Preface

There's a secret world out there. A hidden parallel universe of beauty and elegance, intricately intertwined with ours. It's the world of mathematics. And it's invisible to most of us. This book is an invitation to discover this world.

Consider this paradox: On the one hand, mathematics is woven in the very fabric of our daily lives. Every time we make an online purchase, send a text message, do a search on the Internet, or use a GPS device, mathematical formulas and algorithms are at play. On the other hand, most people are daunted by math. It has become, in the words of poet Hans Magnus Enzensberger, "a blind spot in our culture – alien territory, in which only the elite, the initiated few have managed to entrench themselves." It's rare, he says, that we "encounter a person who asserts vehemently that the mere thought of reading a novel, or looking at a picture, or seeing a movie causes him insufferable torment," but "sensible, educated people" often say "with a remarkable blend of defiance and pride" that math is "pure torture" or a "nightmare" that "turns them off."

How is this anomaly possible? I see two main reasons. First, mathematics is more abstract than other subjects, hence not as accessible. Second, what we study in school is only a tiny part of math, much of it established more than a millennium ago. Mathematics has advanced tremendously since then, but the treasures of modern math have been kept hidden from most of us.

What if at school you had to take an "art class" in which you were only taught how to paint a fence? What if you were never shown the paintings of Leonardo da Vinci and Picasso? Would that make you appreciate art? Would you want to learn more about it? I doubt it. You would probably say something like this: "Learning art at school was a waste of my time. If I ever need to have my fence painted, I'll just hire people to do this for me." Of course, this sounds ridiculous, but this is how math is taught, and so in the eyes of most of us it becomes the equivalent of watching paint dry. While the paintings of the great masters are readily available, the math of the great masters is locked away.

However, it's not just the aesthetic beauty of math that's captivating. As Galileo famously said, "The laws of Nature are written in the language of mathematics." Math is a way to describe reality and figure out how the world works, a universal language that has become the gold standard of truth. In our world, increasingly driven by science and technology, mathematics is becoming, ever more, the source of power, wealth, and progress. Hence those who are fluent in this new language will be on the cutting edge of progress.

One of the common misconceptions about mathematics is that it can only be used as a "toolkit": a biologist, say, would do some field work, collect data, and then try to build a mathematical model fitting these data (perhaps, with some help from a mathematician). While this is an important mode of operation, math offers us *a lot more*: it enables us to make groundbreaking, paradigm-shifting leaps that we couldn't make otherwise.

For example, Albert Einstein was not trying to fit any data into equations when he understood that gravity causes our space to curve. In fact, there was no such data. No one could even imagine at the time that our space is curved; everyone "knew" that our world was flat! But Einstein understood that this was the only way to generalize his special relativity theory to non-inertial systems, coupled with his insight that gravity and acceleration have the same effect. This was a high-level intellectual exercise within the realm of math, one in which Einstein relied on the work of a mathematician, Bernhard Riemann, completed fifty years earlier. The human brain is wired in such a way that we simply cannot imagine curved spaces of dimension greater than two; we can only access them through mathematics. And guess what, Einstein was right – our universe *is* curved, and furthermore, it's expanding. That's the power of mathematics I am talking about!

Many examples like this may be found, and not only in physics, but in other areas of science (we will discuss some of them below). History shows that science and technology are transformed by mathematical ideas at an accelerated pace; even mathematical theories that are initially viewed as abstract and esoteric later become indispensable for applications. Charles Darwin, whose work at first did not rely on math, later wrote in his autobiography: "I have deeply regretted that I did not proceed far enough at least to understand something of the great leading principles of mathematics, for men thus endowed seem to have an extra sense." I take it as prescient advice to the next generations to capitalize on mathematics' immense potential.

When I was growing up, I wasn't aware of the hidden world of mathematics. Like most people, I thought math was a stale, boring subject. But I was lucky: in my last year of high school I met a professional mathematician who opened the magical world of math to me. I learned that mathematics is full of infinite possibilities as well as elegance and beauty, just like poetry, art, and music. I fell in love with math.

Dear reader, with this book I want to do for you what my teachers and mentors did for me: unlock the power and beauty of mathematics, and enable *you* to enter this magical world the way I did, even if you are the sort of person who has never used the words "math" and "love" in the same sentence. Mathematics will get under your skin just like it did under mine, and your worldview will never be the same.

* * *

Mathematical knowledge is unlike any other knowledge. While our perception of the physical world can always be distorted, our perception of mathematical truths can't be. They are objective, persistent, necessary truths. A mathematical formula or theorem means the same thing to anyone anywhere – no matter what gender, religion, or skin color; it will mean the same thing to anyone a thousand years from now. And what's also amazing is that we own all of them. No one can patent a mathematical formula, it's ours to share. There is nothing in this world that is so deep and exquisite and yet so readily available to all. That such a reservoir of knowledge really exists is nearly unbelievable. It's too precious to be given away to the "initiated few." It belongs to all of us.

One of the key functions of mathematics is the ordering of information. This is what distinguishes the brush strokes of Van Gogh from a mere blob of paint. With the advent of 3D printing, the reality we are used to is undergoing a radical transformation: everything is migrating from the sphere of physical objects to the sphere of information and data. We will soon be able to convert information into matter on demand by using 3D printers just as easily as we now convert a PDF file into a book or an MP3 file into a piece of music. In this brave new world, the role of mathematics will become even more central: as the way to organize and order information, and as the means to facilitate the conversion of information into physical reality.

In this book, I will describe one of the biggest ideas to come out of mathematics in the last fifty years: the Langlands Program, considered by many as the Grand Unified Theory of mathematics. It's a fascinating theory that weaves a web of tantalizing connections between mathematical fields that at first glance seem to be light years apart: algebra, geometry, number theory, analysis, and quantum physics. If we think of those fields as continents in the hidden world of mathematics, then the Langlands Program is the ultimate teleportation device, capable of getting us instantly from one of them to another, and back.

Launched in the late 1960s by Robert Langlands, the mathematician who currently occupies Albert Einstein's office at the Institute for Advanced Study in Princeton, the Langlands Program had its roots in a groundbreaking mathematical theory of symmetry. Its foundations were laid two centuries ago by a French prodigy, just before he was killed in a duel, at age twenty. It was subsequently enriched by another stunning discovery, which not only led to the proof of Fermat's Last Theorem, but revolutionized the way we think about numbers and equations. Yet another penetrating insight was that mathematics has its own Rosetta stone and is full of mysterious analogies and metaphors. Following these analogies as creeks in the enchanted land of math, the ideas of the Langlands Program spilled into the realms of geometry and quantum physics, creating order and harmony out of seeming chaos.

I want to tell you about all this to expose the sides of mathematics we rarely get to see: inspiration, profound ideas, startling revelations. Mathematics is a way to break the barriers of the conventional, an expression of unbounded imagination in the search for truth. Georg Cantor, creator of the theory of infinity, wrote: "The essence of mathematics lies in its freedom." Mathematics teaches us to rigorously analyze reality, study the facts, follow them wherever they lead. It liberates us from dogmas and prejudice, nurtures the capacity for innovation. It thus provides tools that transcend the subject itself.

These tools can be used for good and for ill, forcing us to reckon with math's real-world effects. For example, the global economic crisis was caused to a large extent by the widespread use of inadequate mathematical models in the financial markets. Many of the decision makers didn't fully understand these models due to their mathematical illiteracy, but were arrogantly using them anyway – driven by greed – until this practice almost wrecked the entire system. They were taking unfair advantage of the asymmetric access to information and hoping that no one would call their bluff because others weren't inclined to ask how these mathematical models worked either. Perhaps, if more people understood how these models functioned, how the system really worked, we wouldn't have been fooled for so long.

As another example, consider this: in 1996, a commission appointed by the U.S. government gathered in secret and altered a formula for the Consumer Price Index, the measure of inflation that determines the tax brackets, Social Security, Medicare, and other indexed payments. Tens of millions of Americans were affected, but there was little public discussion of the new formula and its consequences. And recently there was another attempt to exploit this arcane formula as a backdoor on the U.S. economy.[1]

Far fewer of these sorts of backroom deals could be made in a mathematically literate society. Mathematics equals rigor plus intellectual integrity times reliance on facts. We should all have access to the mathematical knowledge and tools needed to protect us from arbitrary decisions made by the powerful few in an increasingly math-driven world. Where there is no mathematics, there is no freedom.

* * *

Mathematics is as much part of our cultural heritage as art, literature, and music. As humans, we have a hunger to discover something new, reach new meaning, understand better the universe and our place in it. Alas, we can't discover a new continent like Columbus or be the first to set foot on the Moon. But what if I told you that you don't have to sail across an ocean or fly into space to discover the wonders of the world? They are right here, intertwined with our present reality. In a sense, within us. Mathematics directs the flow of the universe, lurks behind its shapes and curves, holds the reins of everything from tiny atoms to the biggest stars.

This book is an invitation to this rich and dazzling world. I wrote it for readers without any background in mathematics. If you think that math is hard, that you won't get it, if you are terrified by math, but at the same time curious whether there is something there worth knowing – then this book is for you.

There is a common fallacy that one has to study mathematics for years to appreciate it. Some even think that most people have an innate learning disability when it comes to math. I disagree: most of us have heard of and have at least a rudimentary understanding of such concepts as the solar system, atoms and elementary particles, the double helix of DNA, and much more, without taking courses in physics and biology. And nobody is surprised that these sophisticated ideas are part of our culture, our collective consciousness. Likewise, everybody can grasp key mathematical concepts and ideas, if they are explained in the right way. To do this, it is not necessary to study math for years; in many cases, we can cut right to the point and jump over tedious steps.

The problem is: while the world at large is always talking about planets, atoms, and DNA, chances are no one has ever talked to you about the fascinating ideas of modern math, such as symmetry groups, novel numerical systems in which 2 and 2 isn't always 4, and beautiful geometric shapes like Riemann surfaces. It's like they keep showing you a little cat and telling you that this is what a tiger looks like. But actually the tiger is an entirely different animal. I'll show it to you in all of its splendor, and you'll be able to appreciate its "fearful symmetry," as William Blake eloquently said.

Don't get me wrong: reading this book won't by itself make you a mathematician. Nor am I advocating that everyone should become a mathematician. Think about it this way: learning a small number of chords will enable you to play quite a few songs on a guitar. It won't make you the world's best guitar player, but it will enrich your life. In this book I will show you the chords of modern math, which have been hidden from you. And I promise that this will enrich your life.

One of my teachers, the great Israel Gelfand, used to say: "People think they don't understand math, but it's all about how you explain it to them. If you ask a drunkard what number is larger, 2/3 or 3/5, he won't be able to tell you. But if you rephrase the question: what is better, 2 bottles of vodka for 3 people or 3 bottles of vodka for 5 people, he will tell you right away: 2 bottles for 3 people, of course."

My goal is to explain this stuff to you in terms that you will understand.

I will also talk about my experience of growing up in the former Soviet Union, where mathematics became an outpost of freedom in the face of an oppressive regime. I was denied entrance to Moscow State University because of the discriminatory policies of the Soviet Union. The doors were slammed shut in front of me. I was an outcast. But I didn't give up. I would sneak into the University to attend lectures and seminars. I would read math books on my own, sometimes late at night. And in the end, I was able to hack the system. They didn't let me in through the front door; I flew in through a window. When you are in love, who can stop you?

Two brilliant mathematicians took me under their wings and became my mentors. With their guidance, I started doing mathematical research. I was still a college student, but I was already pushing the boundaries of the unknown. This was the most exciting time of my life, and I did it even though I was sure that the discriminatory policies would never allow me to have a job as a mathematician in the Soviet Union.

But there was a surprise in store: my first mathematical papers were smuggled abroad and became known, and I got invited to Harvard University as a Visiting Professor at age twenty-one. Miraculously, at exactly the same time *perestroika* in the Soviet Union lifted the iron curtain, and citizens were allowed to travel abroad. So there I was, a Harvard professor without a Ph.D., hacking the system once again. I continued on my academic path, which led me to research on the frontiers of the Langlands Program and enabled me to participate in some of the major advances in this area during the last twenty years. In what follows, I will describe spectacular results obtained by brilliant scientists as well as what happened behind the scenes.

This book is also about love. Once, I had a vision of a mathematician discovering the "formula of love," and this became the premise of a film *Rites of Love and Math*, which I will talk about later in the book. Whenever I show the film, someone always asks: "Does a formula of love really exist?"

My response: "Every formula we discover is a formula of love." Mathematics is the source of timeless profound knowledge, which goes to the heart of all matter and unites us across cultures, continents, and centuries. My dream is that all of us will be able to see, appreciate, and marvel at the magic beauty and exquisite harmony of these ideas, formulas, and equations, for this will give so much more meaning to our love for this world and for each other.

A Guide for the Reader

I have made every effort to present mathematical concepts in this book in the most elementary and intuitive way. However, I realize that some parts of the book are somewhat heavier on math (particularly, some parts of Chapters 8, 14, 15, and 17). It is *perfectly fine to skip* those parts that look confusing or tedious at the first reading (this is what I often do myself). Coming back to those parts later, equipped with newly gained knowledge, you might find the material easier to follow. But that is usually not necessary in order to be able to follow what comes next.

Perhaps, a bigger point is that *it is perfectly OK if something is unclear*. That's how I feel 90 percent of the time when I do mathematics, so welcome to my world! The feeling of confusion (even frustration, sometimes) is an essential part of being a mathematician. But look at the bright side: how boring would life be if everything in it could be understood with little effort! What makes doing mathematics so exciting is our desire to overcome this confusion; to understand; to lift the veil on the unknown. And the feeling of personal triumph when we do understand something makes it all worthwhile.

My focus in this book is on the big picture and the logical connections between different concepts and different branches of math, not technical details. A more in-depth discussion is often relegated to the endnotes, which also contain references and suggestions for further reading. However, although endnotes may enhance your understanding, they may be safely skipped (at least, at the first reading).

I have tried to minimize the use of formulas – opting, whenever possible, for verbal explanations. Feel free to skip the few formulas that do appear.

A word of warning on mathematical terminology: while writing this book, I discovered, to my surprise, that certain terms that mathematicians use in a specific way actually mean something entirely different to non-mathematicians. Terms like correspondence, representation, composition, loop, manifold, and theory. Whenever I detected this issue, I included an explanation. Also, whenever possible, I changed obscure mathematical terms to terms with more transparent meaning (for example, I would write "Langlands relation" instead of "Langlands correspondence"). You might find it useful to consult the Glossary and the Index whenever there is a word that seems unclear.

Please check out my website http://edwardfrenkel.com for updates and supporting materials, and send me an e-mail to share your thoughts about the book (my e-mail address can be found on the website). Your feedback will be much appreciated.

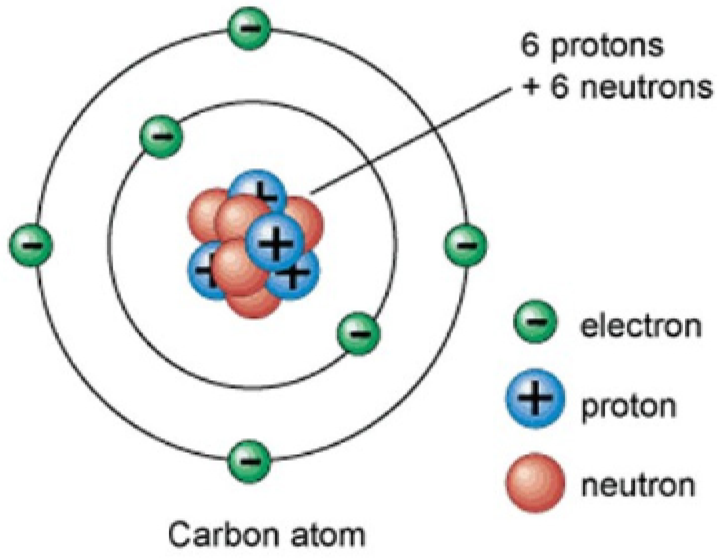
Chapter 1

A Mysterious Beast

How does one become a mathematician? There are many ways that this can happen. Let me tell you how it happened to me.

It might surprise you, but I hated math when I was at school. Well, "hated" is perhaps too strong a word. Let's just say I didn't like it. I thought it was boring. I could do my work, sure, but I didn't understand why I was doing it. The material we discussed in class seemed pointless, irrelevant. What really excited me was physics – especially quantum physics. I devoured every popular book on the subject that I could get my hands on. I grew up in Russia, where such books were easy to find.

I was fascinated with the quantum world. Ever since ancient times, scientists and philosophers had dreamed about describing the fundamental nature of the universe – some even hypothesized that all matter consists of tiny pieces called atoms. Atoms were proved to exist at the beginning of the twentieth century, but at around the same time, scientists discovered that each atom could be divided further. Each atom, it turned out, consists of a nucleus in the middle and electrons orbiting it. The nucleus, in turn, consists of protons and neutrons, as shown on the diagram below.[1]



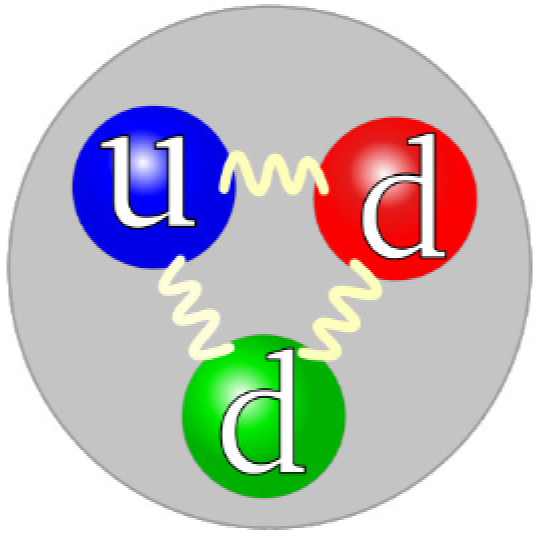
And what about protons and neutrons? The popular books that I was reading told me that they are built of the elementary particles called "quarks."

I liked the name quarks, and I especially liked how this name came about. The physicist who invented these particles, Murray Gell-Mann, borrowed this name from James Joyce's book *Finnegans Wake*, where there is a mock poem that goes like this:

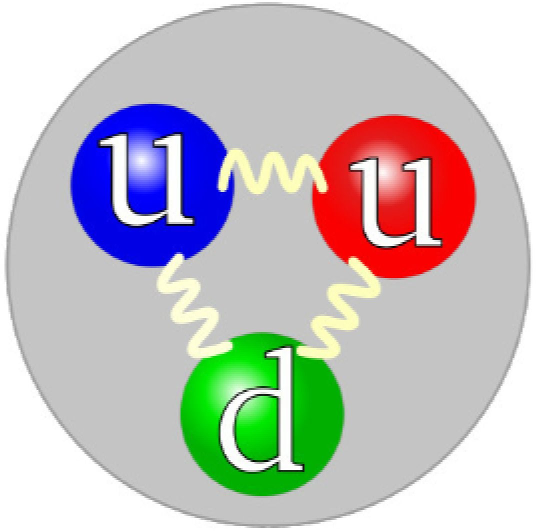
> Three quarks for Muster Mark!
> Sure he hasn't got much of a bark
> And sure any he has it's all beside the mark.

