jacob Koebel,
Geometrei. Von künstlichem Feldmessen und absehen.


Estimate on number of measures in France from Ken Alder, The


Part of the issue was that powerful political forces were at work. The Ministry of Finance wanted to compute taxes for the entire country; scientists at the Academy received large grants to perform the difficult and inaccurate job of measuring the circumference of the earth. But, in the end . . . opposition was so intense that the Emperor Bonaparte rescinded the system, and it was delayed for decades. The metric day, the metric week (a ten day week, with nine work days and one day of rest) and the metric year never caught on.


p74 Quotation from Rachel Courtland’s article is from http://spectrum.ieee.org/consumer-electronics/standards/the-kilogram-reinvented

p75 For the response to the introduction of the metric system, see J.L. Heilbron, The Measure of Enlightenment, in The Quantifying Spirit in the 18th Century, Tore Frangsmyr, J. L. Heilbron, Robin E. Rider (editors); Uppsala Studies in History of Science, University of California Press, 1990.

Chapter 1: Numbers

Section 7 Part 2: Orders of Magnitude

I first heard about orders of magnitude in 1968, at a public astronomy seminar, Cornell University. In 1967, the radio astronomer Jocelyn Bell had detected a signal with a peak repeating every 1.33 seconds (see Figure 70 and p83). The signal followed the rotation of the earth, so she and her advisor Anthony Hewish concluded it was extraterrestrial; they called the signal LGM-1 (for ‘Little Green Men’: a signal so regular could be from an extraterrestrial civilization). The signals were later understood to come from stars, and the stars were named pulsars.

When Frank Drake returned from the Aricebo radio telescope in Puerto Rico (see Figure 78 and p83) with data suggesting a new pulsar in a different part of the sky, Cornell opened a seminar to discuss what might cause these signals.

Memory fades after fifty years. I recall Hans Bethe was there, the Nobel laureate who explained the nuclear reactions that produce the energy of stars. Drake opened, discussing how the telescope needed to be modified to receive these signals; he had a slide with a graduate student doing the updates, hanging upside down 150m above the valley floor. Edwin Saltpeter was there (he’d explained the production of carbon in stars). He discussed the energy required to generate the signal; for the Crab pulsar (Figure 71) the energy is order of magnitude $10^{32}$ joules/year. Saltpeter explained for the non-physicists that a joule was about the order of magnitude of the energy released by a tomato falling one foot onto a floor.

Carl Sagan stood and remarked the signal was about five orders of magnitude greater than the entire energy output of the sun, and therefore, if the signal came from an intelligent civilization, ‘it was a very stupid intelligent civilization’ – there are many ways of sending signals requiring orders of magnitude less energy.

Finally, Tommy Gold (who had developed the theory of the magnetosphere surrounding the earth) stood and proposed that pulsars were highly magnetized rotating neutron stars. As the star rotated, its magnetic field would accelerate particles to near the speed of light, and those would then emit synchrotron radiation like the blue light seen in the Crab nebula, Figure 71. Gold’s explanation is the accepted explanation today (see p84).
It was amazing to see these brilliant scientists doing science, thinking about results, doing computations off the top of their heads. It was clear I had a lot to learn, and not just about orders of magnitude.

The powers of ten ($10^0$, $10^1$, $100 = 10^2$...) form standards for the way we think about numbers – we think of sandwiches as costing about ten dollars, tickets to a theme park about a hundred (see the research earlier referenced as *cognitive reference points*, p28 and the references there). The idea of *orders of magnitude* make this precise.

What we want to do, roughly, is say the pair 4, 7 are the same size, as are the pair 60, 80. In scientific notation, we’d have $4 \times 10^0$, $7 \times 10^0$ versus $6 \times 10^1$, $8 \times 10^1$. The different exponents tell us the two pairs are different orders of magnitude.

But what about 9.9999999 versus 10.00000001? It’s misleading to say one is substantially larger/smaller than the other. Should I be rounding? Where’s the cut off from one order of magnitude to the next? For example, if 9 should belong with 11, what about 8? Maybe 5 is the cut off, since it’s half-way?

Once again, it’s about setting standards: we have to establish an unambiguous system. One way to do this is think about what we said, "different exponents tell us the two pairs are different orders of magnitude". Exponents can always be accessed as logarithms of the numbers; more precisely, if $x = 4.768 \times 10^8$, $\log_{10}(x) = 8.5761 \ldots$, and it’s the 8 we want to pick out (this is called the *characteristic* of $x$). But even this is a bit tricky. $50 = 5 \times 10^1$ and the log is 1.6989..., so we’d give it order of magnitude 1. On the other hand, $.05 = 5 \times 10^{-2}$ but the log is $-1.301029957$, so do we give it order of magnitude $-1$ or $-2$?

The standard we set is this (see p84):

**Order of Magnitude:** To find the order of magnitude of a number $x$, take $\log_{10} x$, and correctly round it.

For example, $x = 2.78$ has $\log_{10} x = 0.4440447959$. Since we have $0.4440447959 < .5$, this is rounded down to zero and 2.78 has order of magnitude zero. For $x = 3.78$, $\log_{10} x = 0.5774917998 > .5$, so this number is rounded up to one, and 3.78 has order of magnitude one.

Most people can’t compute logs in their head. Without logs, the cut off between order of magnitude $n$ and $n + 1$ is $10^{n + \frac{1}{2}} = 10^{\frac{3}{2}} \cdot 10^n = 3.1622776602 \cdot 10^n$. Then $2.78 < 3.1622776602 < 3.78$, though likely we’d use 3.16 instead of 3.1622776602.
Next, we’ll look at some examples – besides astrophysics – where people actually use orders of magnitude.

The spar of an airplane wing is a support beam; see Figure 72. Many different kinds of structures attach directly onto it – for example, the airplane engines and the ailerons that control turning. One consequence is that the spar has to be a precise size, so that each piece will fit in exactly the right place. For a modern 27 meter long airplane wing, the spar has to be constructed to .3mm accuracy, a difference of five orders of magnitude. Such extreme accuracy is needed because of the strong forces acting on the wing during flight; differences between the wings could cause instability, or help cracks form and spread, causing the loss of the wing during flight.

Our second example comes from square-cube ideas. It’s inspired by recent work on environmental remediation. At the height of the space race in the 1960’s launch complexes were used to clean and degrease rocket engines; the chemical used was trichloroethylene (TCE), which is now known to be toxic and carcinogenic. It seeps into the ground and, over long periods of time, can contaminate groundwater. See Figures 73 and 74.