I thought it was really cool that a physicist would name a particle after a novel. Especially such a complex and non-trivial one as *Finnegans Wake*. I must have been around thirteen, but I already knew by then that scientists were supposed to be these reclusive and unworldly creatures who were so deeply involved in their work that they had no interest whatsoever in other aspects of life such as Art and Humanities. I wasn't like this. I had many friends, liked to read, and was interested in many things besides science. I liked to play soccer and spent endless hours chasing the ball with my friends. I discovered Impressionist paintings around the same time (it started with a big volume about Impressionism, which I found in my parents' library). Van Gogh was my favorite. Enchanted by his works, I even tried to paint myself. All of these interests had actually made me doubt whether I was really cut out to be a scientist. So when I read that Gell-Mann, a great physicist, Nobel Prize–winner, had such diverse interests (not only literature, but also linguistics, archaeology, and more), I was very happy.

According to Gell-Mann, there are two different types of quarks, "up" and "down," and different mixtures of them give neutrons and protons their characteristics. A neutron is made of two down and one up quarks, and a proton is made of two up and one down quarks, as shown on the pictures.[2]



Neutron



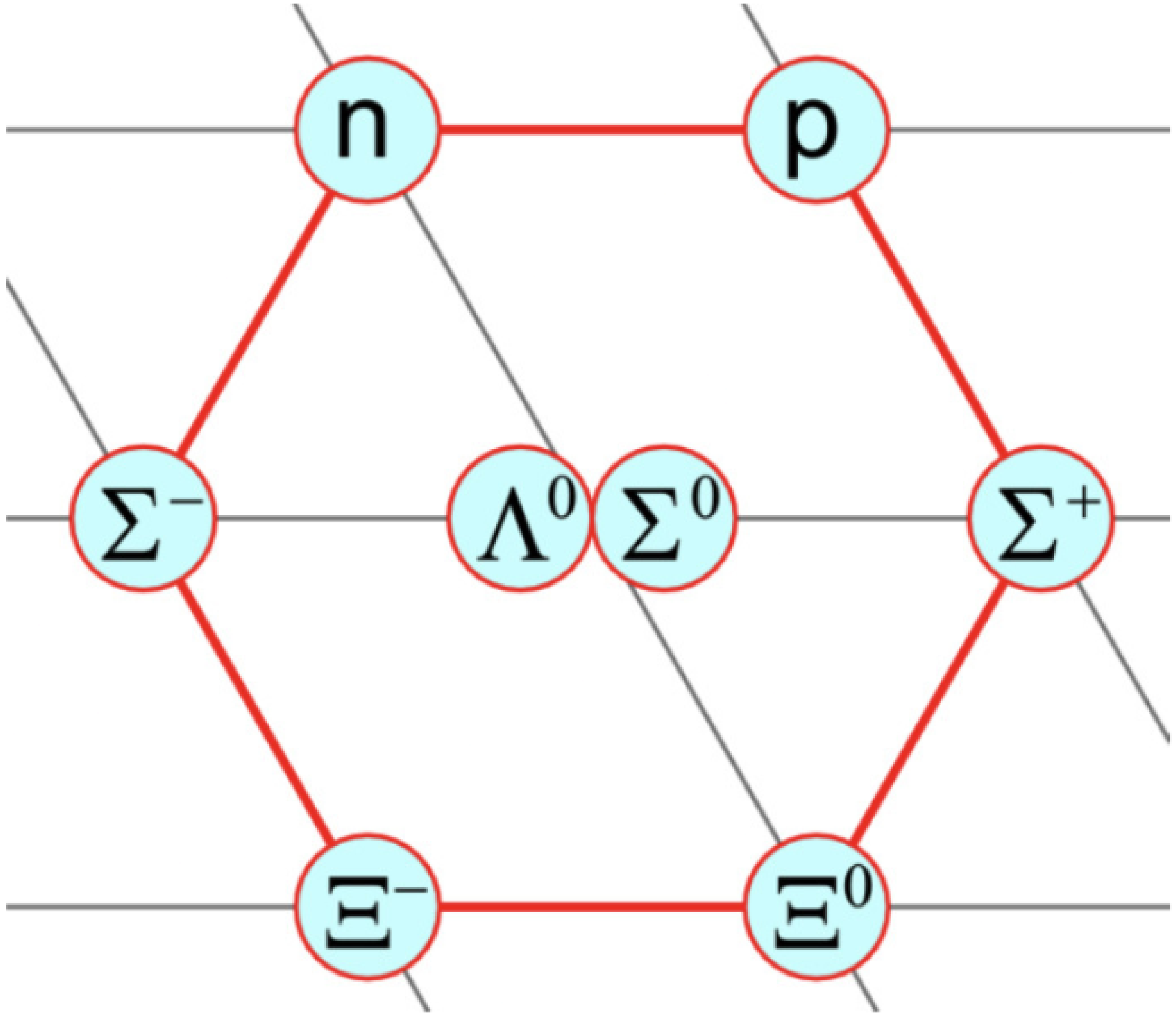
Proton

That was clear enough. But how physicists guessed that protons and neutrons were not indivisible particles but rather were built from smaller blocks was murky.

The story goes that by the late 1950s, a large number of apparently elementary particles, called hadrons, was discovered. Neutrons and protons are both hadrons, and of course they play major roles in everyday life as the building blocks of matter. As for the rest of hadrons – well, no one had any idea what they existed for (or "who ordered them," as one researcher put it). There were so many of them that the influential physicist Wolfgang Pauli joked that physics was turning into botany. Physicists desperately needed to rein in the hadrons, to find the underlying principles that govern their behavior and would explain their maddening proliferation.

Gell-Mann, and independently Yuval Ne'eman, proposed a novel classification scheme. They both showed that hadrons can be naturally split into small families, each consisting of eight or ten particles. They called them octets and decuplets. Particles within each of the families had similar properties.

In the popular books I was reading at the time, I would find octet diagrams like this:



Here the proton is marked as *p*, the neutron is marked as *n*, and there are six other particles with strange names expressed by Greek letters.

But why 8 and 10, and not 7 and 11, say? I couldn't find a coherent explanation in the books I was reading. They would mention a mysterious idea of Gell-Mann called the "eightfold way" (referencing the "Noble Eightfold Path" of Buddha). But they never attempted to explain what this was all about.

This lack of explanation left me deeply unsatisfied. The key parts of the story remained hidden. I wanted to unravel this mystery but did not know how.

As luck would have it, I got help from a family friend. I grew up in a small industrial town called Kolomna, population 150,000, which was about seventy miles away from Moscow, or just over two hours by train. My parents worked as engineers at a large company making heavy machinery. Kolomna is an old town on the intersection of two rivers that was founded in 1177 (only thirty years after the founding of Moscow). There are still a few pretty churches and the city wall to attest to Kolomna's storied past. But it's not exactly an educational or intellectual center. There was only one small college there, which prepared schoolteachers. One of the professors there, a mathematician named Evgeny Evgenievich Petrov, however, was an old friend of my parents. And one day my mother met him on the street after a long time, and they started talking. My mom liked to tell her friends about me, so I came up in conversation. Hearing that I was interested in science, Evgeny Evgenievich said, "I must meet him. I will try to convert him to math."

"Oh no," my mom said, "he doesn't like math. He thinks it's boring. He wants to do quantum physics."

"No worries," replied Evgeny Evgenievich, "I think I know how to change his mind."

A meeting was arranged. I wasn't particularly enthusiastic about it, but I went to see Evgeny Evgenievich at his office anyway.

I was just about to turn fifteen, and I was finishing the ninth grade, the penultimate year of high school. (I was a year younger than my classmates because I had skipped the sixth grade.) Then in his early forties, Evgeny Evgenievich was friendly and unassuming. Bespectacled, with a beard stubble, he was just what I imagined a mathematician would look like, and yet there was something captivating in the probing gaze of his big eyes. They exuded unbounded curiosity about everything.

It turned out that Evgeny Evgenievich indeed had a clever plan how to convert me to math. As soon as I came to his office, he asked me, “So, I hear you like quantum physics. Have you heard about Gell-Mann’s eightfold way and the quark model?”

“Yes, I’ve read about this in several popular books.”

“But do you know what was the basis for this model? How did he come up with these ideas?”

“Well…”

“Have you heard about the group $SU(3)$?”

“*SU* what?”

“How can you possibly understand the quark model if you don’t know what the group $SU(3)$ is?”

He pulled out a couple of books from his bookshelf, opened them, and showed me pages of formulas. I could see the familiar octet diagrams, such as the one shown above, but these diagrams weren’t just pretty pictures; they were part of what looked like a coherent and detailed explanation.

Though I could make neither head nor tail of these formulas, it became clear to me right away that they contained the answers I had been searching for. This was a moment of epiphany. I was mesmerized by what I was seeing and hearing; touched by something I had never experienced before; unable to express it in words but feeling the energy, the excitement one feels from hearing a piece of music or seeing a painting that makes an unforgettable impression. All I could think was “Wow!”

“You probably thought that mathematics is what they teach you in school,” Evgeny Evgenievich said. He shook his head, “No, this” – he pointed at the formulas in the book – “is what mathematics is about. And if you really want to understand quantum physics, this is where you need to start. Gell-Mann predicted quarks using a beautiful mathematical theory. It was in fact a mathematical discovery.”

“But how do I even begin to understand this stuff?”

It looked kind of scary.

“No worries. The first thing you need to learn is the concept of a symmetry group. That’s the main idea. A large part of mathematics, as well as theoretical physics, is based on it. Here are some books I want to give you. Start reading them and mark the sentences that you don’t understand. We can meet here every week and talk about this.”

He gave me a book about symmetry groups and also a couple of others on different topics: about the so-called p-adic numbers (a number system radically different from the numbers we are used to) and about topology (the study of the most fundamental properties of geometric shapes). Evgeny Evgenievich had impeccable taste: he found a perfect combination of topics that would allow me to see this mysterious beast – *Mathematics* – from different sides and get excited about it.

At school we studied things like quadratic equations, a bit of calculus, some basic Euclidean geometry, and trigonometry. I had assumed that all mathematics somehow revolved around these subjects, that perhaps problems became more complicated but stayed within the same general framework I was familiar with. But the books Evgeny Evgenievich gave me contained glimpses of an entirely different world, whose existence I couldn’t even imagine.

I was instantly converted.

Chapter 2

The Essence of Symmetry

In the minds of most people, mathematics is all about numbers. They imagine mathematicians as people who spend their days crunching numbers: big numbers, and even bigger numbers, all having exotic names. I had thought so too – at least, until Evgeny Evgenievich introduced me to the concepts and ideas of modern math. One of them turned out to be the key to the discovery of quarks: the concept of symmetry.

What is symmetry? All of us have an intuitive understanding of it – we know it when we see it. When I ask people to give me an example of a symmetric object, they point to butterflies, snowflakes, or the human body.

Photo by K.G. Libbrecht

But if I ask them what we mean when we say that an object is symmetrical, they hesitate.

Here is how Evgeny Evgenievich explained it to me. “Let’s look at this round table and this square table,” he pointed at the two tables in his office. “Which one is more symmetrical?”

“Of course, the round table, isn’t it obvious?”

“But why? Being a mathematician means that you don’t take ‘obvious’ things for granted but try to reason. Very often you’ll be surprised that the most obvious answer is actually wrong.”

Noticing confusion on my face, Evgeny Evgenievich gave me a hint: “What is the property of the round table that makes it more symmetrical?”

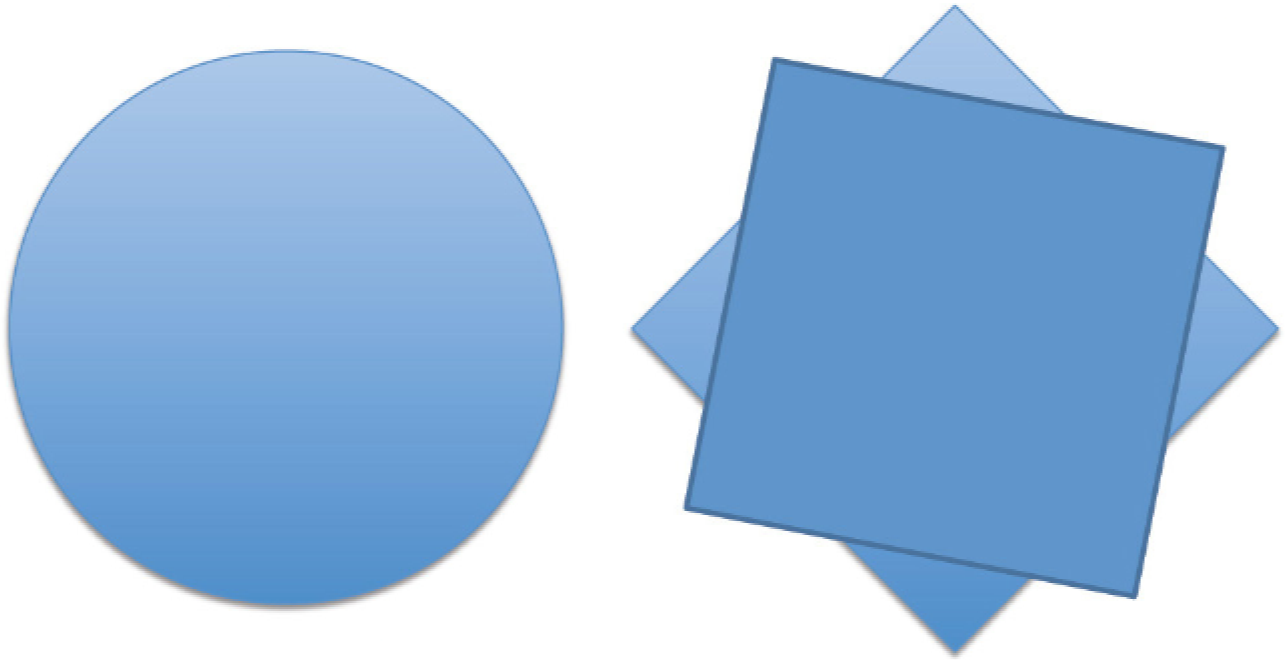
I thought about this for a while, and then it hit me: “I guess the symmetry of an object has to do with it keeping its shape and position unchanged even when we apply changes to it.”

Evgeny Evgenievich nodded.

"Indeed. Let's look at all possible transformations of the two tables which preserve their shape and position," he said. "In the case of the round table..."

I interrupted him: "Any rotation around the center point will do. We will get back the same table positioned in the same way. But if we apply an arbitrary rotation to a square table, we will typically get a table positioned differently. Only rotations by 90 degrees and its multiples will preserve it."

"Exactly! If you leave my office for a minute, and I turn the round table by any angle, you won't notice the difference. But if I do the same to the square table, you will, unless I turn it by 90, 180, or 270 degrees."



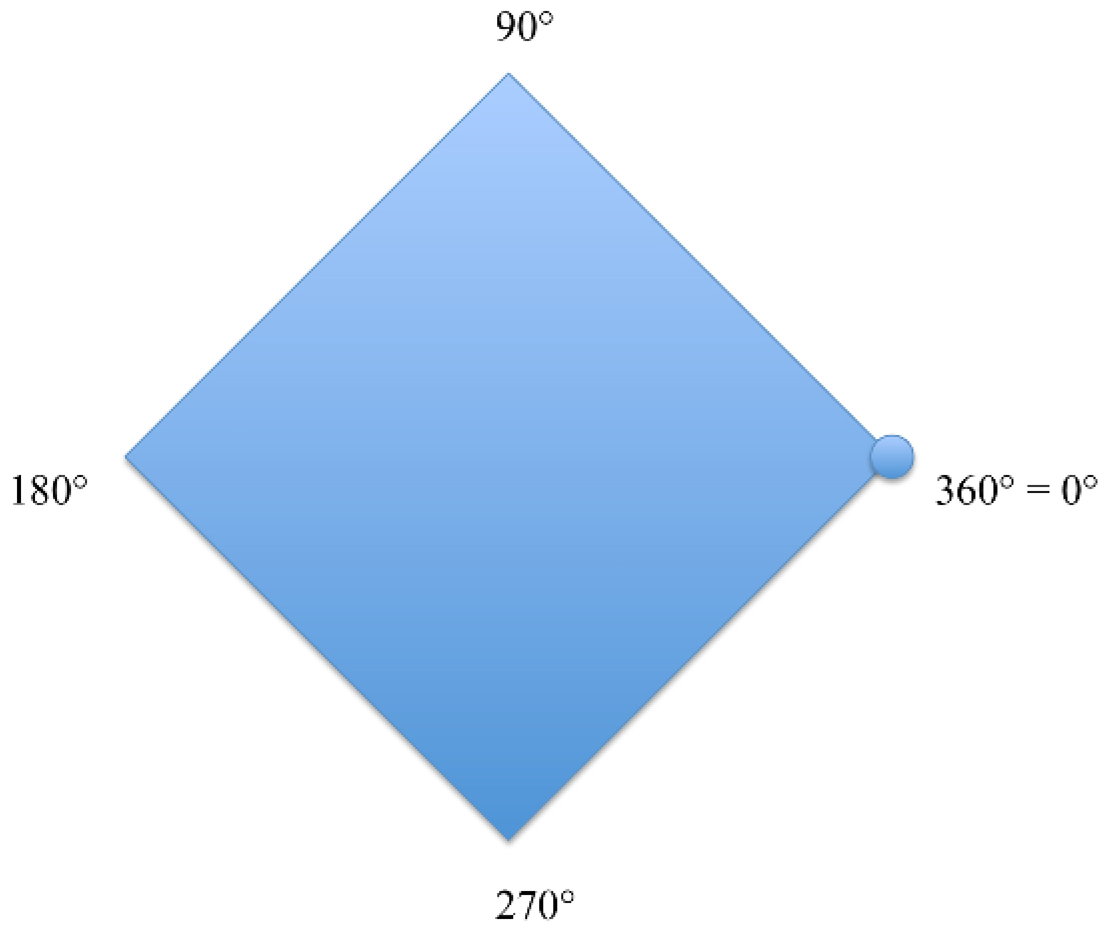
Rotation of a round table by any angle does not change its position, but rotation of a square table by an angle that is not a multiple of 90 degrees does change its position (both are viewed here from above)

He continued: "Such transformations are called symmetries. So you see that the square table has only four symmetries, whereas the round table has many more of them – it actually has infinitely many symmetries. That's why we say that the round table is more symmetrical."

This made a lot of sense.

"This is a fairly straightforward observation," continued Evgeny Evgenievich. "You don't have to be a mathematician to see this. But if you are a mathematician, you ask the next question: what are *all* possible symmetries of a given object?"

Let's look at the square table. Its symmetries[1] are these four rotations around the center of the table: by 90 degrees, 180 degrees, 270 degrees, and 360 degrees, counterclockwise.[2] A mathematician would say that the *set* of symmetries of the square table consists of four elements, corresponding to the angles 90, 180, 270, and 360 degrees. Each rotation takes a fixed corner (marked with a balloon on the picture below) to one of the four corners.



One of these rotations is special; namely, rotation by 360 degrees is the same as rotation by 0 degrees, that is, no rotation at all. This is a special symmetry because it actually does nothing to our object: each point of the table ends up in exactly the same position as it was before. We call it the *identical symmetry*, or just the *identity*.[3]

Note that rotation by any angle greater than 360 degrees is equivalent to rotation by an angle between 0 and 360 degrees. For example, rotation by 450 degrees is the same as rotation by 90 degrees, because 450 = 360 + 90. That's why we will only consider rotations by angles between 0 and 360 degrees.

Here comes the crucial observation: if we apply two rotations from the list {90°, 180°, 270°, 360°} one after another, we obtain another rotation from the same list. We call this new symmetry the *composition* of the two.

Of course, this is obvious: each of the two symmetries preserves the table. Hence the composition of the two symmetries also preserves it. Therefore this composition has to be a symmetry as well. For example, if we rotate the table by 90 degrees and then again by 180 degrees, the net result is the rotation by 270 degrees.

Let's see what happens with the table under these symmetries. Under the counterclockwise rotation by 90 degrees, the right corner of the table (the one marked with a balloon on the previous picture) will go to the upper corner. Next, we apply the rotation by 180 degrees, so the upper corner will go to the down corner. The net result will be that the right corner will go to the down corner. This is the result of the counterclockwise rotation by 270 degrees.

Here is one more example:

$$90^\circ + 270^\circ = 0^\circ.$$

By rotating by 90 degrees and then by 270 degrees, we get the rotation by 360 degrees. But the effect of the rotation by 360 degrees is the same as that of the rotation by 0 degrees, as we have discussed above – this is the "identity symmetry."

In other words, the second rotation by 270 degrees undoes the initial rotation by 90 degrees. This is in fact an important property: any symmetry can be *undone*; that is, for any symmetry *S* there exists another symmetry *S′* such that their composition is the identity symmetry. This *S′* is called the *inverse* of symmetry *S*. So we see that rotation by 270 degrees is the inverse of the rotation by 90 degrees. Likewise, the inverse of the rotation by 180 degrees is the same rotation by 180 degrees.

We now see that what looks like a very simple collection of symmetries of the square table – the four rotations {90°, 180°, 270°, 0°} – actually has a lot of inner structure, or rules for how the members of the set can interact.

First of all, we can compose any two symmetries (that is, apply them one after another).

Second, there is a special symmetry, the identity. In our example, this is the rotation by 0 degrees. If we compose it with any other symmetry, we get back the same symmetry. For example,

$$90^\circ + 0^\circ = 90^\circ, \qquad 180^\circ + 0^\circ = 180^\circ, \qquad \text{etc.}$$

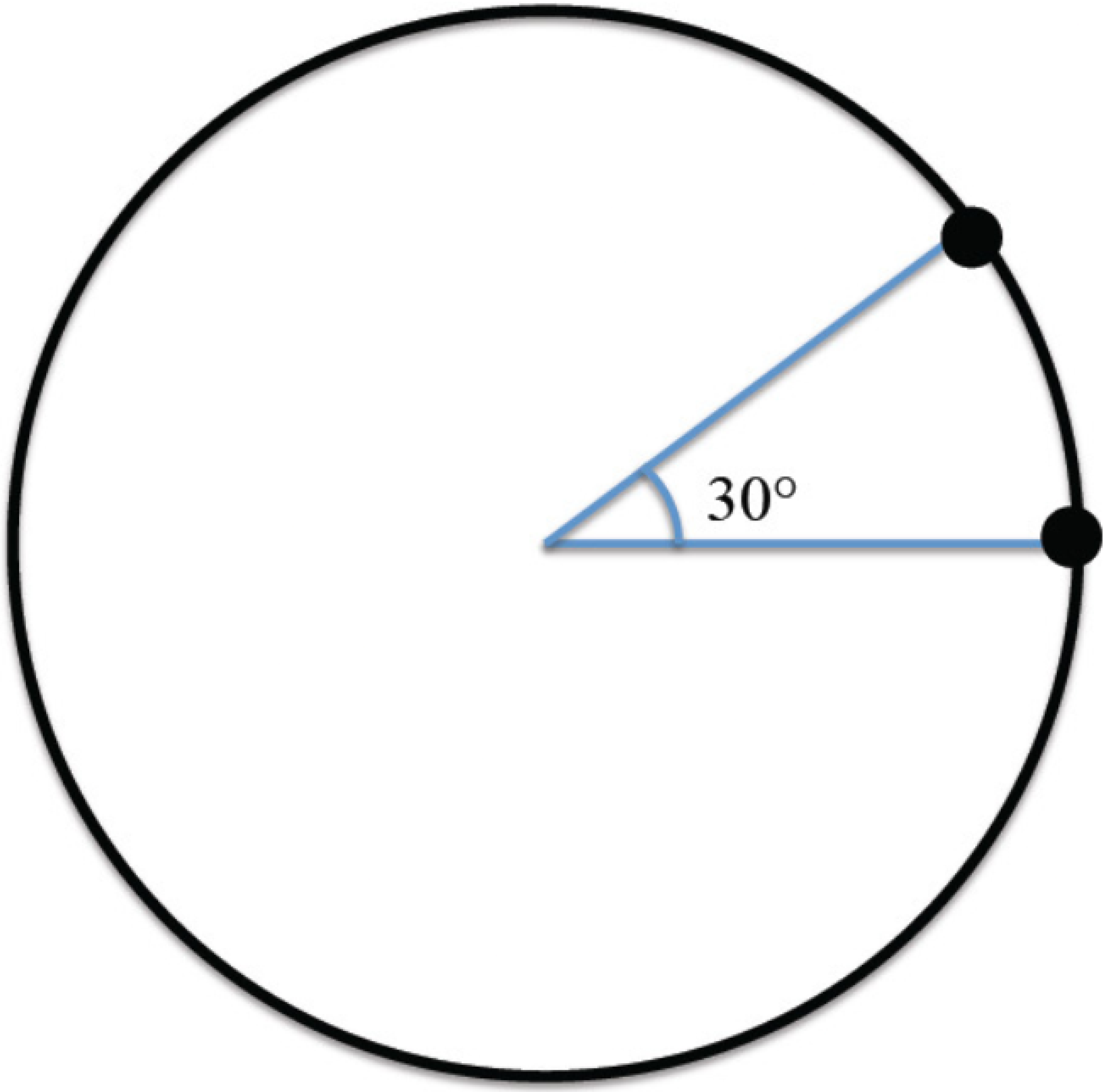
Third, for any symmetry *S*, there is the inverse symmetry *S′* such that the composition of *S* and *S′* is the identity.

And now we come to the main point: the set of rotations along with these three structures comprise an example of what mathematicians call a *group*.

The symmetries of any other object also constitute a group, which in general has more elements – possibly, infinitely many.[4]

Let's see how this works in the case of a round table. Now that we have gained some experience, we can see right away that the set of all symmetries of the round table is just the set of all possible rotations (not just by multiples of 90 degrees), and we can visualize it as the set of all points of a circle.

Each point on this circle corresponds to an angle between 0 and 360 degrees, representing the rotation of the round table by this angle in the counterclockwise direction. In particular, there is a special point corresponding to rotation by 0 degrees. It is marked on the picture below, together with another point corresponding to rotation by 30 degrees.



We should not think of the points of this circle as points of the round table, though. Rather, each point of the circle represents a particular rotation of the round table. Note that the round table does not have a preferred point, but our circle does; namely, the one corresponding to rotation by 0 degrees.[5]

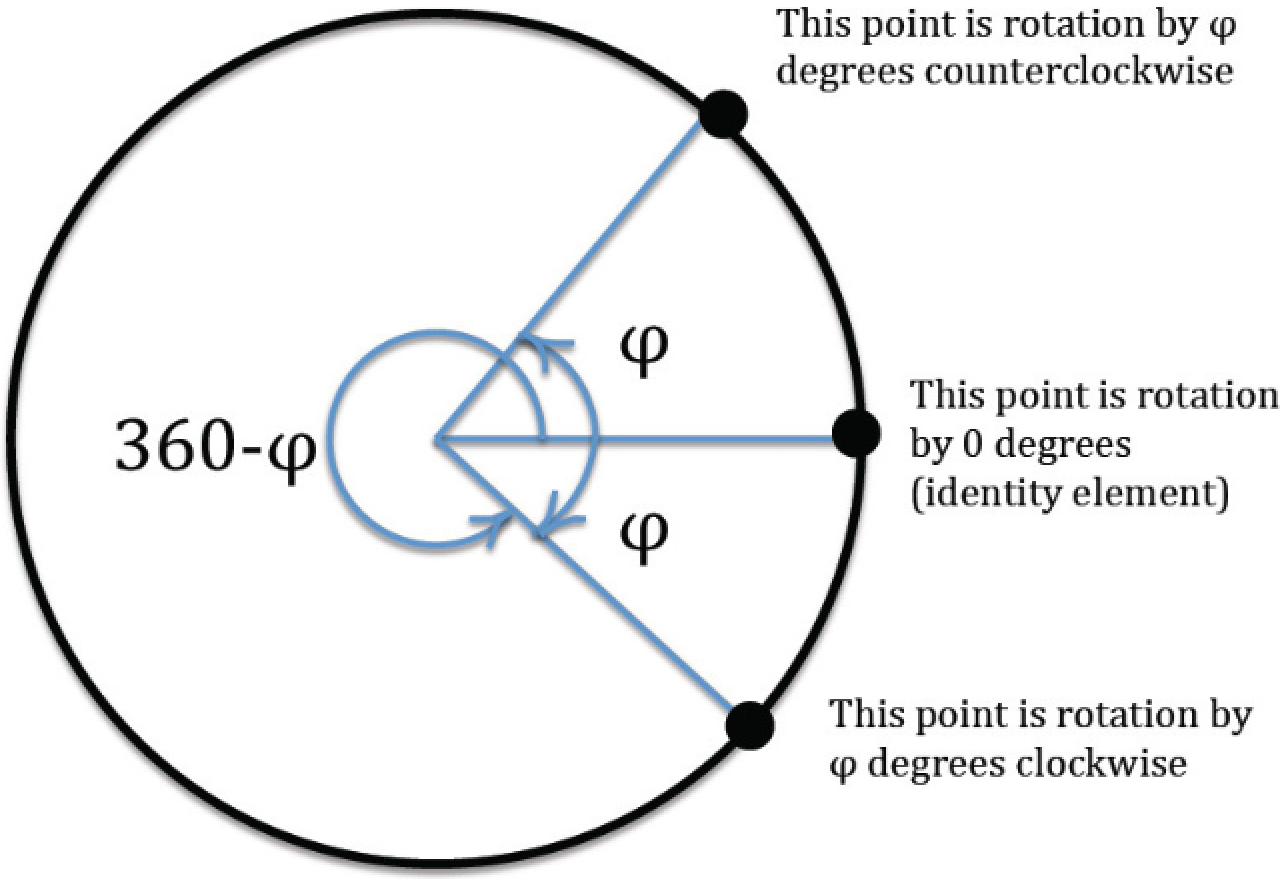
Now let's see if the above three structures can be applied to the set of points of the circle.

First, the composition of two rotations, by φ_1 and φ_2 degrees, is the rotation by $\varphi_1 + \varphi_2$ degrees. If $\varphi_1 + \varphi_2$ is greater than 360, we simply subtract 360 from it. In mathematics, this is called *addition modulo* 360. For example, if $\varphi_1 = 195$ and $\varphi_2 = 250$, then the sum of the two angles is 445, and the rotation by 445 degrees is the same as the rotation by 85 degrees. So, in the group of rotations of the round table we have

$$195^\circ + 250^\circ = 85^\circ.$$

Second, there is a special point on the circle corresponding to the rotation by 0 degrees. This is the identity element of our group.

Third, the inverse of the counterclockwise rotation by φ degrees is the counterclockwise rotation by $(360-\varphi)$ degrees, or equivalently, clockwise rotation by φ degrees (see the drawing).


This point is rotation by φ degrees counterclockwise
φ
360-φ
This point is rotation by 0 degrees (identity element)
φ
This point is rotation by φ degrees clockwise

Thus, we have described the group of rotations of the round table. We will call it the *circle group*. Unlike the group of symmetries of the square table, which has four elements, this group has infinitely many elements because there are infinitely many angles between 0 and 360 degrees.

We have now put our intuitive understanding of symmetry on firm theoretical ground – indeed, we've turned it into a mathematical concept. First, we postulated that a symmetry of a given object is a transformation that preserves it and its properties. Then we made a decisive step: we focused on the set of all symmetries of a given object. In the case of a square table, this set consists of four elements (rotations by multiples of 90 degrees); in the case of a round table, it is an infinite set (of all points on the circle). Finally, we described the neat structures that this set of symmetries always possesses: any two symmetries can be composed to produce another symmetry, there exists the identical symmetry, and for each symmetry there exists its inverse. (The composition of symmetries also satisfies the associativity property described in endnote 4.) Thus, we came to the mathematical concept of a group.

A group of symmetries is an abstract object that is quite different from the concrete object we started with. We cannot touch or hold the set of symmetries of a table (unlike the table itself), but we can imagine it, draw its elements, study it, talk about it. Each element of this abstract set has a concrete meaning, though: it represents a particular transformation of a concrete object, its symmetry.

Mathematics is about the study of such abstract objects and concepts.

Experience shows that symmetry is an essential guiding principle for the laws of nature. For example, a snowflake forms a perfect hexagonal shape because that turns out to be the lowest energy state into which crystallized water molecules are forced. The symmetries of the snowflake are rotations by multiples of 60 degrees; that is, 60, 120, 180, 240, 300, and 360 (which is the same as 0 degrees). In addition, we can "flip" the snowflake along each of the six axes corresponding to those angles. All of these rotations and flips preserve the shape and position of the snowflake, and hence they are its symmetries.[*]

In the case of a butterfly, flipping it turns it upside down. Since it has legs on one side, the flip is not, strictly speaking, a symmetry of the butterfly. When we say that a butterfly is symmetrical, we are talking about an idealized version of it, where its front and back are exactly the same (unlike those of an actual butterfly). Then the flip exchanging the left and the right wings becomes a symmetry. (Alternatively, we can imagine exchanging the wings without turning the butterfly upside down.)

This brings up an important point: there are many objects in nature whose symmetries are approximate. A real-life table is not perfectly round or perfectly square, a live butterfly has an asymmetry between its front and back, and a human body is not fully symmetrical. However, even in this case it turns out to be useful to consider their abstract, idealized versions, or models – a perfectly round table or an image of the butterfly in which we don't distinguish between the front and the back. We then explore the symmetries of these idealized objects and adjust whatever inferences we can make from this analysis to account for the difference between a real-life object and its model.

This is not to say that we do not appreciate asymmetry; we do, and we often find beauty in it. But the main point of the mathematical theory of symmetry is not aesthetic. It is to formulate the concept of symmetry in the most general, and hence inevitably most abstract, terms so that it could be applied in a unified fashion in different domains, such as geometry, number theory, physics, chemistry, biology, and so on. Once we develop such a theory, we can also talk about the mechanisms of symmetry breaking – viewing asymmetry as emergent, if you will. For example, elementary particles acquire masses because the so-called gauge symmetry they obey (which will be discussed in Chapter 16) gets broken. This is facilitated by the Higgs boson, an elusive particle recently discovered at the Large Hadron Collider under the city of Geneva.[6] The study of such mechanisms of symmetry breaking yields invaluable insights into the behavior of the fundamental blocks of nature.

I'd like to point out some of the basic qualities of the abstract theory of symmetry because this is a good illustration of why mathematics is important.

The first is *universality*. The circle group is not only the group of symmetries of a round table, but also of all other round objects, like a glass, a bottle, a column, and so forth. In fact, to say that a given object is round is the same as to say that its group of symmetries is the circle group. This is a powerful statement: we realize that we can describe an important attribute of an object ("being round") by describing its symmetry group (the circle). Likewise, "being square" means that the group of symmetries is the group of four elements described above. In other words, the same abstract mathematical object (such as the circle group) serves many different concrete objects, and it points to universal properties that they all have in common (such as roundness).[7]

The second is *objectivity*. The concept of a group, for example, is independent of our interpretation. It means the same thing to anyone who learns it. Of course, in order to understand it, one has to know the language in which it is expressed, that is, mathematical language. But anyone can learn this language. Likewise, if you want to understand the meaning of René Descartes' sentence "*Je pense, donc je suis*," you need to know French (at least, those words that are used in this sentence) – but anyone can learn it. However, in the case of the latter sentence, once we understand it, different interpretations of it are possible. Also, different people may agree or disagree on whether a particular interpretation of this sentence is true or false. In contrast, the meaning of a logically consistent mathematical statement is not subject to interpretation.[8] Furthermore, its truth is also objective. (In general, the truth of a particular statement may depend on the system of axioms within which it is considered. However, even then, this dependence on the axioms is also objective.) For example, the statement "the group of symmetries of a round table is a circle" is true to anyone, anywhere, at any time. In other words, mathematical truths are the necessary truths. We will talk more about this in Chapter 18.

The third, closely related, quality is *endurance*. There is little doubt that the Pythagorean theorem meant the same thing to the ancient Greeks as it does to us today, and there is every reason to expect that it will mean the same thing to anyone in the future. Likewise, all true mathematical statements we talk about in this book will remain true forever.

The fact that such objective and enduring knowledge exists (and moreover, belongs to all of us) is nothing short of a miracle. It suggests that mathematical concepts exist in a world separate from the physical and mental worlds – which is sometimes referred to as the Platonic world of mathematics (we will talk more about that in the closing chapter). We still don't fully understand what it is and what drives mathematical discovery. But it's clear that this hidden reality is bound to play a larger and larger role in our lives, especially with the advent of new computer technologies and 3D printing.

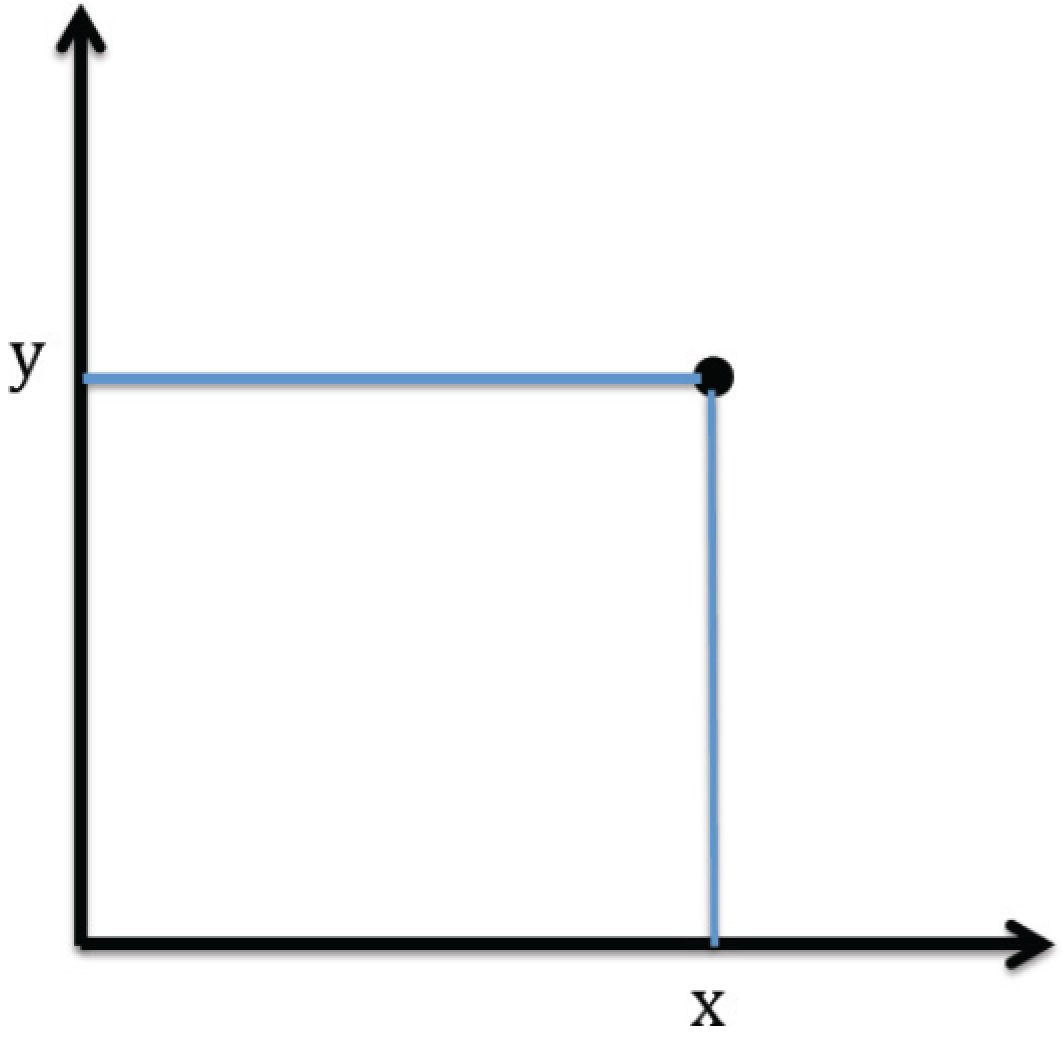
The fourth quality is *relevance* of mathematics to the physical world. For example, a lot of progress has been made in quantum physics in the past fifty years because of the application of the concept of symmetry to elementary particles and interactions between them. From this point of view, a particle, such as an electron or a quark, is like a round table or a snowflake, and its behavior is very much determined by its symmetries. (Some of these symmetries are exact, and some are approximate.)