To de-contaminate the soil, metal particles are injected into the ground; these combine with TCE to form new compounds that are no longer toxic. The question is how the metals should be delivered. If we think of these as small spheres, the TCE will act on the surface of the sphere; for maximum efficiency, the surface area should be as large as possible. We’ll need a couple of formulas: A sphere of radius \( r \) has surface area \( S = 4\pi r^2 \) and volume \( V = \frac{4}{3}\pi r^3 \). Now let’s compare two sizes of metal balls.

In the first, our spheres are about the size of a cell: the radius is 100\( \mu \)m or 10\(^{-4} \)m. This gives it a volume of \( \frac{4}{3}\pi 10^{-12} \)m, and a surface area of \( 4\pi 10^{-8} \)m.

Now, instead of using cell-sized metal balls, let’s use virus-sized particles. Their radius is 100nm or 10\(^{-7} \)m, with a volume of \( \frac{4}{3}\pi 10^{-21} \)m, and a surface area of \( 4\pi 10^{-14} \)m. This is worse – smaller surface area – but, since we’re using smaller metal balls, we’ll have more of them. How many more? Divide the volume of the large balls by that of the small. Cancelling the \( \pi \) and the \( \frac{4}{3} \), it’s 10\(^{-12} \) divided by 10\(^{-21} \), or, 10\(^9 \).

Now I have 10\(^9 \) particles, each with surface area \( 4\pi 10^{-14} \)m, so my new surface area is \( 4\pi 10^{-5} \)m. The single larger ball had a surface area of only \( 4\pi 10^{-8} \)m, so I’ve increased the ability of the metal to deactivate TCE by three orders of magnitude. It’s nano-particle remediation.
Our last example of order of magnitude at work is digestion: we’ll follow a piece of potato as it’s converted into sugars and fats in the body (see p84 for references).

We’ll be taking a journey from organs we can see with our eyes and feel with our hands, to smaller and smaller structures, all the way down to single molecules. We can no longer see or manipulate these tiny structures; what we know about them is the result of centuries of scientific research. At the same time, the closer we zoom in on these micro-structures, the more exact understanding we get on how food is processed.

Digestion actually starts in the mouth, goes on to the esophagus, stomach and small intestine. Some of this is mechanical: food is broken into small pieces. Size does matter here: a variety of enzymes (for example proteases and lipases) are responsible for breaking down proteins and fats. These enzymes attach to the surface of the food, so the larger the surface area presented to them, the faster and more thoroughly they work.

If we look at a piece of food – say a sphere 1cm in radius – the surface area $S$ is about $12.6 \text{ cm}^2$, order of magnitude 1. If the food is broken down to a thousand smaller spheres, $S$ is about $270 \text{ cm}^2$, order of magnitude 3. This is why we chew: the food is preprocessed to be digested two orders of magnitude more efficiently. Snakes and alligators gulp, and it can take days for a snake to digest a mouse.

Our small particles enter the intestine next. An average small intestine is about 6m long – say order of magnitude two. The intestine is lined with circular folds shown in Figure 75. Each fold is roughly 8mm or $8 \times 10^{-2} \text{ m}$ high: order of magnitude $-2$, a jump of five orders of magnitude.

Each circular fold is has its surface coated with thousands of wormy-like tubes called villi, about 1.5mm or $1.5 \times 10^{-3} \text{ m}$ high (see Figure 76); we’ve jumped down one more order of magnitude. But that 1.5mm high tube has a surface area; villi increase the entire surface area of the intestine – the area that can absorb nutrients – by a factor of thirty: one order of magnitude.

The villi themselves are coated with a brush border, made up of further wormy tubes called microvilli (Figure 76 again), $1 \mu\text{m}$ or $10^{-6} \text{ m}$ high, they serve much the same purpose as the villi. For a decrease of three orders of magnitude in size, they increase the surface area by a factor of 600, or three orders of magnitude.

And there it ends – almost. The folds and villi and microvilli can
grab food from the flow through the intestines, but fats still need to be broken down by lipases, proteins by proteases; complex carbohydrates need to be hydrolyzed to monosaccharides. Finally, all these molecules have to get to the bloodstream.

All this occurs in a thin layer covering the microvilli called the *glyco-
calyx*. This is made up of actin filaments, a protein that usually forms the contractile filaments of muscle cells, but in this case forms a protein layer that can contract to keep fluid moving. The actin molecules themselves are about 5 nm or $5 \times 10^{-9}$ m, another two orders of magnitude smaller. The brush border holds digestive enzymes, and the resulting sugars, etc, are transferred from the interior of the intestine to the bloodstream by the cells comprising the villi, shown in Figure 77.

The cells have transport molecules located in their membrane; one transport might carry selected sugars through the cell wall: glucose, for example. This has a molecular diameter (see p85) of 9 Angstroms or $9 \times 10^{-10}$ m, another order of magnitude. These are small enough to get from the microvilli to the blood through cell pores (actual gaps in the cell wall) about 500 to 800 angstroms or $5 \times 10^{-8}$ m to $8 \times 10^{-8}$ m – certainly large enough to let the glucose flow freely into the blood.

Let’s close with a note from a working neuroscientist:

“Single large neurons have physical dimensions observable at low optical magnification, that of a tenth of a millimeter. That is big enough to be dissected by hand with pins, using a good magnifying glass. Moving just two orders of magnitude down to the micrometer level, which requires a good microscope, one is at the scale of synaptic transmission. One may observe synapses at the union between nerve and muscle, for example. Two orders of magnitude further down, at tens of nanometers, with the aid of electron microscopy, we find the realm of single ion channels and of signal transduction and molecular biology. (Rodolfo R. Llinas *I of the Vortex: From Neurons to Self*. The MIT Press, 2001.)

What do we learn from all this? It’s all very complicated, with small built upon smaller and smaller.

We see the enormous range of sizes needed to understand digestion: from the 6m length of the intestine, to eight orders of magnitude smaller for the molecules produced by digestion. And we see the dramatic range that scientific instruments need to span, to understand even the very basic processes of life.

Order of magnitude helps us organize that sense of complexity by examining comparative sizes, discarding details.
Notes for Chapter 1 Section 7 Part 2: Order of Magnitude

p78 The paper by Hewish, Bell et. al. is in Nature, 217(1968) p709. The authors remark:

The remarkable nature of these signals at first suggested an origin in terms of man-made transmissions which might arise from deep-space probes, planetary radar, or the reflection of terrestrial signals from the moon. None of these interpretations can, however, be accepted because the absence of any parallax shows that the source lies far outside the solar system. [...] A tentative explanation of these unusual sources in terms of the stable oscillations of white dwarf or neutron stars is proposed.

p78 The Aricebo telescope consists of a reflecting mesh hung above shade-tolerant vegetation near the town of Aricebo, Puerto Rico (also known as La Villa del Capitan Correa); see Figure 78. The mesh is about 1000m in diameter, and reflects radio waves to a detector about 150m above the dish. Construction of the telescope was an enormous engineering and political undertaking.