The discovery of quarks is a perfect example of how this works. Reading the books Evgeny Evgenievich gave me, I learned that at the root of the Gell-Mann and Ne'eman classification of hadrons that we talked about in the previous chapter is a *symmetry group*. This group had been previously studied by mathematicians – who did not anticipate any connections to subatomic particles whatsoever. The mathematical name for it is $SU(3)$. Here S and U stand for "special unitary." This group is very similar in its properties to the group of symmetries of the sphere, which we will talk about in detail in Chapter 10.

Mathematicians had previously described the representations of the group $SU(3)$, that is, different ways that the group $SU(3)$ can be realized as a symmetry group. Gell-Mann and Ne'eman noticed the similarity between the structure of these representations and the patterns of hadrons that they had found. They used this information to classify hadrons.

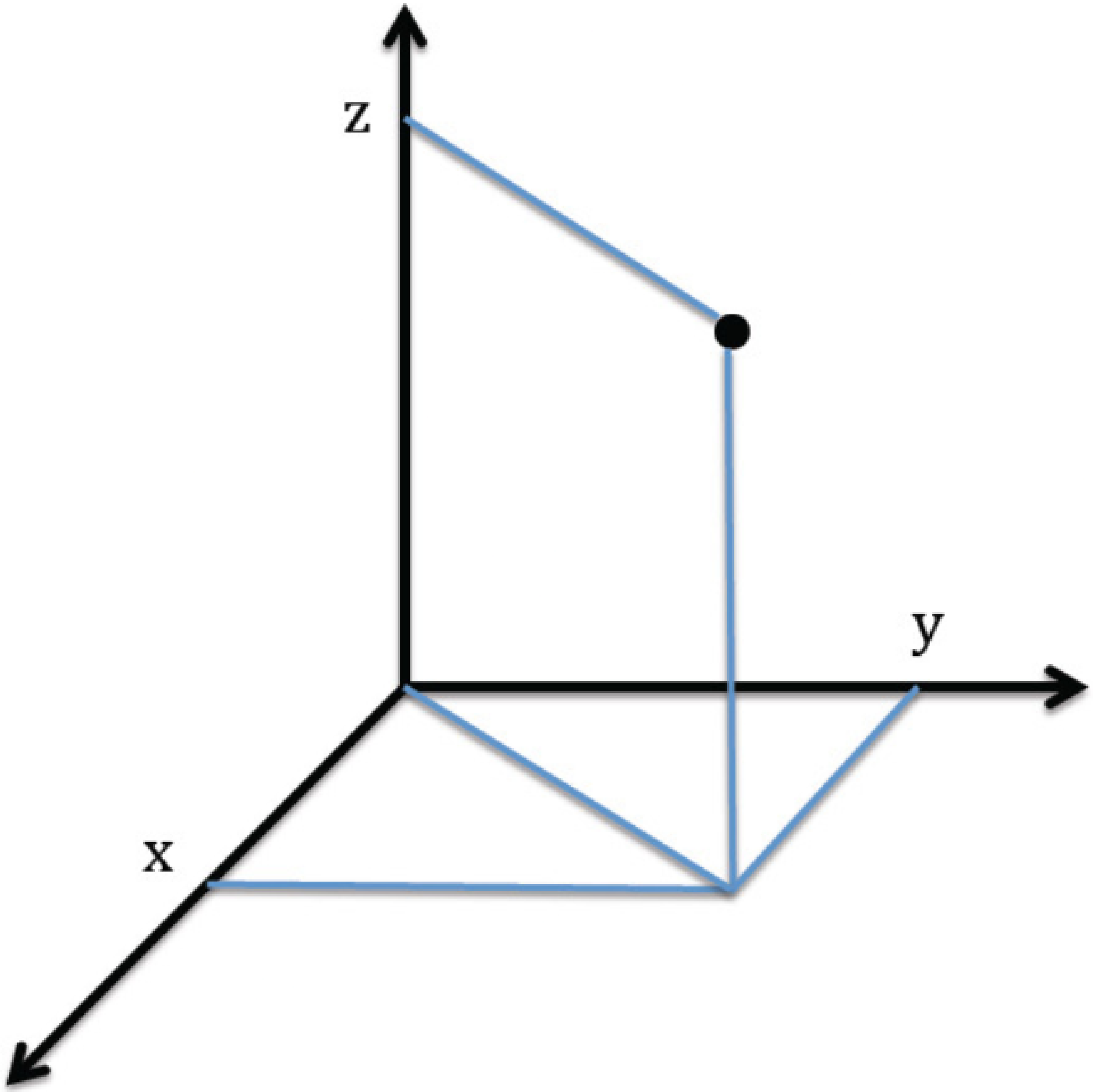
The word "representation" is used in mathematics in a particular way, which is different from its more common usage. So let me pause and explain what this word means in the present context. Perhaps, it would help if I first give an example. Recall the group of rotations of a round table discussed above, the circle group. Now imagine extending the tabletop infinitely far in all directions. This way we obtain an abstract mathematical object: a plane. Each rotation of the tabletop, around its center, gives rise to a rotation of this plane around the same point. Thus, we obtain a rule that assigns a symmetry of this plane (a rotation) to every element of the circle group. In other words, each element of the circle group may be represented by a symmetry of the plane. For this reason mathematicians refer to this process as a *representation* of the circle group.

Now, the plane is two-dimensional because it has two coordinate axes and hence each point has two coordinates:



Therefore, we say that we have constructed a "two-dimensional representation" of the group of rotations. It simply means that each element of the group of rotations is realized as a symmetry of a plane.[9]

There are also spaces of dimension greater than two. For example, the space around us is three-dimensional. That is to say, it has three coordinate axes, and so in order to specify a position of a point, we need to specify its three coordinates (x, y, z) as shown on this picture:



We cannot imagine a four-dimensional space, but mathematics gives us a universal language that allows us to talk about spaces of any dimension. Namely, we represent points of the four-dimensional space by quadruples of numbers (x, y, z, t), just like points of the three-dimensional space are represented by triples of numbers (x, y, z). In the same way, we represent points of an n-dimensional space, for any natural number n, by n-tuples of numbers. If you have used a spreadsheet program, then you have encountered such n-tuples: they appear as rows in a spreadsheet, each of the n numbers corresponding to a particular attribute of the stored data. Thus, every row in a spreadsheet refers to a point in an n-dimensional space. (We will talk more about spaces of various dimensions in Chapter 10.)

If each element of a group can be realized, in a consistent manner,[10] as a symmetry of an n-dimensional space, then we say that the group has an "n-dimensional representation."

It turns out that a given group can have representations of different dimensions. The reason elementary particles can be assembled in families of 8 and 10 particles is that the group $SU(3)$ is known to have an 8-dimensional and a 10-dimensional representation. The 8 particles of each octet constructed by Gell-Mann and Ne'eman (like the one shown on the diagram in the previous chapter) are in one-to-one correspondence with the 8 coordinate axes of an 8-dimensional space which is a representation of $SU(3)$. The same goes for the decuplet of particles. (But particles cannot be assembled in families of, say, 7 or 11 particles because mathematicians have proved that the group $SU(3)$ has no 7- or 11-dimensional representations.)

At first, this was just a convenient way to combine the particles with similar properties. But then Gell-Mann went further. He postulated that there was a deep reason behind this classification scheme. He basically said that this scheme works so well because hadrons consist of smaller particles – sometimes two and sometimes three of them – the quarks. A similar proposal was made independently by physicist George Zweig (who called the particles "aces").

This was a stunning proposal. Not only did it go against the popular belief at the time that protons and neutrons as well as other hadrons were indivisible elementary particles, these new particles were supposed to have electric charges that were fractions of the charge of the electron. This was a startling prediction because no one had seen such particles before. Yet, quarks were soon found experimentally, and as predicted, they had fractional electric charges!

What motivated Gell-Mann and Zweig to predict the existence of quarks? Mathematical theory of representations of the group $SU(3)$. Specifically, the fact that the group $SU(3)$ has two different 3-dimensional representations. (Actually, that's the reason there is a "3" in this group's name.) Gell-Mann and Zweig suggested that these two representations should describe two families of fundamental particles: 3 quarks and 3 anti-quarks. It turns out that the 8- and 10-dimensional representations of $SU(3)$ can be built from the 3-dimensional ones. And this gives us a precise blueprint for how to construct hadrons from quarks – just like in Lego.

Gell-Mann named the 3 quarks "up," "down," and "strange."[11] A proton consists of two up quarks and one down quark, whereas a neutron consists of two down quarks and one up quark, as we saw on the pictures in the previous chapter. Both of these particles belong to the octet shown on the diagram in the previous chapter. Other particles from this octet involve the strange quark as well as the up and down quarks. There are also octets that consist of particles that are composites of one quark and one anti-quark.

The discovery of quarks is a good example of the paramount role played by mathematics in science that we discussed in the Preface. These particles were predicted not on the basis of empirical data, but on the basis of mathematical symmetry patterns. This was a purely theoretical prediction, made within the framework of a sophisticated mathematical theory of representations of the group $SU(3)$. It took physicists years to master this theory (and in fact there was some resistance to it at first), but it is now the bread and butter of elementary particle physics. Not only did it provide a classification of hadrons, it also led to the discovery of quarks, which forever changed our understanding of physical reality.

Imagine: a seemingly esoteric mathematical theory empowered us to get to the heart of the building blocks of nature. How can we not be enthralled by the magic harmony of these tiny blobs of matter, not marvel at the capacity of mathematics to reveal the inner workings of the universe?

As the story goes, Albert Einstein's wife Elsa remarked, upon hearing that a telescope at the Mount Wilson Observatory was needed to determine the shape of space-time: "Oh, my husband does this on the back of an old envelope."

Physicists do need expensive and sophisticated machines such as the Large Hadron Collider in Geneva, but the amazing fact is that scientists like Einstein and Gell-Mann have used what looks like the purest and most abstract mathematical knowledge to unlock the deepest secrets of the world around us.

Regardless of who we are and what we believe in, we all share this knowledge. It brings us closer together and gives a new meaning to our love for the universe.

*Note that flipping a table is not a symmetry: this would turn it upside down – let's not forget that a table has legs. If we were to consider a square or a circle (no legs attached), then flips would be bona fide symmetries. We would have to include them in the corresponding symmetry groups.

Chapter 3

The Fifth Problem

Evgeny Evgenievich's plan worked perfectly: I was "converted" to math. I was learning quickly, and the deeper I delved into math, the more my fascination grew, the more I wanted to know. This is what happens when you fall in love.

I started meeting with Evgeny Evgenievich on a regular basis. He would give me books to read, and I would meet him once a week at the pedagogical college where he taught to discuss what I had read. Evgeny Evgenievich played soccer, ice hockey, and volleyball on a regular basis, but like many men in the Soviet Union in those days, he was a chain smoker. For a long time afterward, the smell of cigarettes was associated in my mind with doing mathematics.

Sometimes our conversations would last well into the night. Once, the auditorium we were in was locked by the custodian who couldn't fathom that there would be someone inside at such a late hour. And we must have been so deep into our conversation that we didn't hear the turning of the key. Fortunately, the auditorium was on the ground floor, and we managed to escape through a window.

The year was 1984, my senior year at high school. I had to decide which university to apply to. Moscow had many schools, but there was only one place to study pure math: Moscow State University, known by its Russian abbreviation MGU, for *Moskovskiy Gosudarstvenny Universitet*. Its famous *Mekh-Mat*, the Department of Mechanics and Mathematics, was the flagship mathematics program of the USSR.

Entrance exams to colleges in Russia are not like the SAT that American students take. At *Mekh-Mat* there were four: written math, oral math, literature essay composition, and oral physics. Those who, like me, graduated from high school with highest honors (in the Soviet Union we were then given a gold medal) would be automatically accepted after getting a 5, the highest grade, at the first exam.

I had by then progressed far beyond high school math, and so it looked like I would sail through the exams at MGU.

But I was too optimistic. The first warning shot came in the form of a letter I received from a correspondence school with which I had studied. This school had been organized some years earlier by, among others, Israel Gelfand, the famous Soviet mathematician (we will talk much more about him later). The school intended to help those students who, like me, lived outside of major cities and did not have access to special mathematical schools. Every month, participating students would receive a brochure elucidating the material that was studied in school and going a little beyond. It also contained some problems, more difficult than those discussed at school, which a student was supposed to solve and mail back. Graders (usually undergrads of Moscow University) read those solutions and returned them, marked, to the students. I was enrolled in this school for three years, as well as in another school, which was more physics-oriented. It was a helpful resource for me, though the material was pretty close to what was studied at school (unlike the stuff I was studying privately with Evgeny Evgenievich).

The letter I received from this correspondence school was short: “If you would like to apply to Moscow University, stop by our office, and we will be happy to give you advice,” and it gave the address on the campus of MGU and the visiting hours. Shortly after receiving the letter, I took the two-hour train ride to Moscow. The school’s office was a big room with a bunch of desks and a number of people working, typing, and correcting papers. I introduced myself, produced my little letter, and was immediately led to a diminutive young woman, in her early thirties.

“What’s your name?” she said by way of greeting.

“Eduard Frenkel.” (I used the Russian version of Edward in those days.)

“And you want to apply to MGU?”

“Yes.”

“Which department?”

“*Mekh-Mat*.”

“I see.” She lowered her eyes and asked:

“And what’s your nationality?”

I said, “Russian.”

“Really? And what are your parents’ nationalities?”

“Well… My mother is Russian.”

“And your father?”

“My father is Jewish.”

She nodded.

This dialogue might sound surreal to you, and as I am writing it now, it sounds surreal to me too. But in the Soviet Union circa 1984 – remember Orwell?* – it was not considered bizarre to ask someone what his or her “nationality” was. In the interior passport that all Soviet citizens had to carry with them, there was in fact a special line for “nationality.” It came after (1) first name, (2) patronymic name, (3) last name, and (4) the date of birth. For this reason, it was called *pyataya grafa*, “the fifth line.” Nationality was also recorded in one’s birth certificate, as were the nationalities of the parents. If their nationalities were different, as in my case, then the parents had a choice of which nationality to give to their child.

For all intents and purposes, the fifth line was a code for asking whether one was Jewish or not. (People of other nationalities, such as Tatars and Armenians, against whom there were prejudices and persecution – though not nearly at the same scale as against the Jews – were also picked up this way.) My fifth line said that I was Russian, but my last name – which was my father’s last name, and clearly sounded Jewish – gave me away.

It is important to note that my family was not religious at all. My father was not brought up in a religious tradition, and neither was I. Religion in the Soviet Union was in fact all but non-existent in those days. Most Christian Orthodox churches were destroyed or closed. In the few existing churches, one could typically only find a few old *babushkas* (grandmothers), such as my maternal grandmother. She occasionally attended service at the only active church in my hometown. There were even fewer synagogues. There were none in my hometown; in Moscow, whose population was close to 10 million, officially there was only one synagogue.[1] Going to a service in a church or a synagogue was dangerous: one could be spotted by special plain-clothed agents and would then get in a lot of trouble. So when someone was referred to as being Jewish, it was meant not in the sense of religion but rather in the sense of ethnicity, of "blood."

Even if I hadn't been using my father's last name, my Jewish origin would be picked up by the admissions committee anyway, because the application form specifically asked for the full names of both parents. Those full names included patronymic names; that is, the first names of the grandparents of the applicant. My father's patronymic name is Joseph, which sounded unmistakably Jewish in the Soviet Union of that era, so this was another way to find out (if his last name hadn't given me away). The system was set up in such a way that it would flag those who were at least one-quarter Jewish.

Having established that by this definition I was a Jew, the woman said, "Do you know that Jews are not accepted to Moscow University?"

"What do you mean?"

"What I mean is that you shouldn't even bother to apply. Don't waste your time. They won't let you in."

I didn't know what to say.

"Is that why you sent me this letter?"

"Yes. I'm just trying to help you."

I looked around. It was clear that everyone in the office was aware of what this conversation was about, even if they weren't listening closely. This must have already happened dozens of times, and everybody seemed used to it. They all averted their eyes, as if I were a terminally ill patient. My heart sank.

I had encountered anti-Semitism before, but at a personal, not institutional, level. When I was in fifth grade, some of my classmates took to taunting me with *evrey, evrey* ("Jew, Jew"). I don't think they had any idea what this meant (which was clear from the fact that some of them confused the word *evrey*, which meant "Jew," with *evropeyets*, which meant "European") – they must have heard anti-Semitic remarks from their parents or other adults. (Unfortunately, anti-Semitism was deeply rooted in the Russian culture.) I was strong enough and lucky enough to have a couple of true friends who stood by me, so I was never actually beaten up by these bullies, but this was an unpleasant experience. I was too proud to tell the teachers or my parents, but one day a teacher overheard and intervened. As a result, those kids were immediately called to the principal, and the taunting stopped.

My parents had heard of the discrimination against Jews in entrance exams to universities, but somehow they did not pay much attention to this. In my hometown, there weren't many Jews to begin with, and all the purported discrimination cases my parents had heard of concerned programs in physics. A typical argument went that Jews weren't accepted there because the studies in such a program were related to nuclear research and hence to national defense and state secrets; the government didn't want Jews in those areas because Jews could emigrate to Israel or somewhere else. By this logic, there shouldn't have been a reason to care about those who studied pure math. Well, apparently, someone did.

Everything about my conversation at MGU was strange. And I am not just talking about the Kafkaesque aspect of it. It is possible to conclude that the woman I talked to simply tried to help me and other students by warning us of what's going to happen. But could this really be the case? Remember, we are talking about 1984, when the Communist Party and the KGB still tightly controlled all aspects of life in the Soviet Union. The official policy of the state was that all nationalities were equal, and publicly suggesting otherwise would put one in danger. Yet, this woman calmly talked about this to me, a stranger she had just met, and she didn't seem to be worried about being overheard by her colleagues.

Besides, the exams at MGU were always scheduled one month ahead of all other schools. Therefore, students who were failed at MGU would still have a chance to apply elsewhere. Why would someone try to convince them not even to try? It sounded like some powerful forces were trying to scare me and other Jewish students away.

But I would not be deterred. After talking about all this at great length, my parents and I felt that I had nothing to lose. We decided that I would apply to MGU anyway and just hope for the best.

The first exam, at the beginning of July, was a written test in mathematics. It always consisted of five problems. The fifth problem was considered deadly and unsolvable. It was like the fifth element of the exam. But I was able to solve all problems, including the fifth. Aware as I was of the strong likelihood that whoever graded my exam could be biased against me and would try to find gaps in my solutions, I wrote everything out in excruciating detail. I then checked and double-checked all my arguments and calculations to make sure that I hadn't made any mistakes. Everything looked perfect! I was in an upbeat mood on the train ride home. The next day I told Evgeny Evgenievich my solutions, and he confirmed that everything was correct. It seemed like I was off to a good start.

My next exam was oral math. It was scheduled for July 13, which happened to be a Friday.

I remember very clearly many details about that day. The exam was scheduled for the early afternoon, and I took the train from home with my mother that morning. I entered the room at MGU a few minutes before the exam. It was a regular classroom, and there were probably between fifteen and twenty students there and four or five examiners. At the start of the exam each of us had to draw a piece of paper from a big pile on the desk at the front of the room. Each paper had two questions written on it, and it was turned blank side up. It was like drawing a lottery ticket, so we called this piece of paper *bilet*, ticket. There were perhaps one hundred questions altogether, all known in advance. I didn't really care which ticket I would draw as I knew this material inside-out. After drawing the ticket, each student had to sit down at one of the desks and prepare the answer, using only the provided blank sheets of paper.

The questions on my ticket were: (1) a circle inscribed in a triangle and the formula for the area of the triangle using its radius; and (2) derivative of the ratio of two functions (the formula only). I was so ready for these questions, I could have answered them in my sleep.

I sat down, wrote down a few formulas on a sheet of paper, and collected my thoughts. This must have taken me about two minutes. There was no need to prepare more; I was ready. I raised my hand. There were several examiners present in the room, and they were all waiting for the students to raise their hands, but, bizarrely, they ignored me, as if I did not exist. They looked right through me. I was sitting with my hand raised for a while: no response.

Then, after ten minutes or so, a couple of other kids raised their hands, and as soon as they did, the examiners rushed to them. An examiner would take a seat next to a student and listen to him or her answer the questions. They were quite close to me, so I could hear them. The examiners were very polite and were mostly nodding their heads, only occasionally asking follow-up questions. Nothing out of the ordinary. When a student finished answering the questions on the ticket (after ten minutes or so), the examiner would give him or her one additional problem to solve. Those problems seemed rather simple, and most students solved them right away. And that was it!

The first couple of students were already happily gone, having obviously earned a 5, the highest grade, and I was still sitting there. Finally, I grabbed one of the examiners passing by, a young fellow who seemed like he was a fresh Ph.D., and asked him pointedly: "Why aren't you talking to me?" He looked away and said quietly: "Sorry, we are not allowed to talk to you."

An hour or so into the exam, two middle-aged men entered the room. They briskly walked up to the table at the front of the room and presented themselves to the guy who was sitting there. He nodded and pointed at me. It became clear that these were the people I'd been waiting for: my inquisitors.

They came up to my desk and introduced themselves. One was lean and quick, the other slightly overweight and with a big mustache.

"OK," the lean man said – he did most of the talking – "what have we got here? What's the first question?"

"The circle inscribed in a triangle and..."

He interrupted me: "What is the definition of a circle?"

He was quite aggressive, which was in sharp contrast to how other examiners treated students. Besides, the other examiners never asked anything before the student had a chance to fully present their answer to the question on the ticket.

I said, "A circle is the set of points on the plane equidistant from a given point."

This was the standard definition.

"Wrong!" declared the man cheerfully.

How could this possibly be wrong? He waited for a few seconds and then said, "It's the set of *all* points on the plane equidistant from a given point."

That sounded like excessive parsing of words – the first sign of trouble ahead.

"OK," the man said, "What is the definition of a triangle?"

After I gave that definition, and he thought about it, no doubt trying to see if he could do some more nit-picking, he continued: "And what's the definition of a circle inscribed in a triangle?"

That led us to the definition of the tangent line, then just "a line," and that led to other things, and soon he was asking me about Euclid's fifth postulate about the uniqueness of parallel lines, which wasn't even part of the high school program! We were talking about issues that were not even close to the question on the ticket and far beyond what I was supposed to know.

Every word I said was questioned. Every concept had to be defined, and if another concept was used in the definition, then I was immediately asked to define it as well.

Needless to say, if my last name were Ivanov, I would never be asked any of these questions. In retrospect, the prudent course of action on my part would have been to protest right away and tell the examiners that they were out of line. But it's easy to say this now. I was sixteen years old, and these men were some twenty-five years my senior. They were the officials administering an exam at Moscow State University, and I felt obligated to answer their questions as best I could.

After nearly an hour-long interrogation, we moved to the second question on my ticket. By then, other students had left, and the auditorium was empty.

Apparently, I was the only student in that room who required "special care." I guess they tried to place Jewish students so that there would be no more than one or two of them in the same room.

The second question asked me to write the formula for the derivative of the ratio of two functions. I was not asked to give any definitions or proofs. The question said specifically, the formula only. But of course, the examiners insisted that I explain to them a whole chapter of the calculus book.

"What is the definition of derivative?"

The standard definition I gave involved the concept of limit.

"What is the definition of limit?" Then "What is a function?" and on and on it went again.

The question of ethnic discrimination at the entrance exams to MGU has been the subject of numerous publications. For example, in his insightful article[2] in the *Notices of the American Mathematical Society*, mathematician and educator Mark Saul used my story as an example. He aptly compared my exam to the Red Queen interrogating Alice in *Alice in Wonderland*. I knew the answers, but in this game, in which everything I said was turned against me, I couldn't possibly win.

In another article[3] on this subject in the *Notices*, journalist George G. Szpiro gave this account:

> Jews – or applicants with Jewish-sounding names – were singled out at the entrance exams for special treatment.... The hurdles were raised in the oral exam. Unwanted candidates were given "killer questions" that required difficult reasoning and long computations. Some questions were impossible to solve, were stated in an ambiguous way, or had no correct answer. They were not designed to test a candidate's skill but meant to weed out "undesirables." The grueling, blatantly unfair interrogations often lasted five or six hours, even though by decree they should have been limited to three and a half. Even if a candidate's answers were correct, reasons could always be found to fail him. On one occasion a candidate was failed for answering the question "what is the definition of a circle?" with "the set of points equidistant to a given point." The correct answer, the examiner said, was "the set of all points equidistant to a given point." On another occasion an answer to the same question was deemed incorrect because the candidate had failed to stipulate that the distance had to be nonzero. When asked about the solutions to an equation, the answer "1 and 2" was declared wrong, the correct answer being, according to an examiner, "1 or 2." (On a different occasion, the same examiner told another student the exact opposite: "1 or 2" was considered wrong.)

But back to my exam. Another hour and a half had gone by. Then one of the examiners said:

"OK, we are done with the questions. Here is a problem we want you to solve."

The problem he gave me was pretty hard. The solution required the use of the so-called Sturm principle, which was not studied in school.[4] However, I knew about it from my correspondence courses, so I was able to solve it. As I was working my way through the final calculations, the examiner came back.

"Are you done yet?"

"Almost."

He looked at my writings and no doubt saw that my solution was correct and that I was just finishing my calculations.

"You know what," he said, "let me give you another problem."

Curiously, the second problem was twice as hard as the first one. I was still able to solve it, but the examiner again interrupted me halfway through.

"Not done yet?" he said, "Try this one."

If this were a boxing match, with one of the boxers pressed in the corner, bloodied, desperately trying to hold his own against the barrage of punches falling on him (many of them below the belt, I might add), that would be the equivalent of the final, deadly blow. The problem looked innocent enough at first glance: given a circle and two points on the plane outside the circle, construct another circle passing through those two points and touching the first circle at one point.

But the solution is in fact quite complicated. Even a professional mathematician would not necessarily be able to solve it right away. One must either use a trick called inversion or follow an elaborate geometric construction. Neither method was studied in high school, and hence this problem should not have been allowed on this exam.

I knew about inversion, and I realized that I could apply it here. I started to work on the problem, but a few minutes later my interrogators came back and sat down next to me. One of them said:

"You know, I've just talked to the deputy chairman of the admissions committee and I told him about your case. He asked me why we are still wasting our time… Look," he pulled out an official looking form with some notes scribbled on it – this was the first time I saw it. "On the first question on your ticket, you did not give us a complete answer, you didn't even know the definition of a circle. So we have to put a minus. On the second question, your knowledge was also shaky, but OK, we give you minus plus. Then you couldn't completely solve the first problem, did not solve the second problem. And on the third? You haven't solved it either. See, we have no choice but to fail you."

I looked at my watch. More than four hours had passed by since the beginning of the exam. I was exhausted.

"Can I see my written exam?"

The other man went back to the main table and brought my exam. He put it in front of me. As I was turning the pages, I felt like I was in a surrealistic movie. All answers were correct, all solutions were correct. But there were many comments. They were all made in pencil – so that they could be easily erased, I guess – but they were all ridiculous, like someone was playing a practical joke on me. One of them still stands out in my mind: in the course of a calculation, I wrote "$\sqrt{8} > 2$." And there was a comment next to it: "not proved." Really? Other comments were no better. And what grade did they give me, for all five problems solved, with all correct answers? Not 5, not 4. It was a 3, the Russian equivalent of a C. They gave me a C for this?

I knew it was over. There was no way I could fight this system. I said, "All right."

One of the men asked, "Aren't you going to appeal?"

I knew that there was an appeal board. But what would be the point? Perhaps, I could raise my grade on the written exam from 3 to 4, but appealing the result of the oral exam would be more difficult: it would be their word against mine. And even if I could raise the grade to 3, say, then what? There were still two more exams left at which they could get me.

Here is what George Szpiro wrote in the *Notices*:[5]

> And if an applicant, against all odds, managed to pass both the written and the oral test, he or she could always be failed on the required essay on Russian literature with the set phrase "the theme has not been sufficiently elaborated." With very rare exceptions, appeals against negative decisions had no chance of success. At best they were ignored, at worst the applicant was chastised for showing "contempt for the examiners."

A bigger question was: did I really want to enroll in a university that did everything in its power to prevent me from being there? I said, "No. Actually, I'd like to withdraw my application."

Their faces lit up. No appeal meant less hassle for them, less potential for trouble.

"Sure," the talkative one said, "I'll get your stuff for you right away."

We walked out of the room and entered the elevator. The doors closed. It was just the two of us. The examiner was clearly in a good mood. He said, "You did great. A really impressive performance. I was wondering: did you go to a special math school?"

I grew up in a small town; we didn't have special math schools.

"Really? Perhaps your parents are mathematicians?"

No, they are engineers.

"Interesting… It's the first time I have seen such a strong student who did not go to a special math school."

I couldn't believe what he was saying. This man had just failed me after an unfairly administered, discriminatory, grueling, nearly five-hour long exam. For all I knew, he killed my dream of becoming a mathematician. A 16-year-old student, whose only fault was that he came from a Jewish family… And now this guy was giving me compliments and expecting me to open up to him?!

But what could I do? Yell at him, punch him in the face? I was just standing there, silent, stunned. He continued: "Let me give you advice. Go to the Moscow Institute of Oil and Gas. They have an applied mathematics program, which is quite good. They take students *like you* there."

The elevator doors opened, and a minute later he handed me my thick application folder, with a bunch of my school trophies and prizes oddly sticking out of it.

"Good luck to you," he said, but I was too exhausted to respond. My only wish was to get the hell out of there!

And then I was outside, on the giant staircase of the immense MGU building. I was breathing fresh summer air again and hearing the sounds of the big city coming from a distance. It was getting dark, and there was almost no one around. I immediately spotted my parents, who were waiting anxiously for me on the steps this whole time. By the look on my face, and the big folder I was holding in my hands, they knew right away what had happened inside.

[*] This was one year before Mikhail Gorbachev came to power in the Soviet Union, and another couple of years before he launched his *perestroika*. The totalitarian Soviet regime in 1984 was in many ways a haunting facsimile of George Orwell's prescient book.

Chapter 4

Kerosinka

That night, after the exam, my parents and I came home quite late. We were still in the state of initial shock and disbelief about what happened.

This was a gut-wrenching experience for both of my parents. I have always been very close to them, and they always gave me unconditional love and support. They never pushed me to study harder or choose a particular profession, but they encouraged me to pursue my passion. And of course they were proud of my accomplishments. They were devastated by what had happened at my exam, both because of the sheer unfairness of it and because they were unable to do anything to protect their son.

Thirty years earlier, in 1954, my father's dream of becoming a theoretical physicist had been shattered just as ruthlessly, for a different reason. Like millions of innocent people, his father, my grandfather, had been a victim of Joseph Stalin's persecution. He was arrested in 1948 on bogus charges that he wanted to blow up the big automobile plant in Gorky (now Nizhny Novgorod), where he worked as the head of supplies. The only "evidence" presented in his indictment was that he had in his possession at the time of his arrest a box of matches. He was sent to a hard-labor camp at a coal mine in the northern part of Russia, part of the Gulag Archipelago that Alexander Solzhenitsyn and other writers described so vividly years later. He was deemed an "enemy of the people," and my father was therefore a "son of the enemy of the people."

My dad was obligated to write this on his application to the physics department of Gorky University. Even though he finished high school with honors and was supposed to be accepted automatically, he was failed at the interview, whose sole purpose was to screen out the relatives of the "enemies of the people." My dad was forced to go to an engineering school instead. (Like other prisoners, his father was rehabilitated and released by Nikita Khrushchev's decree in 1956, but by then it was too late to undo the injustice.)

Now, thirty years later, his son had to go through a similar experience.

But there was no time for self-pity. We had to decide quickly what to do next, and the first question was which school I should apply to. All of them held their exams at the same time, in August, about two weeks later, and I could only apply to one.

The next morning my dad woke up early and went back to Moscow. He took the recommendation from my examiner at MGU seriously. It sounded like the examiner was trying to help me, perhaps, as some partial compensation for the injustice he had done. So when my father arrived in Moscow, he went straight to the admissions office at the Institute of Oil and Gas.* Somehow my dad managed to find someone there willing to talk to him privately and described my situation. The fellow said that he was aware of the anti-Semitism at MGU but said the Institute of Oil and Gas had none of this. He went on to say that the level of applicants to their applied mathematics program was quite high due to a large number of students like me, who were not accepted at MGU. The entrance exam would be no cakewalk. But, he said, "if your son is as bright as you say he is, he will be admitted. There is no discrimination against Jews at the entrance exams here."

"I have to warn you, though," he said at the end of the conversation, "Our post-graduate studies are handled by different people, and I think your son probably won't be accepted to the grad school."

But that was something to worry about in five years, too far ahead.

My father went to a couple of other schools in Moscow with applied math programs, but there was nothing like the attitude he found at the Institute of Oil and Gas. So when he came back home that evening and told me and my mom the news, we decided right away that I would apply to the Institute of Oil and Gas, to their applied mathematics program.

The Institute was one of a dozen schools in Moscow preparing technicians for various industries, such as the Institute of Metallurgy and the Institute of Railway Engineers (in the Soviet Union, many colleges were called "institutes"). From the late 1960s, anti-Semitism at MGU "created a market for placements in mathematics for Jewish students," writes Mark Saul in his article.[1] The Institute of Oil and Gas "began to cater to these markets, benefiting from the anti-Semitic policies of other universities to get highly qualified students." Mark Saul explains:

> Its nickname, Kerosinka, reflected [their] pride and cynicism. A kerosinka is a kerosene-burning space heater, a low-tech but effective response to adversity. The students and graduates of the institute quickly became known as "kerosineshchiks," and the school became a haven for Jewish students with a passion for mathematics.
>
> How did fate choose Kerosinka as the repository of so much talent? This question is not easy to answer. We know that there were other institutions that benefited from the exclusion of Jews from MGU. We also know that the establishment of this exclusionary policy was a conscious act, which probably met with some resistance at first. It may have been easier for some institutions to continue accepting Jewish students than for them to institute a new policy. But once the phenomenon grew and there was a cadre of Jewish students at Kerosinka, why was it tolerated? There are dark whispers of a plot by the secret police (KGB) to keep the Jewish students under surveillance in one or two places. But some of the motivation may have been more positive: the administration of the institute may have seen a good department developing and done what it could to preserve the phenomenon.

I believe the last sentence is more accurate. The President (or Rector, as he was called) of the Institute of Oil and Gas, Vladimir Nikolaevich Vinogradov, was a clever administrator known for recruiting professors who were engaged in innovative teaching and research and for using new technologies in the classrooms. He instituted the policy that all exams (including the entrance exams) were given *in writing*. Of course, there might still be some opportunity for abuse even with written tests (as was the case with my written exam at MGU), but the policy would prevent the kind of debacle that happened at my oral examination at MGU. I would not be surprised if it was Vinogradov's personal decision not to discriminate against Jewish applicants, and if so, it must have required some good will, and perhaps even some courage, on his part.

As predicted, there seemed to be no discrimination at the entrance exams. I was accepted after the first exam (written math), on which I got a 5, that is, an A (gold medalists were accepted outright if they got an A at the first exam). In a bizarre twist, this 5 did not come easy to me because apparently some of my solutions were entered incorrectly into the automated grading system, and as a result my grade was initially recorded as 4, or B. I had to go through the appeals process, which meant waiting in line for hours, with all kinds of bad thoughts swirling in my head. But once I got in to speak with the appeals committee, the error was found and fixed swiftly, an apology was offered, and my entrance exams saga came to a close.

On September 1, 1984, the school year began, and I met my new classmates. Only fifty students were accepted every year to this program (in contrast, at *Mekh-Mat* the number was close to 500). Many of my fellow students went through the same experience as I did. These were some of the brightest, most talented math students around.

Everybody, except for me and another student, Misha Smolyak from Kishinev who became my roommate at the dorm, were from Moscow. Those who lived outside of Moscow could apply only if they had graduated from high school with a gold medal, which fortunately I had.

Many of my fellow students had studied at the best Moscow schools with special math programs: schools No. 57, No. 179, No. 91, and No. 2. Some of them went on to become professional mathematicians and now work as professors at some of the best universities in the world. Just in my class, we had some of the best mathematicians of our generation: Pasha Etingof, now professor at MIT; Dima Kleinbock, professor at Brandeis University; and Misha Finkelberg, professor at the Higher School of Economics in Moscow. It was a very stimulating environment.

Mathematics was taught at Kerosinka at a high level, and basic courses, such as analysis, functional analysis, and linear algebra, were taught at the same level of rigor as at MGU. But courses in other areas of pure math, such as geometry and topology, were not available. Kerosinka only offered the applied mathematics program, so our education was geared toward concrete applications, particularly, to oil and gas exploration and production. We had to take quite a few courses of more applied orientation: optimization, numerical analysis, probability, and statistics. There was also a large computer science component.

I was glad that I had the opportunity to be exposed to these applied math courses. This taught me that there isn't really a sharp distinction between "pure" and "applied" math; good-quality applied math is always based on sophisticated pure math. But, however useful this experience was, I could not forget my true love. I knew I had to find a way to learn the pure math subjects that were not offered at Kerosinka.