Aricebo was conceived by DARPA (the Defense Advanced Research Projects Agency) as a means to detect incoming guided missiles: entering high speed objects ionize the upper atmosphere, and this can be detected on radar. As the upper atmosphere was poorly understood, the telescope was designed to carry out atmospheric research; W.E. Gordon of Cornell Engineering wrote in his proposal:

The discovery that free electrons in the Earth’s ionosphere incoherently scatter signals that are weak but detectable [...] makes possible the exploration of the upper atmosphere [...] that the radar components [...] are all within the state of the art means that the exploration can begin as soon as the radar is assembled. From "Design study of a radar to explore the earth’s ionosphere and surrounding space" by W. E. Gordon, H. G. Booker, & B. Nichols, Cornell.

Ward Low at DARPA realized the telescope could also be used to study the behavior of radio waves, and to intercept Soviet communications. The proposed atmospheric telescope would be built as a general-purpose radio telescope. The perfect location would have to be in the tropics, so that all planets passed over the telescope; a natural valley would save on construction costs ... and Braulio Dueno from the University of Puerto Rico at Mayaguez was studying at Cornell. Aricebo was chosen, and the telescope opened in 1964. The first big scientific result was the discovery that the rotational period of Mercury was 59 days, not 88 days as previously thought. For military intelligence, it detected Soviet radar waves reflected off the moon,
and gave the location of the radar station. For the history, see Daniel R. Altschuler, The National Astronomy and Ionosphere Center’s (NAIC) Arecibo Observatory in Puerto Rico, http://www.astro.wisc.edu/ sstanim/Students/daltschuler_2.pdf


‘At a conference in London in 1951 I had argued that dense, collapsed stars would be ideally suited to emit strong radio signals, since their magnetic fields may be enormously strengthened by the collapse and extend out into low density space.

‘The pulsars now seemed to represent just the stellar objects I had discussed then. Calculations existed for the collapsed “neutron stars” that indicated approximately their size, as small as a few kilometers, and their mass, on the order of a solar mass .4 Astronomers generally thought that even if they existed, they could never be discovered. However hot, a star so small could not be seen at astronomical distances. But they had not considered the energy concentration resulting from the collapse: enormous magnetic field strengths and a spin energy quite comparable with the entire content of nuclear energy of the star before its collapse. With these considerations it was not unreasonable to expect them to be observable.

‘I had another clue to the nature of the new objects: While there was some irregularity in the pulse-to-pulse timing, the long-term accuracy of these “clocks” was enormously better than a statistical addition of the pulse irregularities would have allowed.

‘I proposed the model of a rapidly spinning neutron star, which, as a result of some asymmetries, sent out a strong beam of radiation from one region of longitude. This beam would sweep over the Earth once in each rotation period and would be seen as a short pulse. The underlying long-term accuracy would then be produced by the spin of the object, while short-term timing fluctuations would just reflect details of the emitting mechanism. I compared it to the rotating beacon of a lighthouse, whose lamp, hanging from the rotating shaft, could wobble a little; the long-term accuracy would still be that of the rotation of the shaft.’


p81 This discussion of digestion follows Thomas Fischer Weiss’ Cellular Biophysics: Transport, MIT Press 1996.
A single molecule is an object in constant motion, spinning, contracting, expanding, etc; it doesn’t have any single size. Instead, quantum mechanics assigns a probability that the molecule has a given size. The idea of ‘molecular diameter’ is a substitute. There are many ways to do the computation; see for example ‘kinetic diameter’ at https://en.wikipedia.org/wiki/Kinetic_diameter; atomic radius at https://en.wikipedia.org/wiki/Atomic_radius; hard spheres models, http://www.kayelaby.npl.co.uk/general_physics/2_2/2_2_4.html, and more.
Chapter 1: Numbers

Section 8: Computer Numbers

This section is about how computers handle numbers. For many people, this is right up there with asking how variable-speed transmissions work in cars: they do their thing, and who cares how? It takes a bit of work to replace that complacency. One of the very early events in that work was Intel’s release of the Pentium P5 chip, Figure 80.

Figure 81 shows the problem: Nicely discovered that certain inputs of $x$ and $y$ would cause an error in the computation

$$x - \left( \frac{x}{y} \right) \cdot y = 0$$

As an example, for $x = 4195835$ and $y = 3145727$, the chip returned 256 instead of 0. We don’t think about a computer making errors like this – after all, it’s just a bunch of wires, which would be either always wrong or always right. That’s why we use computers. In reality, modern computer chips are very complex circuits, and they have to be programmed to do computations. As the programs are written by humans … there can be mistakes; see p95.

Even though errors like this are inevitable (in 2018 the problem was with SPECTRE and MELTDOWN errors), these aren’t the main issue. Since 1994, computer and chip manufacturers have systems in place to recall and replace defective units (Intel went ahead and did this in 1994).

There’s another level of problem – defective software. It’s estimated that the cost of dealing with bugs ran to over one trillion US$ in 2016 alone. Engineers and scientists typically do not have the same kinds of losses, but there can be losses to reputation for scientists, or lives in engineering projects. For example, the scientists C.L.Reyes and G. Chang retracted their paper *Structure of the ABC transporter MsbA in complex with ADP:vanadate and lipopolysaccharide*, Science, May 13, 2005. The structure was analyzed with a computer program, and an error in ± led to a structure where some portions of the molecule were backwards; see Figure 82. The structure in question was a transporter molecule: it allows harmful bacteria to defend themselves against drug treatment, by transporting the drugs out through the
bacterium cell wall. Understanding the mechanism used by ABC transporters might lead to techniques to disable it and eliminate resistance to antibiotics; see p95.

What went wrong? *Structure of MsbA from E. coli: A Homolog of the Multidrug Resistance ATP Binding Cassette (ABC) Transporters* Science 7 Sept 2001 vol 293, discusses how the structure of this molecule was determined:

Iterative eightfold noncrystallographic symmetry averaging, solvent flattening, flipping, phase extension, and amplitude sharpening using in-house programs yielded electron density maps of excellent quality for tracing a polypeptide chain.

This involves intensive computer manipulation of the data, and, as it happened, one of the programs used had an error.