The solution presented itself as I became friends with the other students, including those who went to the prestigious special math schools in Moscow. We exchanged our stories. Those who were Jewish (according to the standards I described earlier) were also failed at the exams, as ruthlessly as I was, while all of their classmates who were not Jewish were accepted to MGU without any problems. Through these other students, they knew what was happening at the *Mekh-Mat*, which courses were good, and where and when the lectures were held. So my second week at Kerosinka, my classmate (I think it was Dima Kleinbock) came up to me: "Hey, we are going to Kirillov's course at MGU. Wanna come with us?"

Kirillov was a famous mathematician, and of course I wanted to attend his lectures. But I had no idea how this would be possible. The grand building of MGU was heavily guarded by police. One needed to have a special ID to get in.

"No worries," my classmate said, "we'll scale the fence."

That sounded dangerous and exciting, so I said, "Sure."

The fence on the side of the building was quite high, easily twenty feet, but at one point the metal was bent, and it was possible to sneak in. Then what? We entered the building through a side door and after following some long corridors ended up in the kitchen. From there, through the kitchen, trying not to attract too much attention of the staff working there, to the cafeteria, and then to the main entrance hall. Elevator to the fourteenth floor, where the auditorium was.

Alexander Alexandrovich Kirillov (or San Sanych, as he was affectionately called) is a charismatic lecturer and a great human being, whom I got to know quite well years later. I think he was teaching a standard undergraduate course on representation theory along the lines of his well-known book. He also had a seminar for graduate students, which we attended as well.

We got away with this thanks to Kirillov's good heart. His son Shurik (now professor at the Stony Brook University) studied at the special math school No. 179 together with my classmates Dima Kleinbock and Syoma Hawkin. Needless to say, San Sanych knew about the situation with admissions at MGU. He told me many years later that there was nothing he could do about this – they wouldn't let him anywhere near the admissions committee, which was largely staffed with the Communist Party apparatchiks. So all he could do was let us sneak into his classes.

Kirillov did all he could to make Kerosinka students coming to his lectures feel welcome. One of the best memories of my first college year was coming to his lively lectures and seminars. I also attended a seminar given by Alexander Rudakov, which was also a great experience.

In the meantime, I was learning whatever math I could learn at Kerosinka. I was living in a dorm but coming home for the weekends, and I was still meeting Evgeny Evgenievich every couple of weeks. He advised me on what books to read, and I reported to him on my progress. But I was quickly reaching the point where to maintain my momentum, as well as the motivation for it, I would need an advisor with whom I would meet more regularly and not only learn from, but also get a problem to work on. Because I was not at *Mekh-Mat*, I could not take advantage of the vast resources that it had to offer. And I was too shy to come up to someone like A.A. Kirillov and ask him to work with me individually, or give me a problem to work on. I felt like an outsider. By the spring semester of 1986 (my second year at Kerosinka), complacency and stagnation were beginning to set in. With all the odds stacked against me, I started to doubt that I could fulfill my dream of becoming a mathematician.

* At the time, it was known as the Gubkin Institute of Petrochemical and Gas Industry (it was named after the long-time head of the Ministry of Oil and Gas in the USSR, I.M. Gubkin). After I became a student there, it was renamed Gubkin Institute of Oil and Gas, and later, Gubkin University of Oil and Gas.

Chapter 5

Threads of the Solution

I was beginning to despair when, one day, during a break in the lecture at Kerosinka, one of our most respected math professors, Alexander Nikolaevich Varchenko, approached me in the corridor. Varchenko is a former student of Vladimir Arnold, one of the leading Soviet mathematicians, and he is a world-class mathematician himself.

"Would you be interested in working on a math problem?" he asked.

"Yes, of course," I said, "What kind of problem?" as if I would not have been happy to do just anything.

"There is this question that came up in my research, which I think is a good problem to give to a bright student like you. The expert on this matter is Dmitry Borisovich Fuchs." That was the name of a famous mathematician, which I had heard before. "I have already spoken to him, and he has agreed to supervise a student's research on this topic. Here is his phone number. Give him a call, and he'll tell you what to do."

It is quite common for experienced mathematicians like Varchenko to encounter all kinds of unsolved mathematical problems in their research. If Varchenko's problem had been closely tied to his own research program, he might have tried to solve it himself. But no mathematician does everything alone, so mathematicians often delegate some of such unsolved problems (typically, the ones they consider to be simpler) to their students. Sometimes a problem might be outside of the professor's immediate interests, but he or she might nonetheless be curious about it, as was the case with my problem. That's why Varchenko enlisted Fuchs, an expert in this area, to supervise me. All in all, this was for the most part a typical "transaction" in the social workings of the mathematical world.

What was actually unusual was that Fuchs was not formally affiliated with teaching at any university. But for many years Fuchs had been, along with a number of other top mathematicians, trying to alleviate the effect of the discrimination against Jewish students by privately teaching young talented kids who were denied entry to MGU.

As part of those efforts, Fuchs was involved in what became known as "Jewish People's University," an unofficial evening school, where he and his colleagues gave courses of lectures to students. Some of those lectures had even been held at Kerosinka, although this was before my time.

The school had been organized by a courageous woman, Bella Muchnik Subbotovskaya, who was its heart and soul. Unfortunately, the KGB got on the case, alarmed that there were unauthorized gatherings of Jewish people. She was eventually called to the KGB and interrogated. Soon after that interview, she was killed by a truck under suspicious circumstances, which led many people to suspect that this was in fact a cold-blooded murder.[1] Without her at the helm, the school collapsed.

I came to Kerosinka two years after this tragic chain of events. Though the evening school did not exist anymore, there was still a small network of professional mathematicians who helped misfortunate outcasts like myself on an individual basis. They sought out promising students and gave them advice, encouragement, and in some cases, full-fledged mentoring and advising. This was the reason that Varchenko gave that problem to me, a student at Kerosinka, rather than a student at *Mekh-Mat*, where, through his connections, he could have easily found a student willing to take it up. This was also why Fuchs was willing to invest his personal time to supervise me.

And I am glad he did. Looking back, it is clear to me that without Fuchs' kindness and generosity, I would have never become a mathematician. I was studying math at Kerosinka and sitting in at the lectures at MGU, but by itself that was not enough. In fact, it is virtually impossible for students to do their own research without someone guiding their work. Having an advisor is absolutely essential.

At the time, though, all I knew was that I had in my hand the phone number of Fuchs, a renowned mathematician, and I was about to embark on a project supervised by him. This was unbelievable! I didn't know where this would end up, but I knew right away that something big had happened.

That evening, having mustered all my courage, I called Fuchs from a pay phone and explained who I was.

"Yes, I know," said Fuchs, "I have to give you a paper to read."

We met the next day. Fuchs had the physical appearance of a giant, not at all how I imagined him. He was very business-like.

"Here," he said, handing me an offprint of an article, "try to read this, and as soon as you see a word that you don't understand, call me."

I felt like he had just handed me the Holy Grail.

This was an article, a dozen pages long, which he had written some years earlier, on the subject of "braid groups." That evening I started reading it.

The preceding three years of studying with Evgeny Evgenievich and on my own were not spent in vain. Not only did I understand all the words in the title, I could make sense of the content as well. I decided to try to read the whole thing on my own. It was a matter of pride. I was already imagining how impressed Fuchs would be when I told him that I understood everything on my own.

I had heard of the "braid groups" before. These are excellent examples of groups, the concept we discussed in Chapter 2. Evgeny Evgenievich had introduced this concept in the context of symmetries, and so elements of the groups that we considered were symmetries of some object. For example, the circle group consisted of the symmetries of a round table (or any other round object), and the group of four rotations was the group of symmetries of a square table (or any other square object). Once we have the notion "group," we can look for other examples. It turns out there are many examples of groups that have nothing to do with symmetries, which was our motivation to introduce the concept of a group in the first place. This is actually a typical story. The creation of a mathematical concept may be motivated by problems and phenomena in one area of math (or physics, engineering, and so forth), but later it may well turn out to be useful and well adapted to other areas.

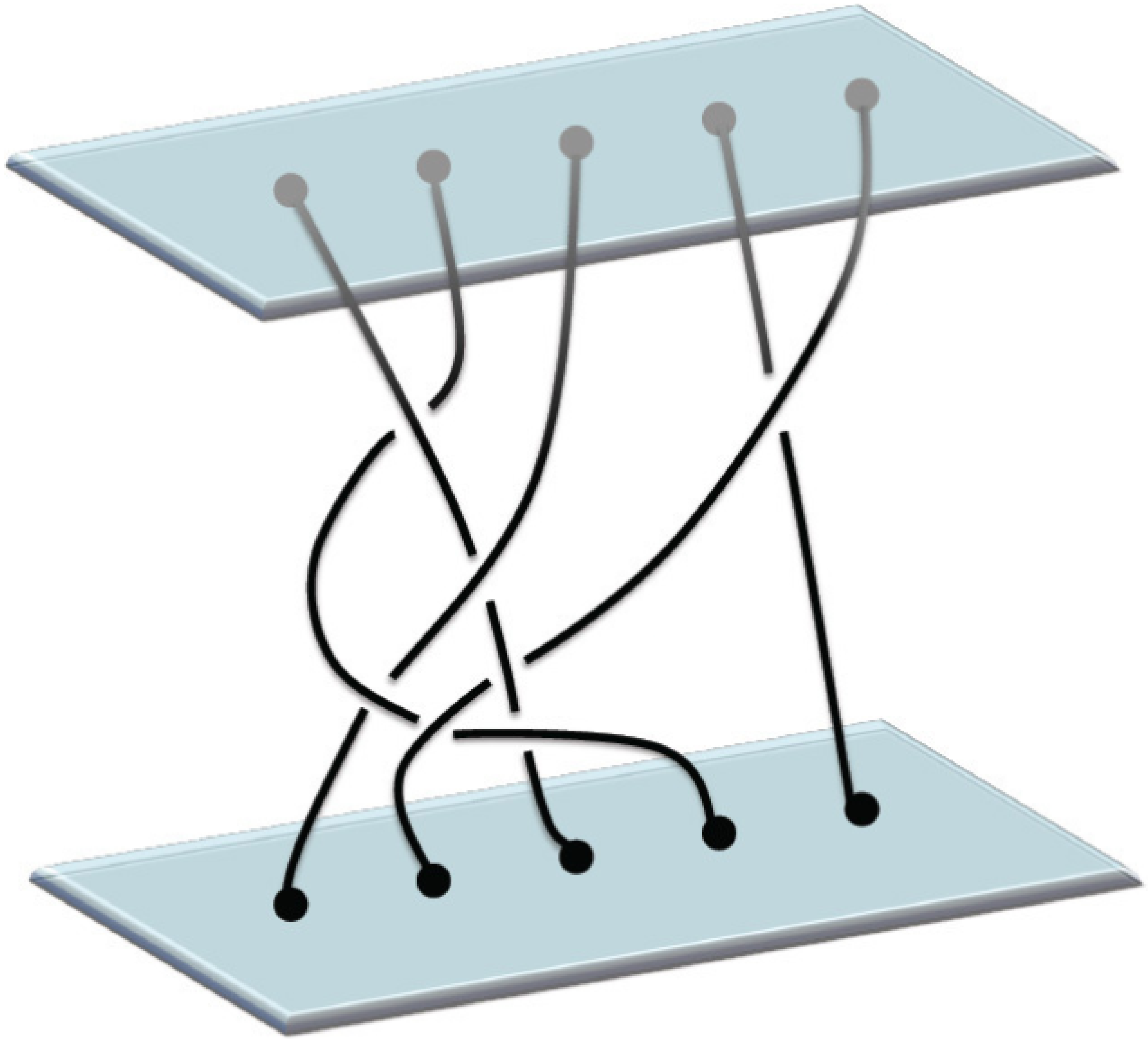
It turns out that many groups do *not* come from symmetries. And the braid groups are such groups.

I did not know yet about the real-world applications of braid groups to such areas as cryptography, quantum computing, and biology, which we will talk about later. But I was mesmerized by the innate beauty of these mathematical abstractions.

There is one braid group for each natural number $n = 1, 2, 3,...$ We can use those numbers to get a name for each braid group. In general, we call them B_n, and so for $n = 1$ we have a group called B_1, for $n = 2$ we have a group called B_2, and so on.

To describe the group B_n, we have to describe first its elements, as we did with the rotational symmetries of the round and square tables. The elements of the group B_n are the so-called *braids with n threads*, such as the one shown on the picture below, with $n = 5$. Imagine two solid, transparent plates with five nails in each, with one thread connecting each nail in one plate to one nail in the other. Since the plates are transparent, we can see each of the threads in its entirety. Each thread is allowed to weave around any other thread any way we like but is *not* allowed to get entangled with itself. Each nail must connect to exactly one thread. The positions of the plates are fixed once and for all.

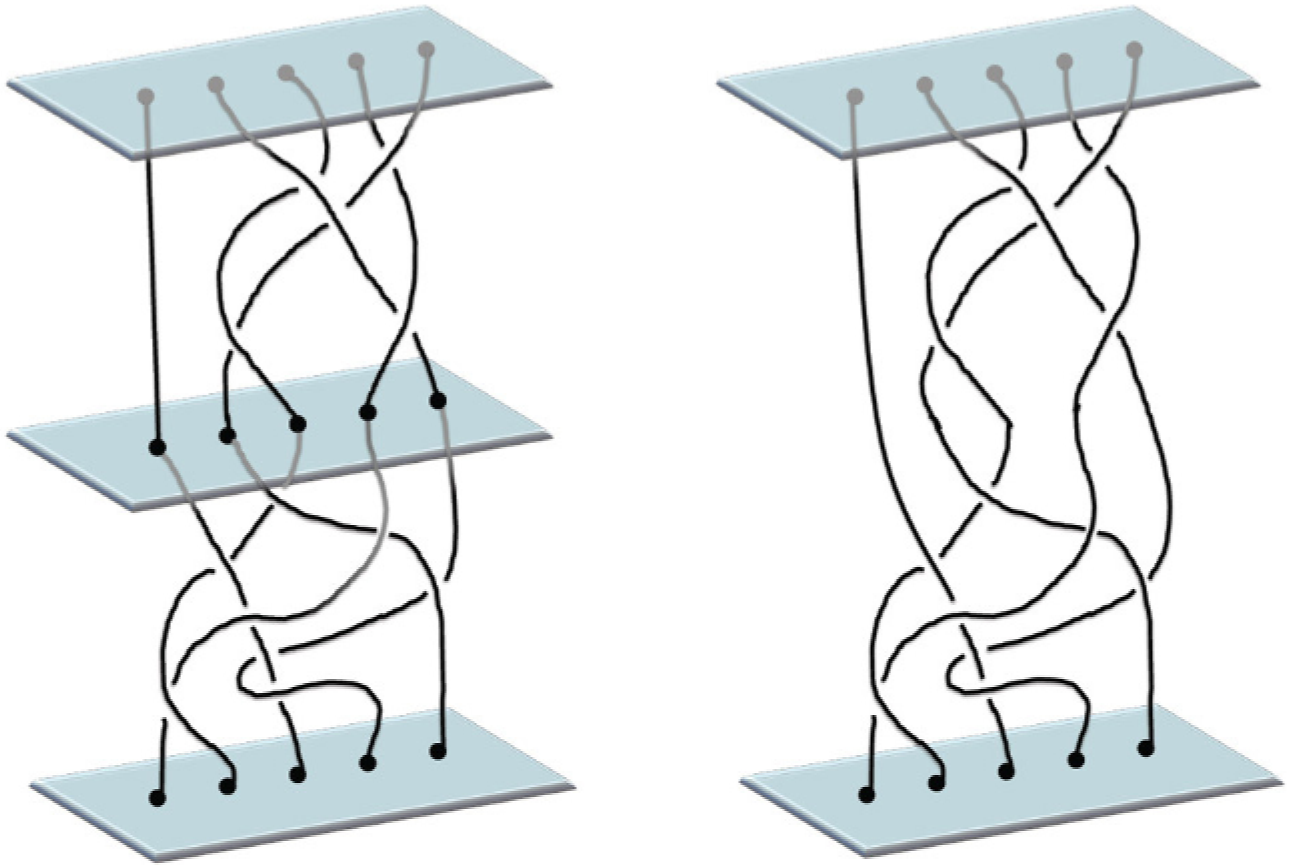
This whole thing – two plates and however many threads – constitutes a single braid, just as a car has four wheels, one transmission, four doors, and so forth. We are not considering those parts separately; we are focusing on the braid as a whole.



Those are the braids with n threads. Now we need to show that all braids with n threads form a group. This means that we need to describe how to make the composition of two such braids. In other words, for each pair of braids with n threads, we have to produce another braid with n threads, just as applying two rotations one after another gave us a third rotation. And then we will have to check that this composition satisfies the properties listed in Chapter 2.

So suppose we have two braids. In order to produce a new braid out of them, we put one of them on top of the other, aligning the nails, as shown on the picture. And then we remove the middle plates while connecting the upper threads to the lower ones attached to the matching nails.

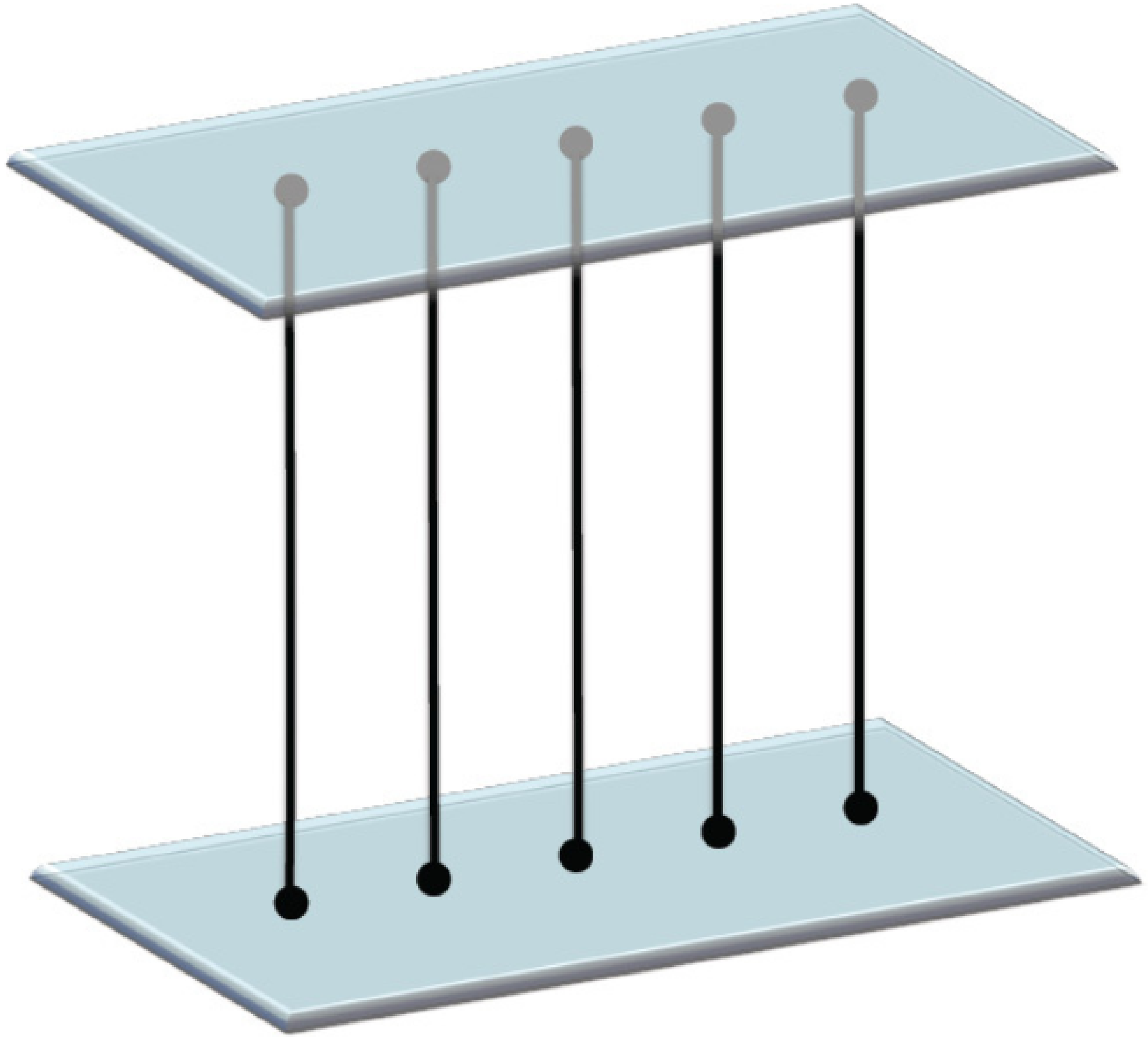
The resulting braid will be twice as tall, but this is not a problem. We'll just shorten the threads so that the resulting braid will have the same height as the original ones, while preserving the way the threads go around each other. *Voilà!* We started out with two braids and produced a new one. This is the rule of composition of two braids in our braid group.



Since a braid group does not come from symmetries, it is sometimes better to think of this operation not as "composition" (which was natural in the case of groups of symmetries), but as "addition" or "multiplication," similar to the operations that we perform on numbers. From this point of view, braids are like numbers – these are some "hairy numbers," if you will.

Given two whole numbers, we can add them to each other and produce a new number. Likewise, given two braids, we produce a new one by the rule described above. So we can think of this as the "addition" of two braids.

Now we need to check that this addition of braids satisfies all properties (or axioms) of a group. First, we need the identity element. (In the circle group, this was the point corresponding to the rotation by 0 degrees.) This will be the braid with all threads going straight down without any weaving as shown on the next picture. It is a kind of "trivial" braid, in which no braiding actually occurs, the same way rotation by 0 degrees makes no rotation at all.[2]



Next, we need to find the inverse braid of a given braid b (in the case of the circle group, this was the rotation by the same angle but in the opposite direction). It should be such that if we add this braid to the braid b, according to the rule described above, we will get the identity braid.

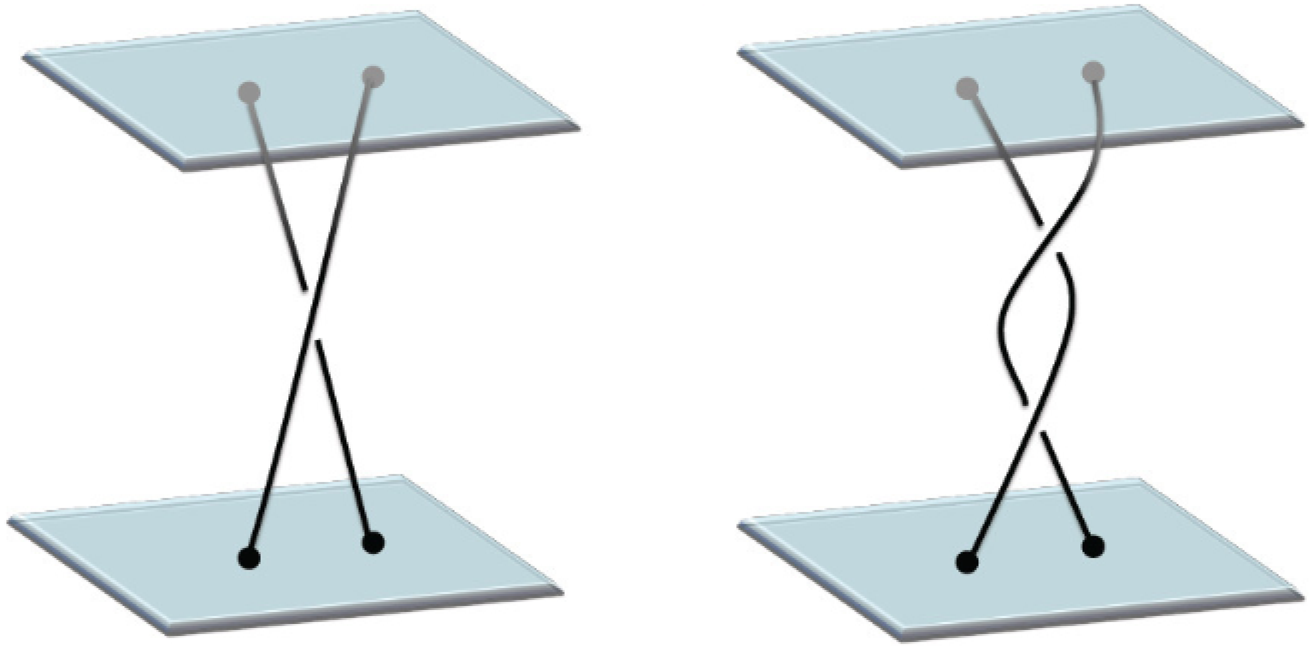
This inverse braid will be the reflection of b with respect to the bottom plate. If we compose it with the original braid according to our rule, we will be able to rearrange all threads so that the result will be the identity braid.

Here I need to make an important point, which up to now I have kind of swept under the rug: we will not distinguish the braids that can be obtained from one another by pulling the threads, stretching and shrinking them any way we like so long as we do not cut or resew the threads. In other words, the threads should be attached to the same nails, and we do not allow the threads to go through each other, but otherwise we can tweak them any way we like. Think of this as grooming our braid. When we do that, it will still be the same braid (only prettier!). It is in this sense that the addition of a braid and its mirror image is "the same" as the identity braid; it is not literally the same but becomes one after we tweak the threads.[3]

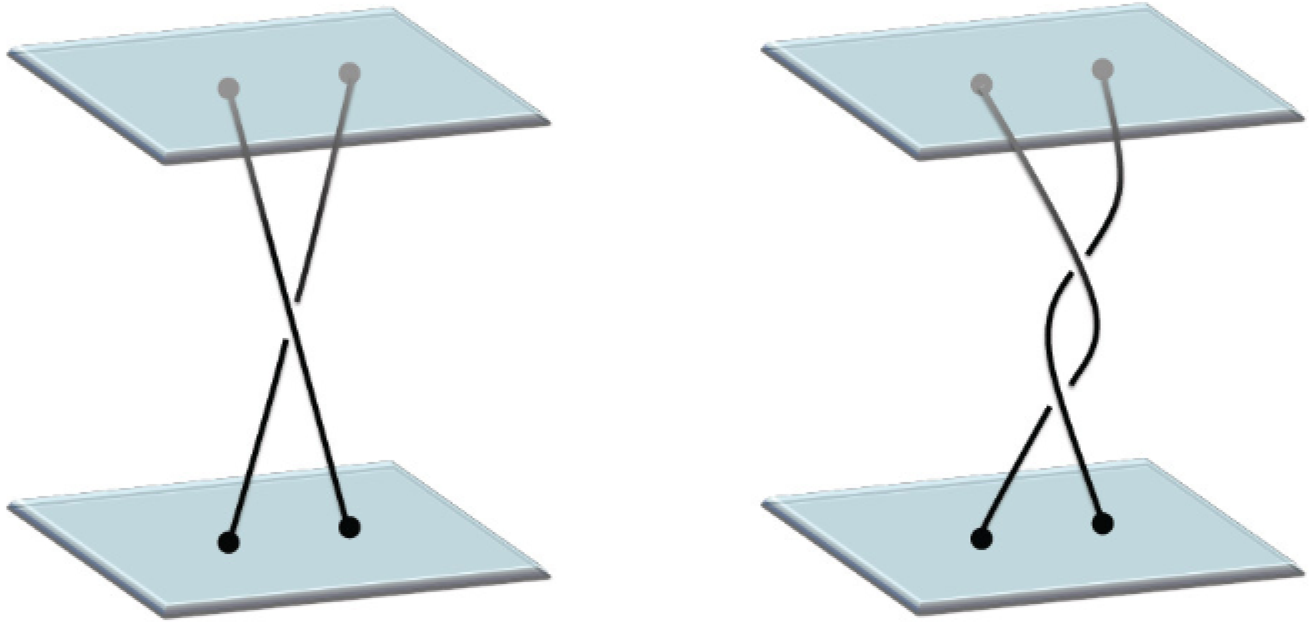
So we see now that the axioms of a group – composition (or addition), identity, and inverse – are satisfied. We have proved that braids with n threads form a group.[4]

To see what the braid groups are more concretely, let's look closely at the simplest one: the group B_2 of braids with two threads. (The group with B_1 with one thread has only one element, so there is nothing to discuss.[5]) We will assign to each such braid an integer N. By an integer, here I mean a natural number: 1, 2, 3,...; or 0; or a negative of a natural number: −1, −2, −3,...

First of all, to the identity braid we will assign the number 0. Second, if the thread starting at the left nail on the top plate goes underneath the other thread, then we assign to it 1. If it goes around it, then we assign to it 2, and so on, as shown on the pictures.



If this thread goes on top of the other thread, then we assign to the braid the negative number −1, if it goes around it as shown on the picture below, we assign to it −2, and so on.



Let's call the number assigned to a braid in this way the "number of overlaps." If we have two braids with the same number of overlaps, we can transform one into another by "tweaking" the threads. In other words, the braid is completely determined by the number of overlaps. So we have a one-to-one correspondence between braids with two threads and integers.

Here it is useful to note something we always take for granted: the set of all integers is itself a group! Namely, we have the operation of addition, the "identity element" is number 0, and for any integer N, its "inverse" is $-N$. Then all properties of a group listed in Chapter 2 are satisfied. Indeed, we have $N + 0 = N$ and $N + (-N) = 0$.

What we have just found is that the group of braids with two threads has the same structure as the group of integers.[6]

Now, in the group of integers the sum of two integers a and b is the same in two different orders:

$$a + b = b + a.$$

This is also so in the braid group B_2. Groups satisfying this property are called "commutative" or "abelian" (in honor of the Norwegian mathematician Niels Henrik Abel).

In a braid with 3 threads or more, the threads can be entangled among themselves in a much more complicated fashion than in a braid with only 2 threads. The knotting pattern can no longer be described merely by the numbers of overlaps (look at the above picture of a braid with 5 threads). The pattern in which the overlaps occur is also important. Furthermore, it turns out that the addition of two braids with 3 or more threads does depend on the order in which it is taken (that is to say, which of the two braids is on top in the picture above, describing the addition of braids). In other words, in the group B_n with $n = 3,4,5,...$ we have in general

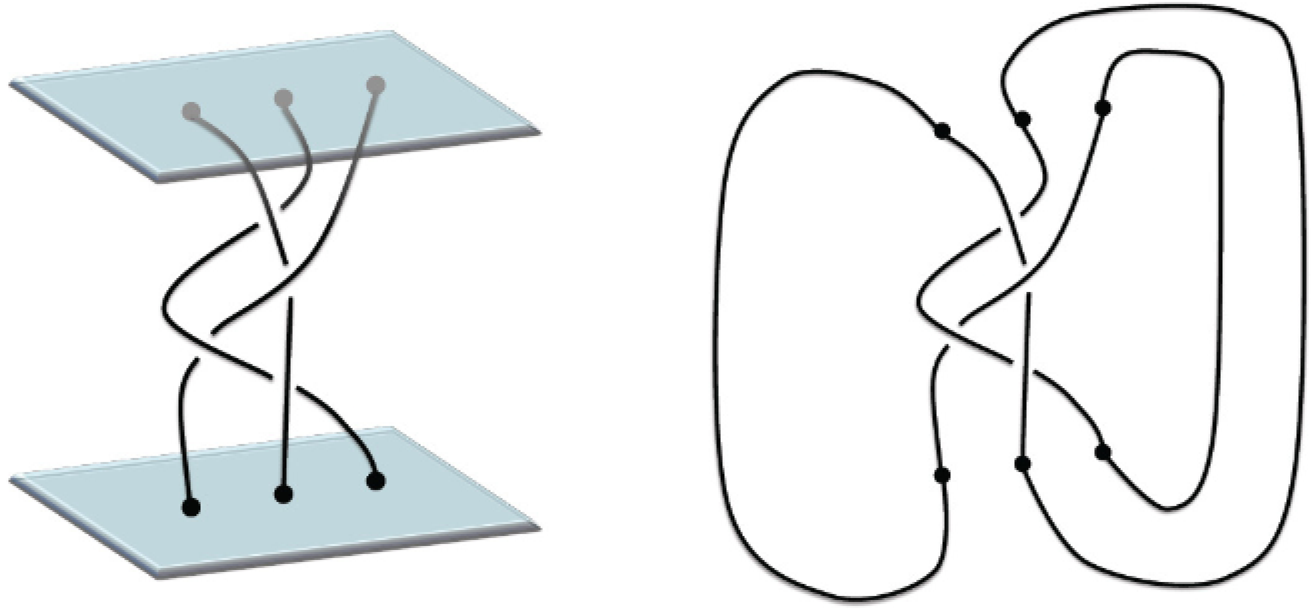
$$a + b \neq b + a.$$

Such groups are called "non-commutative" or "non-abelian."

Braid groups have many important practical applications. For example, they are used to construct efficient and robust public key encryption algorithms.[7]

Another promising direction is designing quantum computers based on creating complex braids of quantum particles known as anyons. Their trajectories weave around each other, and their overlaps are used to build "logic gates" of the quantum computer.[8]

There are also applications in biology. Given a braid with n threads, we can number the nails on the two plates from 1 to n from left to right. Then, connect the ends of the threads attached to the nails with the same number on the two plates. This will create what mathematicians call a "link": a union of loops weaving around each other.



In the example shown on this picture, there is only one loop. Mathematicians' name for it is "knot." In general, there will be several closed threads.

The mathematical theory of links and knots is used in biology: for example, to study bindings of DNA and enzymes.[9] We view a DNA molecule as one thread, and the enzyme molecule as another thread. It turns out that when they bind together, highly non-trivial knotting between them may occur, which may alter the DNA. The way they entangle is therefore of great importance. It turns out that the mathematical study of the resulting links sheds new light on the mechanisms of recombination of DNA.

In mathematics, braids are also important because of their geometric interpretation. To explain it, consider all possible collections of n points on the plane. We will assume that the points are distinct; that is, for any two points, their positions on the plane must be different. Let's choose one such collection; namely, n points arranged on a straight line, with the same distance between neighboring points. Think of each point as a little bug. As we turn on the music, these bugs come alive and start moving on the plane. If we view the time as the vertical direction, then the trajectory of each bug will look like a thread. If the positions of the bugs on the plane are distinct at all times – that is, if we assume that the bugs don't collide – then these threads will never intersect. While the music is playing, they can move around each other, just like the threads of a braid. However, we demand that when we stop the music after a fixed period of time, the bugs must align on a straight line in the same way as at the beginning, but each bug is allowed to end up in a position initially occupied by another bug. Then their collective path will look like a braid with n threads.

Thus, braids with n threads may be viewed as paths in the space of collections of n distinct points on the plane.[10]

The problem that Varchenko gave me, and on which I was about to start working with Fuchs, concerned a part of the braid group called the "commutator subgroup." Remember that for braids with two threads, we have defined the overlap number. A similar number can be assigned to a braid with any number of threads.[11] We use this to define the commutator subgroup B'_n of the braid group with n threads. It consists of all braids whose total overlap number is zero.[12]

The problem I had to solve was to compute the so-called "Betti numbers" of the group B'_n. These numbers reflect deep properties of this group, which are important in applications. As an analogy, think of a physical object, like a house. It has various characteristics: some more obvious, like the numbers of floors, rooms, doors, windows, etc., and some less so, like the proportions of the materials from which it is built. Likewise, a group also has various characteristics, and these are the Betti numbers.[13] Fuchs had earlier computed the Betti numbers of the braid group B_n itself. He gave me his paper so that I could learn the basics of the subject.

Within a week, I was able to read the entire Fuchs paper on my own, occasionally looking up previously unknown-to-me concepts and definitions in my by-then fairly large library of math books. I called Fuchs.

"Oh, it's you," he said. "I was wondering why you hadn't called. Have you started reading the article?"

"Yes, Dmitry Borisovich. Actually, I have finished it."

"Finished it?" Fuchs sounded surprised. "Well, then we should meet. I want to hear what you've learned."

Fuchs suggested we meet on the next day at MGU, after a seminar he was going to attend. As I was preparing for the meeting, I kept re-reading the article and practicing my answers to the kinds of questions that I thought Fuchs was likely to ask. A world-class mathematician like Fuchs wouldn't just take up a new student out of pity. The bar was set high. I understood that my first conversation with Fuchs would be something of an audition, and that's why I was so eager to make a good impression on him.

We met at the appointed hour and walked the corridors of *Mekh-Mat* to find a bench where we wouldn't be bothered. After we sat down, I started telling Fuchs what I learned from his article. He listened attentively, occasionally asking me questions. I think he was pleased by what he was hearing. He was curious where I learned all this stuff, and I told him about my studies with Evgeny Evgenievich, reading books, and attending lectures at *Mekh-Mat*. We even talked about my exam at the MGU (this was of course nothing new to Fuchs).

Luckily, our meeting went well. Fuchs seemed impressed with my knowledge. He told me that I was ready to tackle Varchenko's problem and that he would help me with it.

I was elated when I was leaving MGU that evening. I was about to start working on my first math problem, guided by one of the best mathematicians in the world. Less than two years had passed since my entrance exam at *Mekh-Mat*. I was back in the game.

Chapter 6

Apprentice Mathematician

Solving a mathematical problem is like doing a jigsaw puzzle, except you don't know in advance what the final picture will look like. It could be hard, it could be easy, or it could be impossible to solve. You never know until you actually do it (or realize that it's impossible to do). This uncertainty is perhaps the most difficult aspect of being a mathematician. In other disciplines, you can improvise, come up with different solutions, even change the rules of the game. Even the very notion of what constitutes a solution is not clearly defined. For example, if we are tasked with improving productivity in a company, what metrics do we use to measure success? Will an improvement by 20 percent count as a solution of the problem? How about 10 percent? In math, the problem is always well defined, and there is no ambiguity about what solving it means. You either solve it or you don't.

For Fuchs' problem, I had to compute the Betti numbers of the groups B'_n. There was no ambiguity in what this meant. It means the same thing today to everyone familiar with the language of math as it did in 1986 when I first learned about this problem, and will mean the same thing a hundred years from now.

I knew that Fuchs had solved a similar problem, and I knew how he'd done it. I prepared for my own task by working on similar problems for which solutions had already been known. This gave me intuition, skills, and a toolkit of methods. But I could not know *a priori* which of these methods would work or which way I should approach the problem – or even whether I could solve it without creating an essentially new technique or an entirely different method.

This quandary besets all mathematicians. Let's look at one of the most famous problems in mathematics, Fermat's Last Theorem, to see how one can go about doing math when the problem is easy to state but the solution is far from obvious. Fix a natural number n, that is, 1, 2, 3,..., and consider the equation

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

on the natural numbers x, y, and z.

If $n = 1$, we get the equation

$$x + y = z,$$

which surely has many solutions among natural numbers: just take any x and y and set $z = x + y$. Note that here we use the operation of addition of natural numbers that we discussed in the previous chapter.