In lab science, errors in procedure are to be expected. A scientist will spend years or decades learning correct lab procedure; what is different here is that scientists typically do not have decades of experience in computer programming. Here, a program was imported from another lab, but, Chang remarks, "you just trust the code to do the right job". He reports that he now triple checks everything (see Zeeya Merali, *Nature* 14 October 2010 vol 467, p 775).

Figure 83 gives another example where defective software destroyed (literally) a scientific mission: the $245 million dollar Mars Climate Orbiter. Launched in 1998 to study the climate and atmosphere of Mars, the mission plan was to use onboard thruster rockets to position the spacecraft to enter Mars orbit. Instead, the thrusters sent the craft deep into the Martian atmosphere, where it burned up before crashing.

Here, the problem is more complex: a mission such as this involves thousands of individuals, across multiple corporations and government institutions. On the other hand, the same is true for many modern engineering projects: the construction of highways, skyscrapers, sports arenas, or air-traffic control systems, and even smartphones. Engineering firms taking on these kinds of tasks need to have quality control systems in place. Compare the discussion on p43.

For the Mars Orbiter failure, two studies of the failure were made: the first, NASA's internal study; a second in *Why the Mars Probe went off course*, IEEE Spectrum 1 Dec 1999. For details of the reports, see p95.

In these two examples, training and management systems can help reduce errors and losses. But there are some actual mathematical issues here.
One of the core issues is inconsistent use of units for the thrust of the orbiter’s rockets. The actual numbers generated differed by only a small amount. But two important points emerge:
i) A small initial error, over a million mile trajectory, can result in a large final error. Compared to the million mile trajectory, the final path was off by only a few kilometers – a tiny percentage, but large enough to destroy the orbiter.  

ii) Flight engineers made many course corrections; this meant that the small error was made many times. A small error repeated many times can result in a large final error.

We’ll see these two issues again.

There’s a third major issue in using computers: operator error, illustrated in Figure 84. This is an example of a CAC or a computer aided catastrophe. As shown in Figure 85, the oil platform was intended to rest on the sea floor, meaning that the long beams supporting the platform had to carry not only the weight of the platform, but also deal with water pressure and storms. As the supports were lowered to the sea floor, one cracked, causing immediate flooding, dragging the platform deeper, where buoyancy tanks imploded under pressure; see https://wikivisually.com/wiki/Sleipner_A.

The problem was traced to the design of the supports. Large complex structures acted on by many forces are typically designed using computer tools. The standard mathematical technique is called finite element analysis; in this technique, a complex structure is analyzed as smaller parts interacting with each other through forces. It’s then possible to see how a force on one part of the structure flows through the other parts. NASA developed a program, NASTRAN, used in modeling forces on multi-stage rockets, which is famously successful.

In the Slepnir CAC, the designer did not fully understand how to implement the elements of the structure, which resulted in a design that underestimated forces by almost 50%. The Slepnir designer is not alone. Ivo Babuska, a professor of civil engineering at the University of Texas at Austin, tried a small experiment: he sent an engineering problem to a group of engineers. The problem is the Girkman problem, illustrated by the structure in Figure 86: how are the massive concrete walls to be supported under their own weight and wind forces?

Babuska specified that the engineers had to use professional-level computer software to solve the problem; of the 15 licensed professional engineers who submitted a solution, half were wrong. They
gave us numbers which were completely wrong and they believed in them” said Babuska. He commented, “How is it possible that this happened is a good question. There could be many various reasons. Nevertheless, in this case the reason was only one. Some of the analysts did not have sufficient engineering intuition and mathematical and engineering knowledge and possibly used the software incorrectly.”

There’s a – prejudice, or myth, that computers always give accurate answers. Babuska phrases the issue as ‘signing the blueprints.’ Only humans can sign blueprints and take financial and legal responsibility for structures. Operator error is a more difficult problem to overcome, as it depends not on project management, or program design, but on the integrity of users.

All of the above examples of computer fail are really examples of human failure. Is the myth really true, that computers don’t make errors? What we’ll see in the following is that indeed, using computers can lead directly to certain kinds of errors. These kind of errors are a consequence of how computers store numbers; they can only be avoided by understanding the issue, and planning against it – much like Babuska’s “mathematical and engineering knowledge.” This will be the take-away: serious users of computers must develop an intuition about their behavior.

We’ll start with an example:

\[ \lim_{x \to 0^+} \frac{1 - \cos(x)}{x^2} \]

With a calculator: we let \( x \) approach 0+, by taking \( x = .1, .01, .001 \) etc, or, \( x = 10^{-1}, 10^{-2}, 10^{-3}, \ldots \). Next, recall \( \lim_{x \to c} f(x) = L \) means the more decimal places \( x \) and \( c \) agree, the more decimal places \( f(x) \) and \( L \) agree. So, with a Texas Instruments TI-85 calculator:

<table>
<thead>
<tr>
<th>( x )</th>
<th>( (1 - \cos x)/x^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>.1</td>
<td>0.45970</td>
</tr>
<tr>
<td>.01</td>
<td>0.50000</td>
</tr>
<tr>
<td>.001</td>
<td>0.50000</td>
</tr>
<tr>
<td>( 10^{-5} )</td>
<td>0.50000</td>
</tr>
<tr>
<td>( 10^{-6} )</td>
<td>0.50004</td>
</tr>
<tr>
<td>( 10^{-7} )</td>
<td>0.49960</td>
</tr>
<tr>
<td>( 10^{-8} )</td>
<td>0</td>
</tr>
<tr>
<td>( 10^{-9} )</td>
<td>0</td>
</tr>
</tbody>
</table>
We believe that, the closer $x = 10^{-n}$ is to $0^+$, the better answer we’ll get. Therefore, the correct answer is the zero, not the .5.

It’s a kind of a suspicious answer, because of the sudden jump from 0.50000 to zero. Let’s use our mathematical knowledge and intuition to check it in another way, with L’Hospital’s rule:

$$\lim_{x \to 0^+} \frac{1 - \cos(x)}{x^2} = \lim_{x \to 0^+} \frac{[1 - \cos(x)]'}{[x^2]'} = \lim_{x \to 0^+} \frac{\sin(x)}{2x} = \lim_{x \to 0^+} \frac{[\sin(x)]'}{[2x]'} = \lim_{x \to 0^+} \frac{\cos(x)}{2} = \frac{1}{2}$$

On the whole, I’d go with L’Hospital’s answer, because that’s a well-known theorem that’s been around for centuries, with no reported errors. But then – that means something very strange must be happening inside the calculator. And to sort this out, we’re going to have to go inside the calculator.

Computers and calculators are electronic devices – they work by shifting electrical charges around. Just as batteries can store and release charge, devices called capacitors (Figure 87) can store and release charges quickly enough for modern computers.

We’ll conceptualize a computer number as a row of capacitors, and the charge in each capacitor represents a number; it would look something like the Heng/Zong system from p13, Figure 89. This picture shows the problem: there are only five boxes to store digits, so numbers larger than 99999 or smaller than .00001 can’t be written. We can add more boxes, but in a computer, there’s a limit; this limit is referred to as the word size of the computer.

There are other limitations. First, there’s an issue of making the best use of the small word size. For example, if I wrote .000003 as $3 \times 10^{-6}$ I could store it as 3, –, 6. All I need to know is the third place stores the exponent, the second the sign of the exponent, and the first the actual number.

While we can imagine our capacitors storing ten different levels of charge to represent the numbers 0, 1, . . . 9, in fact subtle variations in charge are very hard to detect. Modern computers use only two levels, which are traditionally thought of as the numbers one and zero (though in fact the two charges are more like ±.5). Then a number like 1001 could represent $1 \cdot 2^3 + 0 \cdot 2^2 + 0 \cdot 2^1 + 1 \cdot 2^0$, or 81. That is, we have to use binary arithmetic.

We also have to have standards, to know how to write numbers. Usually, to subtract 5.42 from 21.06, we’d align the numbers:

\[
\begin{array}{c}
2 & 1 & 0 & 6 \\
5 & 4 & 2 \\
1 & 5 & 6 & 4
\end{array}
\]
This is called fixed point arithmetic, and was used in the earliest computers (see p 96). Of course, we could use many other kinds of representation: 21.06 would be written as \( 0.2106 \times 10^2 \), or, in scientific notation, \( 2.106 \times 10^1 \). In the 1960’s, when computers began to be used in business (see Figure 90), standards for storing numbers varied. Each representation had it’s own problems and associated errors; multiplication by \( 1.0 \) could cause loss of the last four decimal places; programmers would use tricks such as replacing \( x \) by \( (x+x) - x \) to fool the computer into getting the answer right. Programming each individual computer was a craft in its own right, but as long as manufacturers like IBM kept one standard, programmers could adjust, and companies paid for programmer-craftsmen. See Charles Severance’s interview with ‘The Old Man of Floating Point’, William Kahane, at https://people.eecs.berkeley.edu/wkahan/ieee754status/754story.html.

To see how we can get into trouble, we’ll invent a silly machine, the Kathytron 5, Figure 91. The machine uses floating point arithmetic: for the number 21.06 the decimal point floats to the front, to give \( .2106 \times 10^2 \). With the bit structure described, the Kathytron has machine numbers of the form \( \pm b_1 b_2 \times 2^{\pm e_1} \), where \( b_1, b_2, e_1 \) can be 0, 1. These numbers are:

<table>
<thead>
<tr>
<th>(b_1 b_2)</th>
<th>( \times 2^{-1} )</th>
<th>( \times 2^{0} )</th>
<th>( \times 2^{1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>.01</td>
<td>.01</td>
<td>.01</td>
<td>.01</td>
</tr>
<tr>
<td>.10</td>
<td>.10</td>
<td>.10</td>
<td>.10</td>
</tr>
<tr>
<td>.11</td>
<td>.11</td>
<td>.11</td>
<td>.11</td>
</tr>
</tbody>
</table>

As a fraction, \( .11 = \frac{1}{2^2} + \frac{1}{2^4} = \frac{1}{2} + \frac{1}{4} = \frac{3}{4} \) and similarly for the others:

<table>
<thead>
<tr>
<th>(b_1 b_2)</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>.1</td>
<td>.5</td>
<td>.5</td>
<td>.1</td>
</tr>
<tr>
<td>.2</td>
<td>.3</td>
<td>.3</td>
<td>.2</td>
</tr>
<tr>
<td>.3</td>
<td>.6</td>
<td>.6</td>
<td>.3</td>
</tr>
<tr>
<td>.4</td>
<td>.8</td>
<td>.8</td>
<td>.4</td>
</tr>
<tr>
<td>.5</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>.6</td>
<td>1.5</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>.7</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

What we get, in decimals, are 0, .125, .375, .5, .75, 1, 1.5.
On a number line, the positive machine numbers look like this:

These numbers are not equally spaced, and, although the Kathytron 5 is a small machine, this unevenness is characteristic of this kind of standard; we’ll talk about that when we discuss errors in approximating numbers by machine numbers. Another feature is that there’s a smallest non-zero number the machine can represent, called machine epsilon. In many programming languages, you can access machine epsilon by running the command eps; on the Kathytron, you’d get eps=.125 This book is being written on a MacBook Air running the public domain language Octave; there eps = $2.2204 \times 10^{-16}$, which will give rather more accurate computations than the Kathytron 5.

The Kathytron rounds a number smaller than machine epsilon to zero. A number larger than the largest machine number terminates the computation with an error message. This inequality in treatment can cause problems.

Now, we’re not finished. The Kathytron saves one bit by using normalized floating point notation: the initial number can’t be a zero; our Kathytron 5 numbers are now $\pm.1 \times 2^{\pm e}$ In theory this allows for more numbers – but there’s a catch. Look at the second row, .01 $\times 2^b$: .01 $\times 2^{-1}$ .01 $\times 2^0$.01 $\times 2^1$. When I normalize those, I get $\pm 2^{b-1}$:.10 $\times 2^{-2}$ .10 $\times 2^{-1}$ .10 $\times 2^0$. But the exponent $2^{-2}$ exceeds the allowed exponents (-1, 0, 1), so .10$\times 2^{-2}$ is not a Kathytron number; it rounds to zero, and my table now looks like this:

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1/4</td>
<td>1/2</td>
<td></td>
</tr>
<tr>
<td>1/3</td>
<td>1/3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1/2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Normalized Kathytron numbers are 0, .375, .5, .75, 1, 1.5; the .125 is missing, and machine epsilon is now .375. The numbers are still not equally spaced, especially the distance between zero and machine epsilon.
The picture below shows the new number line:

This is called the \textit{gap at zero}, and occurs in machines that use a similar standard. And it is the cause of the original problem: when I subtract the machine numbers .5 and .375, I get .125, which is smaller than machine epsilon. This machine rounds this down to zero.

This is what went wrong when I computed \( \lim_{x \to 0^+} \frac{1-\cos(x)}{x^2} \). When \( x \) was very close to zero, the subtraction \( 1 - \cos(x) \) caused the leading terms in the decimal to cancel, and the decimal places that remained were less than machine epsilon. Without notice, the machine rounded down to zero, and my table has a sudden jump from .5 to 0, where \( 1 - \cos(x) \) becomes less than machine epsilon.

You can see this is going to be a problem: the user has to know this can happen, and plan against it. Here’s where we have programmers-craftsmen. The trick is to eliminate possible subtractions where many leading terms in a decimal will cancel. There’s a rule of thumb: ‘Don’t subtract nearly equal numbers.’ That’s one of our ‘mathematical intuitions’.

Here’s an example of a trick of the craft, for \( 1 - \cos(x) \):

\[
\frac{1 - \cos(x)}{x^2} = \frac{1 - \cos(x)}{x^2} \left( \frac{1 + \cos(x)}{1 + \cos(x)} \right) = \frac{\sin^2(x)}{x^2} \left( \frac{1}{1 + \cos(x)} \right)
\]

Back to the TI-85 calculator:

<table>
<thead>
<tr>
<th>( x )</th>
<th>( \frac{\sin^2(x)}{x^2} \left( \frac{1}{1 + \cos(x)} \right) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.49958</td>
</tr>
<tr>
<td>0.01</td>
<td>0.49999</td>
</tr>
<tr>
<td>0.001</td>
<td>0.50000</td>
</tr>
<tr>
<td>10^{-5}</td>
<td>0.50000</td>
</tr>
<tr>
<td>10^{-6}</td>
<td>0.50000</td>
</tr>
<tr>
<td>10^{-7}</td>
<td>0.50000</td>
</tr>
<tr>
<td>10^{-8}</td>
<td>0.50000</td>
</tr>
<tr>
<td>10^{-9}</td>
<td>0.50000</td>
</tr>
</tbody>
</table>
This is one example of unexpected rounding, down to zero. There's an additional problem, that of repeated rounding of computations. Early thinkers believed that repeated rounding would create such large errors that number systems like those in the Kathytron would be useless. (to be continued)

When computers became available for medical, scientific and engineering research, the lack of a single standard became an issue. Software written for one computer would not run properly on other computers. Estimates of the accuracy of a computation were only good on a specific computer, and could be unreliable on others.

But the real motivation for setting a standard came from the development of microcomputers. IBM might sell a few hundred machines; Cray a dozen, but a chip designer like Intel expected to sell millions. Moreover, the users would not be large businesses with expensive craftsmen, but small businesses with perhaps no programmers, but who still expected correct answers.

A group of scientists began meeting to resolve the problem. Intel, and Motorola, wanted not just a standard, but the best possible standard – that didn’t slow down their chips too much. In 1974, the IEEE 794 standard for representing numbers on a computer came out. (in reality it’s a set of standards; see again p96). In fact, it wasn’t a standard, but a set of suggestions, but as all the large chip-makers followed it, it became the standard by default.
Notes for Chapter 1 Section 8: Computer Numbers

p86 The P5 problem arose because of the way computers do division. In a way, it’s very much like we do division. Try dividing 376 by 7. We have a technique – an algorithm. We’d first divide 7 into 37. We remember that $7 \times 5 = 35$ but $7 \times 6 = 42$. So the divisor is 5, with a remainder of 26. Now divide 7 into 26. Again we remember that $7 \times 3 = 21$ but $7 \times 4 = 28$. So the divisor is 3, with a remainder of 5.

When I was younger – a lot younger – I’d look up some of those multiplications, in a multiplication table. The P5 used a division algorithm called the SRT algorithm, named after the inventors, Sweeney, Robertson, and Tocher. It also uses a lookup table, though everything is binary, so it seems rather strange. It’s similar to Figure 93. See An analysis of division and implementations by Stuart F Oberman and Michael J Flynn at http://i.stanford.edu/pub/cstr/reports/csl/tr/95/675/CSL-TR-95-675.pdf

p87 For the importance of the ABC transporters, see Christopher F. Higgins and Kenneth J. Linton, The xyz of ABC Transporters, Science 7 Sept 2001 vol 293 p1782. They remark,

A cell must selectively translocate molecules across its plasma membrane to maintain the composition of its cytoplasm distinct from that of the surrounding milieu. The most intriguing, and, arguably, the most important membrane proteins for this purpose are the ABC (ATP-binding cassette) transporters. These proteins, found in all species, use the energy of ATP hydrolysis to translocate specific substrates across cellular membranes.

Overexpression of certain ABC transporters is the most frequent cause of resistance to cytotoxic agents including antibiotics, antifungals, herbicides, and anticancer drugs.

p87 In the analyses of the failure of the Mars Orbiter Mission, NASA identified the problem as a software issue: a contractor wrote a software module to determine the thrust of the orbiters rockets. The module gave thrust in $ft/sec^2$; the software to which it reported expected force to be reported in Newtons. NASA blamed its own internal procedures for not finding the units mismatch.

The IEEE Spectrum analysis asserts that the problems in NASA went much deeper: flight controllers believed that something was wrong with the trajectory, but were over-ruled by higher management. The article claims that the engineers were told to “stop thinking like engineers and think like managers.” They claim that the expectation was that all systems worked perfectly, and if problems were suspected, it was up to the engineers to prove there was a problem. This is the opposite of standards for airline safety.
For a brief history, see Michael L. Overton *Numerical Computing with IEEE Floating Point Arithmetic*, SIAM Press 2001. Overton notes: "Von Neumann […] promoted the use of fixed point representation. He was well aware that the range limitations of fixed point would be too severe to be practical, but he believed that the necessary scaling by a power of 2 should be done by the programmer, not the machine; he argued that bits were too precious to be wasted on storing an exponent when they could be used to extend the precision of the significand." See Figure 94, von Neuman with the 1949 EDVAC computer, using both binary arithmetic and stored programs.
List of Figures

Chapter 0

p6 Figure 2 ECG machine picture from http://en.wikipedia.org/wiki/Image:Willem_Einthoven_ECG.jpg.

p6 Figure 3 The first human ECG. From Einthoven W. The galvanometric registration of the human electrocardiogram. Pfluger’s Arch f.d. ges Physiol 1903;99:472D480.

p7 Figure 4 Picture of ECG indicating irregular heartbeat, from http://www.bem.fi/book/19/19.htm.

p7 Figure 5 Picture of portable ECG unit from http://www.skymall.com/handheld-ecg-monitor/HDW107.html

p7 Figure 6 is from Alfonso, Valentino X ECG QRS Detection, in Tompkins, Ed. Biomedical Digital Signal Processing, C-Language Examples and Laboratory Experiments for the IBM PC, Prentice Hall PTR, 1995.


p8 The drawing linking the regions of the heart with their electrical signature is from Frank H. Netter, Heart (CIBA Collection of Medical Illustrations, Volume 5, CIBA 1978.

Chapter 1

p10 Figure 7 from Simon Hunt, Jason Low and K. C. Burns, Adaptive numerical competency in a food-hoarding songbird, Proc. R. Soc. B (2008) 275, 2373D2379.

p11 Figure 10 is from http://www.gregapodaca.com/numerography/files/o20.html

p11 Figure 9 is from http://afrolegends.com/2013/08/29/the-ishango-bone-craddle-of-mathematics/

p12 The image in Figure 11 is by Ben M. Harvey, Utrecht University.

p12 Figure 12 is from T. Cuyler Young, Jr., of the Royal Ontario Museum, Toronto, Canada.

p12 Figure 13 is licensed under Public Domain via Commons https://commons.wikimedia.org/wiki/File:Zero_in_Rod_Calculus.png#/media/File:Zero_in_Rod_Calculus.png
Hobbes; engraving by Abraham Bosse, public domain.

Figure 36 A photo of Wittgenstein is by Moritz Nühr, an Austrian photographer also known for his portraits of the composer Gustav Mahler and the artist Gustav Klimt. From the archives of the Austrian National Library, and is public domain.

Figure 37 A paramecium is from https://commons.wikimedia.org/wiki/File:Paramecium.jpg, by Barfooz from the English Wikipedia. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License.

Figure 38 A coloring of a map is from https://commons.wikimedia.org/wiki/File:Four_Colour_Map_Example.svg based on a raster image by chas zzz brown on en.wikipedia. Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License.

Figure 39 A photo of Ahmes is from the University College, London, http://www.ucl.ac.uk/museums-static/digitalegypt/lahun/uc32159.html

Figure 40 A photo of Lui Hui is from http://www.slideshare.net/sbrierton/liuhui1

Figure 41 Lui Hui’s construction for \(\pi\) is from CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=4322998

Figure 42 A painting of Euclid from Galleria Nazionale delle Marche, Italy; public domain. This picture is meant to represent the famous mathematician Euclid of Alexandria, who was, in medieval times, wrongly identified with Euclid of Megara, the disciple of Socrates.

Figure 43 A photo of Ashurbannipal is from https://commons.wikimedia.org/wiki/File:Table_of_Shamash_relief.jpg licensed under Creative Commons.

Figure 44 The original tablet of the Pythagorean theorem is from the personal collection of Bill Casselman; see http://www.math.ubc.ca/~cass/Euclid/ycb/ycb.html. The original tablet is in the Yale Babylonian Collection; see http://nelc.yale.edu

Figure 45 A photo of diffraction rings is from https://commons.wikimedia.org/wiki/File:Laser_Interference.JPG


Figure 47 Illustrating surreal numbers is by Lukas Lansky

Figure 48 A photo of John Conway is from Princeton University Office of Communications, by Denise Applewhite
Figure 51 on Archimedes’ estimates for π is from http://mrstevesscience.blogspot.com/2011/03/help-6th-graders-figure-out-for.html

Figure 52 on a geometric proof of the binomial expansion is from http://www.mathaware.org/mam/00/master/essays/B3D/2/binomial.html

Figure 54 of Cauchy is from http://www-groups.dcs.st-and.ac.uk/history/PictDisplay/Cauchy.html

Figure 55 is from a portrait of Archimedes in the Museo del Prado, Spain. This is from https://www.cs.drexel.edu/~crorres/Archimedes/contents.html

Figure 57 of the Cray computer is from macobserver.com/tmo/article/The_Fastest_Mac_Compared_to_Todays_Supercomputers

Figure 58 of the Sunway Taihen Light is from http://www.gizmag.com/supercomputer-china-taihulight-processor-petaflops/43982/

Figure 59 of Lara Croft is from https://www.flickeringmyth.com/2015/06/e3-gameplay-trailer-for-rise-of-the-tomb-raider/rise-of-the-tomb-raider-3/

Figure 60 of the Xbox is from http://www.xbox.com/en-us/xbox-one-x

Figure 61, is from http://www.factknow.com/2016/03/some-of-greatest-discoveries-made-by.html

Figure 62 of Japanese money is from http://sputniknews.com/business/20150903/1026525213.html

Figure 63 of crystalline structure is taken from http://www.ebsd.com/introduction

Figure 64 Picture of child with money envelope from https://www.sassandbelle.co.uk/blog/2015/02/19

Figure 65 Picture of British length standards from The Metrology Handbook, American Society for Quality Measurement, by Bennett, Keith Bucher, Jay L., ed.

Figure 45 Image of Sun God from https://en.wikipedia.org/wiki/Measuring rod

Figure 67 Picture of MMDA tablets is copyrighted by Erowid.org. See https://erowid.org/chemicals/mdma/mdma.shtml

Figure 68 Gravity map from http://www.nasa.gov/vision/earth/lookingatearth/earth_drag.html

Figure 69 Mars Orbiter picture from NASA JPL.

Figure 70 Photo of Jocelyn Bell in 1967 from http://campus.pari.edu/radiosky/lessons/pulsars/02.shtml

Figure 71 Crab Nebula photo from http://imagesci.com/img/2013/11/crab-
nebula-1613-hd-wallpapers.jpg

p80 Figure 72 Diagram of airplane spar from https://www.quora.com/What-is-aircraft-wing-spar

p80 Figure 73 Launch Complex photo from http://www.hq.nasa.gov/office/pao/History/alsj/misc/apmisc-SAT-3-24.jpg This file is in the public domain in the United States because it was solely created by NASA.

p80 Figure 74 Deserted Launch Complex photo from http://zimmer.csufresno.edu/fring-wal/ksc.html

p81 Figure 75 Image of circular folds in the small intestine from https://s3.amazonaws.com/classconnection/699/flashcards/3329699/png/plicae-14C4570528C65150E72.png

p81 Figure 76 Sketch of villi from https://commons.wikimedia.org/wiki/File:Esquema_del_epitelio_del_intestino_delgado.png

p82 Figure 77 Sketch of single villus from http://biology4ibdp.weebly.com/61-digestion.html


Figure 80 CPU Intel Pentium A80501 with GoldCap, 66 MHz, Vcore = 5V, SX837 = with FDIV-Bug. Konstantin Lanzet - CPU Collection Konstantin Lanzet. Permission details This work is free and may be used by anyone for any purpose. If you wish to use this content, you do not need to request permission as long as you follow any licensing requirements mentioned on this page. Wikimedia has received an e-mail confirming that the copyright holder has approved publication under the terms mentioned on this page. This correspondence has been reviewed by an OTRS member and stored in our permission archive.

Figure 81 of the autographed Pentium chip is from https://www.viva64.com/en/b/0454/.

Figure 82 of the (erroneous) structure of an ABC transporter molecule is from https://bl831.als.lbl.gov/mcfuser/publications/New%20Journal%20Papers/Reyes_Science.pdf

Figure 83 of the Mars Orbiter flyby is from http://www.teoti.com/funny-stuff/130568-great-miscalculations-the-french-railway-error-and-10-others.html

Figure 85 of the Sleipnir project is from https://wikivisually.com/wiki/Sleipner_A

Figure 86 of the Sydney Opera House is from https://www.ices.utexas.edu/about/news/254/

Figure 87 of a modern CMOS imaging chip is from https://www.dpreview.com/news/0671207908/canon-is-now-selling-cmos-image-sensors-including-a-120mp-aps-h-beast

Figure 88, schematic view of one imaging cell in a CMOS sensor is from https://micro.magnet.fsu.edu/primer/digitalimaging/cmosimagesensors.html

Figure 90 of the IBM System/360 computer is from https://www.theregister.co.uk/2014/04/07/ibm_s_360_50_anniversary/

Figure 93 of a lookup table for performing computer arithmetic is from http://i.stanford.edu/pub/cstr/reports/csl/tr/95/675/CSL-TR-95-675.pdf

Figure 94 of the mathematician John von Neumann is from https://www.eldiario.es/turing/John-Neumann-revolucionando-computacion-Manhattan_0_380412943.html
Index

A
abacus, 13
ABC molecules, 87
abstraction, 18
accounting, 11, 39, 44
Achebe, Chinua, 76
aerodynamics, 66
Al-Andalus, 13
algebraic numbers, 48, 58
algorithm, 13
anthropologists, 11
approximation, 40
Archimedean Principle, 29, 42, 62, 71
Archimedes, 52
Aricebo, 78, 83
Aristotle, 18, 25–27, 31, 32, 67, 70, 71
arithmetisation, 54, 67
Armstrong, Karen, 19
astrology, 15
astronomy, 78
atom, 69
Avial Age, 19
axiom, 25, 27, 44

B
base sixty, 11
Babuska, Ivar, 88
Babylonian, 44, 73
Babylonian Astronomy, 15
Baghdad, 13
Babushal Manuscript, 13
beer, 14, 46
Behaviorism, 20, 22
Bell, Jocelyn, 78
Berlin, Brent, 76
Bessel function, 48
Bethe, Hans, 78
binomial expansion, 52

C
Bitcoins, 66
Boole, George, 32
Buddha, Gautama, 19
CAC, 88
Calculus, 89, 93
Calendar, 17
calender, 14
Caliphate, 13
Cantor, 54, 58–60, 71
Cantor, Georg, 38, 59
carbohydrates, 82
carcinogenic, 80
Cauchy, 59
Cauchy Sequence, 55
Cauchy, Augustin, 54
CFIT, 75, 87
characteristic, 79
Chinese, 12, 42, 66
Chomsky, Noam, 20
circular folds, intestinal, 81
cognitive reference points, 38
Cognitive Science, 38
completeness, 33, 57, 58
complex numbers, 58
computable, 58, 59
conditioned behavior, 20
consistency, 33
costellations, 14
cognitiveist history, 41
continued fractions, 60
Continuity, 58, 70
Conway, John, 50
cortisol, 36
counting, 27
Crab Nebula, 78
Cray, 66
cuneiform, 15
cut, Dedekind, 70

D
Dead Sea Scroll, 17
Dedekind, Richard, 59, 70, 71
deduction, 25
degrees, 15
Dehaene, Stanislaw, 21, 23
density of rationals, 57
diffraction, 48
disease, genetic, 66
divination, 15
Drake, Frank, 78
drug design, 66
drug resistance, 87
Dumbledore, 35

E
Easter, 17, 72
ECG, 6, 8
EDVAC, 96
Egypt, 39
eighteen-wheelers, 66
Einthoven, 6
Elements of Euclid, 30
Elements, Euclid’s, 26
Engineering, 13
Enlightenment, 74
Enuma Anu Enlil, 15
environmental remediation, 80
epistemology, 17, 30
epsilon, 41
equivalence, 37
equivalence class, 37, 59
error, 40
Essene, 17
Euclid, 26, 30–32, 51, 52, 60
Eudoxus, 47
Euler, 48
Parmenides, 18
Passover, 72
Peano axioms, 28
Peano, Giuseppe, 27, 67
pendulum, 74
Pentium, 95
Pentium chip, 86
photosynthesis, 65
Pinker, Steven, 35
Planck, 69
Plato, 17, 18, 25, 30, 32
Prior Analytics, 26, 31
protease, 82
proto-Indo-European, 25
ProtoWorld, 23
Puerto Rico, 83
pulsar, 78
Pythagoras, 18, 19, 25, 52
Pythagorean Theorem, 45

Q
Qin, 12
quadrivium, 32
quantum mechanics, 85
Quran, 72

R
radar, 83
Ramadan, 72
rational numbers, 52
real line, 67
real numbers, 56
reasoning, deductive, 25
relativism, 76
remediation, 80
representation, 11
representation, 11, 35
Robin, 10
Robson, Eleanor, 41
Roman numerals, 12
Rosch, Eleanor, 38, 76
rounding, 64
ruler and compass, 52
Russell, Bertrand, 33

S
Salt peter, Edwin, 78
Schmandt-Besserat, Denise, 23
scientific notation, 65
Seleucid, 44
sequence, 43, 44, 55
Shamash, 45, 73
Shannon, Claude, 34
Shining One, 41
Skinner, Verbal Behavior, 20
Soviet, 83
Space, 45
Spain, 13
spar, of a wing, 80
SPECTRE, 86
square-cube, 80
SRT algorithm, 95
stock market trading, 65
stored program, 96
strong induction, 28, 38, 51
surface area of sphere, 80
surreal numbers, 50
surveying, 73
sylllogism, 26, 31
symbolic, 11
synapse, 82
System A and B, 15, 24

T
tax, 45
TCE, 80
teleology, 43
Time, 45
Tomkins, Willis, 97
transcendental functions, 48
transcendental numbers, 48
transfinite numbers, 38, 60
trivium, 32
Turing machine, 34
Turing, Alan, 34
Tyler, G. Leonard, 78

V
Vaieika, 69
villi, 81
volume of a sphere, 80
von Neumann, John, 96

W
warming, global, 65
weapons, nuclear, 66
weather forecasting, 66
Whitehead, Alfred North, 33
Wigner, Eugene, 16, 18
Wittgenstein, Ludwig, 35
word size, computer, 90

Y
Yahad, 17
Yaweh, 73