If $n = 2$, we get the equation

$$x^2 + y^2 = z^2.$$

This equation also has many solutions in natural numbers; for instance,

$$3^2 + 4^2 = 5^2.$$

All of this has been known since antiquity. What was unknown was whether the equation had any solutions for *n* greater than 2. Sounds pretty simple, right? How hard could it be to answer a question like this?

Well, as it turned out, pretty hard. In 1637, a French mathematician, Pierre Fermat, left a note on the margin of an old book saying that if *n* is greater than 2, then the equation had no solutions *x*, *y*, *z* that are natural numbers. In other words, we cannot find three natural numbers *x*, *y*, *z* such that

$$x^3 + y^3 = z^3,$$

cannot find natural numbers *x*, *y*, *z* such that

$$x^4 + y^4 = z^4,$$

and so on.

Fermat wrote that he had found a simple proof of this statement, for all *n* greater than 2, but "this margin is too small to contain it." Many people, professional mathematicians as well as amateurs, took Fermat's note as a challenge and tried to reproduce his "proof," making this the most famous mathematical problem of all time. Prizes were announced. Hundreds of proofs were written and published, only to be crashed later on. The problem remained unsolved 350 years later.

In 1993, a Princeton mathematician, Andrew Wiles, announced his own proof of Fermat's Last Theorem. But his proof, at first glance, had nothing to do with the original problem. Instead of proving Fermat's Last Theorem, Wiles tackled the so-called Shimura–Taniyama–Weil conjecture, which is about something entirely different and is a lot more complicated to state. But a few years earlier, a Berkeley mathematician named Ken Ribet had proved that the statement of this conjecture implies Fermat's Last Theorem. That's why a proof of the conjecture would also prove Fermat's Last Theorem. We will talk about all this in detail in Chapter 8; the point I want to make now is that what looks like a simple problem may not necessarily have an elementary solution. It is clear to us now that Fermat could not have possibly proved the statement attributed to him. Entire fields of mathematics had to be created in order to do this, a development that took a lot of hard work by many generations of mathematicians.[1]

But is it possible to predict all that, given this innocent-looking equation?

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

Not at all!

With any math problem, you never know what the solution will involve. You hope and pray that you will be able to find a nice and elegant solution, and perhaps discover something interesting along the way. And you certainly hope that you will actually be able to do it in a reasonable period of time, that you won't have to wait for 350 years to reach the conclusion. But you can never be sure.

In the case of my problem, I was lucky; there was in fact an elegant solution that I was able to find in a relatively short period of time, about two months. But it didn't come easily to me. It never does. I tried many different methods. As each of them failed, I felt increasingly frustrated and anxious. This was my first problem, and inevitably I questioned whether I could be a mathematician. This problem was my first test of whether I had what it takes.

Working on this problem didn't excuse me from taking classes and passing exams at Kerosinka, but my highest priority was the problem, and I spent endless hours with it, nights and weekends. I was putting way too much pressure on myself. I was starting to have trouble sleeping, the first time this ever happened to me. The insomnia I acquired while working on this problem was the first "side effect" of my mathematical research. It haunted me for many months afterward, and from that point on I never allowed myself to get lost so completely in a math problem.

I met with Fuchs every week or so at the *Mekh-Mat* building, where I told him about my progress, or lack of it (by then he was able to get me an ID, so I did not have to scale the fence anymore). Fuchs was always supportive and encouraging, and each time we met he would tell me about a new trick or suggest a new insight, which I would try to apply to my problem.

And then, suddenly, I had it. I found the solution, or perhaps more accurately, the solution presented itself, in all of its splendor.

I was trying to use one of the standard methods for computing Betti numbers, which Fuchs had taught me, called "spectral sequence." I was able to apply it in a certain way, which allowed me in principle to compute the Betti numbers of the group B'_n from the knowledge of the Betti numbers of all the groups B'_m with $m < n$. The caveat was, of course, that I did not know what those other Betti numbers were either.

But this gave me a way to attack the problem: if I could *guess* the right answer, I would then have a path to *proving* it by following this method.

That's easy to say, but coming up with such a guess required many sample computations, which only became more and more complicated. For a long time, no pattern seemed to emerge.

Suddenly, as if in a stroke of black magic, it all became clear to me. The jigsaw puzzle was complete, and the final image was revealed to me, full of elegance and beauty, in a moment that I will always remember and cherish. It was an incredible feeling of high that made all those sleepless nights worthwhile.

For the first time in my life, I had in my possession something that *no one else in the world* had. I was able to say something new about the universe. It wasn't a cure for cancer, but it was a worthy piece of knowledge, and no one could ever take it away from me.

If you experience this feeling once, you will want to go back and do it again. This was the first time it happened to me, and like the first kiss, it was very special. I knew then that I could call myself mathematician.

The answer was actually quite unexpected, and much more interesting than what Fuchs or I could imagine. I found that for each divisor of the natural number n (the number of threads in the braids we are considering), there is a Betti number of the group B'_n that is equal to the celebrated "Euler function" of that divisor.[2]

The Euler function assigns to any natural number d another natural number, called $\varphi(d)$. This is the number of integers between 1 and d that are *relatively prime* with d; that is, have no common divisors with d (apart from 1, of course).

For example, take $d = 6$. Then 1 is relatively prime with 6, 2 is not (it is a divisor of 6), 3 is not (it is also a divisor of 6), 4 is not (4 and 6 share a common divisor; namely, 2), 5 is relatively prime with 6, and 6 is not. So there are two natural numbers between 1 and 6 that are relatively prime with 6: namely, 1 and 5. Hence the Euler function of 6 is equal to 2. We write this as $\varphi(6) = 2$.

The Euler function has many applications. For example, it is employed in the so-called RSA algorithm used to encrypt credit card numbers in online transactions (this is explained in endnote 7 to Chapter 14). It is named in honor of the eighteenth-century Swiss mathematician Leonhard Euler.

The fact that the Betti numbers I found were given by the Euler function suggested the existence of some hidden connections between braid groups and number theory. Therefore, the problem I had solved could potentially have implications far beyond its original scope.

Of course, I was eager to tell Fuchs about my results. It was already June 1986, almost three months after he and I first met. By then, Fuchs had left Moscow with his wife and two young daughters to spend the summer at his dacha near Moscow. Luckily for me, it was situated along the same train line as my hometown, about halfway, and so it was easy for me to visit him there on my way home.

After offering me a customary cup of tea, Fuchs asked me about my progress.

“I solved the problem!”

I couldn’t contain my excitement, and I guess the account of the proof that I gave was pretty rambling. But no worries – Fuchs understood everything quickly. He looked pleased.

“This is great,” he said, “Well done! Now you have to start writing a paper about this.”

It was the first time I wrote a math paper, and it turned out to be no less frustrating than my mathematical work, but much less fun. Searching for new patterns on the edge of knowledge was captivating and exciting. Sitting at my desk, trying to organize my thoughts and put them on paper, was an entirely different process. As someone told me later, writing papers was the punishment we had to endure for the thrill of discovering new mathematics. This was the first time I was so punished.

I came back to Fuchs with different drafts, and he read them carefully, pointing out deficiencies and suggesting improvements. As always, he was extremely generous with his help. From the beginning, I put Fuchs’ name as one of the coauthors, but he flatly refused. “This is your paper,” he said. Finally, Fuchs declared that the article was ready, and he told me that I should submit it to *Functional Analysis and Applications*, the math journal run by Israel Moiseevich Gelfand, the patriarch of the Soviet mathematical school.

A compact charismatic man, then in his early seventies, Gelfand was a legend in the Moscow mathematical community. He presided over a weekly seminar held at a grand auditorium on the fourteenth floor of the main MGU building. This was an important mathematical and social event, which had been running for more than fifty years and was renowned all over the world. Fuchs was a former collaborator of Gelfand (their work on what became known as “Gelfand–Fuchs cohomology” was widely known and appreciated) and one of the most senior members of Gelfand’s seminar. (The others included A.A. Kirillov, who was Gelfand’s former student, and M.I. Graev, Gelfand’s longtime collaborator).

The seminar was unlike any other seminars I have ever attended. Usually, a seminar has fixed hours – in the U.S. one hour or an hour and a half – and there is a speaker who prepares a talk on a particular topic chosen in advance. Occasionally, the audience members ask questions. It was not at all like this at the Gelfand seminar. It met every Monday evening, and the official starting time was 7:00 pm. However, the seminar rarely started before 7:30, and it usually began around 7:45 to 8:00. During the hour or so before the start, the members of the seminar, including Gelfand himself (who usually arrived around 7:15–7:30), would wander around and talk to each other inside the auditorium and in the large foyer outside. Clearly, this was what Gelfand had intended. This was as much a social event as a math seminar.

Most of the mathematicians coming to the Gelfand seminar worked at various places that were not affiliated with MGU. Gelfand's seminar was the only place where they could meet their peers, find out what was happening in the world of mathematics, share their ideas, and forge collaborations. Since Gelfand was himself Jewish, his seminar was considered as one of "safe havens" for Jews and even hailed as "the only game in town" (or one of very few) in which Jewish mathematicians could participate (though, in fairness, many other seminars at MGU were open to the public and were run by people who were not prejudiced against any ethnicities). No doubt, Gelfand gladly took advantage of this.

The anti-Semitism that I had experienced at my entrance exam to MGU spread to all levels of academia in the Soviet Union. Earlier, in the 1960s and early 1970s, even though there were restrictions, or "quotas," for students of Jewish background, they could still get in as undergraduates at the *Mekh-Mat* (the situation gradually worsened throughout the 1970s and early 1980s, to the point where in 1984, when I was applying to *Mekh-Mat*, almost no Jewish students were accepted).[3] But even in those years, it was nearly impossible for these students to enter graduate school. The only way Jewish students could do this was to go to work somewhere for three years after getting the bachelor's degree, and then one could be sent to graduate school by their employer (often, located somewhere in a faraway province). And even if they managed to overcome this hurdle and get a Ph.D., it was impossible for them to find an academic job in mathematics in Moscow (at MGU, for example). Either they had to settle for a job somewhere in the province or join one of many research institutes in Moscow that had little or nothing to do with mathematics. The situation was even more difficult for those who were not originally from Moscow, because they did not have *propiska*, a Moscow residency stamp in their interior passport, which was required for employment in the capital.

Even the most exceptional students got such treatment. Vladimir Drinfeld, a brilliant mathematician and future Fields Medal winner about whom we will talk more later, was allowed to become a graduate student at *Mekh-Mat* right after obtaining his bachelor's degree (though from what I've heard it was very difficult to arrange), but being a native of Kharkov, Ukraine, he could not be employed in Moscow. He had to settle for a teaching job at a provincial university in Ufa, an industrial city in the Ural Mountains. Eventually, he got a job as a researcher at the Institute for Low Temperature Physics in Kharkov.

Those who stayed in Moscow were employed at places like the Institute for Seismic Studies or Institute for Signal Processing. Their day jobs consisted of some tedious calculations related to a particular industry to which their institute was attached (though some actually managed to break new ground in those areas, multi-talented as they were). They had to do the kind of mathematical research that was their true passion on the side, in their spare time.

Gelfand himself was forced out of his teaching job at *Mekh-Mat* in 1968 after he signed the famous letter of ninety-nine mathematicians demanding the release of the mathematician and human rights activist Alexander Esenin-Volpin (the son of the poet Sergei Esenin) from a politically motivated detention in a psychiatric hospital. That letter was so skillfully written that after it was broadcast on the BBC radio, the worldwide outrage embarrassed the Soviet leadership to release Esenin-Volpin almost immediately.[4] But this also greatly angered the authorities. They subsequently found ways to punish everyone who signed it. In particular, many of the signatories were fired from their teaching jobs.[5]

So Gelfand was no longer professor of mathematics at MGU, though he was able to preserve his seminar that was still being held at the main MGU building. His official job was at a biological lab of MGU that he had founded to conduct research in biology, which was also his passion.[*] Fuchs was employed at the same lab.

Fuchs had earlier urged me to start attending Gelfand's seminar, and I did come to a couple of meetings at the very end of the spring semester. Those meetings made a great impression on me. Gelfand ran his seminar in the most authoritarian way. He decided its every aspect, and though to an untrained eye his seminars could appear chaotic and disorganized, he actually devoted an enormous amount of time and energy to the preparation and choreographing of the weekly meetings.

Three years later, when Gelfand asked me to speak about my work, I had the opportunity to see the inner workings of the seminar up close. For now, I was observing it from the vantage point of a seventeen-year-old student just starting his mathematical career.

The seminar was in many ways the theater of a single actor. Officially, there would be a designated speaker reporting on a designated topic, but typically only part of the seminar would be devoted to it. Gelfand would usually bring up other topics and call other mathematicians, who had not been asked to prepare in advance, to the blackboard to explain them. But he was always at the center of it all. He and only he controlled the flow of the seminar and had the absolute power to interrupt the speaker at any moment with questions, suggestions, and remarks. I can still hear him say "*Dayte opredelenie*" – "Give the definition" – his frequent admonition to a speaker.

He also had the habit of launching into long tirades on various topics (sometimes unrelated to the material discussed), telling jokes, anecdotes, and stories of all kinds, many of them quite entertaining. This was where I heard the parable that I mentioned in the Preface: a drunkard may not know which number is larger, 2/3 or 3/5, but he knows that 2 bottles of vodka for 3 people is better than 3 bottles of vodka for 5 people. One of Gelfand's skills was his ability to "rephrase" questions asked by others in such a way that the answer became obvious.

Another joke he liked to tell involved the wireless telegraph: "At the beginning of the twentieth century, someone asks a physicist at a party: can you explain how it works? The physicist replies that it's very simple. First, you have to understand how the ordinary, wired, telegraph works: imagine a dog with its head in London and its tail in Paris. You pull the tail in Paris, and the dog barks in London. The wireless telegraph, says the physicist, is the same thing, but without a dog."

After recounting the joke and waiting for the laughter to subside (even from those people in the audience who had heard it a thousand times), Gelfand would pivot to whatever math problem was being discussed. If he thought that the solution of the problem required a radically new approach, he would comment, "What I'm trying to say is we need to do it without a dog."

A frequently used device at the seminar was to appoint a *kontrol'nyj slushatel'*, a test listener, usually a junior member of the audience, who was supposed to repeat at regular intervals what the speaker was saying. If it was deemed that the "test listener" was following the lecture well, this meant that the speaker was doing a good job. Otherwise, the speaker had to slow down and explain better. Occasionally, Gelfand would even discharge a particularly incomprehensible speaker in disgrace and replace him or her with another member of the audience. (Of course, Gelfand would poke fun at the test listener as well.) All of this made the seminar very entertaining.

Most seminars proceed at a steady pace, with people in the audience listening politely (and some perhaps dozing off) – too complacent, too polite, or simply afraid to ask the speaker any questions, and perhaps learning little. There is no doubt that the uneven pace and the generally subversive character of the Gelfand seminar not only kept people awake (not an easy task given that the seminar often lasted till midnight), but stimulated them in ways that other seminars simply couldn't. Gelfand demanded a lot of his speakers. They worked hard, and so did he. Whatever one can say about Gelfand's style, people never left the seminar empty-handed.

However, it seems to me that a seminar like this could only exist in a totalitarian society, like the Soviet Union. People were accustomed to the kind of dictatorial powers and behavior that Gelfand displayed. He could be cruel, at times insulting, to people. I don't think many would tolerate this kind of treatment in the West. But in the Soviet Union, this was not considered to be out of the ordinary, and no one protested. (Another famous example like this was Lev Landau's seminar on theoretical physics.)

When I first started coming to the seminar, Gelfand had a young physicist, Vladimir Kazakov, present a series of talks about his work on so-called matrix models. Kazakov used methods of quantum physics in a novel way to obtain deep mathematical results that mathematicians could not obtain by more conventional methods. Gelfand had always been interested in quantum physics, and this topic had traditionally played a big role at his seminar. He was particularly impressed with Kazakov's work and was actively promoting it among mathematicians. Like many of his foresights, this proved to be golden: a few years later this work became famous and fashionable, and it led to many important advances in both physics and math.

In his lectures at the seminar, Kazakov was making an admirable effort to explain his ideas to mathematicians. Gelfand was more deferential to him than usual, allowing him to speak without interruptions longer than other speakers.

While these lectures were going on, a new paper arrived, by John Harer and Don Zagier, in which they gave a beautiful solution to a very difficult combinatorial problem.[6] Zagier has a reputation for solving seemingly intractable problems; he is also very quick. The word was that the solution of this problem took him six months, and he was very proud of that. At the next seminar, as Kazakov was continuing his presentation, Gelfand asked him to solve the Harer–Zagier problem using his work on the matrix models. Gelfand had sensed that Kazakov's methods could be useful for solving this kind of problem, and he was right. Kazakov was unaware of the Harer–Zagier paper, and this was the first time he heard this question. Standing at the blackboard, he thought about it for a couple of minutes and immediately wrote down the Lagrangian of a quantum field theory that would lead to the answer using his methods.

Everyone in the audience was stunned. But not Gelfand. He asked Kazakov innocently, "Volodya, how many years have you been working on this topic?"

"I am not sure, Israel Moiseevich, perhaps six years or so."

"So it took you six years plus two minutes, and it took Don Zagier six months. Hmmm... You see how much better he is?"

And this was a mild "joke," compared with some others. You had to have thick skin in order to survive in this environment. Unfortunately, some speakers took these kinds of public put-downs personally, and this caused them a lot of torment. But I have to add that Gelfand always had a sharper tongue for the older, more established mathematicians, and he was much more gentle to young mathematicians, especially to students.

He used to say that at the seminar he welcomed all undergraduates, talented graduate students, and only brilliant professors. He understood that in order to keep the subject moving, it was very important to prepare new generations of mathematicians, and he always surrounded himself with young talent. They kept him young as well (he was actively doing cutting-edge research till he was in his late eighties). Often, he would even invite high school students to the seminar and make them sit in the front row to make sure that they were following what was going on. (Of course, these were no ordinary high school students. Many of them went on to become world-renowned mathematicians.)

By all accounts, Gelfand was very generous with his students, spending hours talking to them on a regular basis. Very few professors do this. It wasn't easy to be his student; he gave them a kind of tough love, and they had to cope with his various quirks and dictatorial habits. But my impression from talking to many of them is that they were all loyal to him and felt they owed him a tremendous debt.

I was not Gelfand's student – I was his "grand-student," as both of my teachers, Fuchs and Feigin (who was not yet in my life), were at least partially Gelfand's students. Hence I always considered myself as being part of the "Gelfand mathematical school." Much later, when he and I were in the United States, Gelfand asked me directly about this, and by the look of satisfaction on his face when I said yes, I could tell how important the issue of his school and recognition of who was part of it was to him.

This school, of which the seminar was the focal point, its window to the world, had an enormous impact not only on mathematics in Moscow, but around the world. Foreign mathematicians came to Moscow just to meet Gelfand and attend his seminar, and many considered it an honor to lecture there.

Gelfand's fascinating and larger-than-life personality played a big role in the seminar's reputation. A few years later, he became interested in my work and asked me to speak at his seminar. I spent many hours talking to him, not just about mathematics, but about a lot of other things. He was very interested in the history of mathematics and his own legacy in particular. I remember vividly how, when I first came to visit him at his Moscow apartment (I had just turned twenty-one), he informed me that he considered himself the Mozart of mathematics.

"Most composers are remembered for particular pieces they wrote," he said. "But in the case of Mozart, that's not so. It's the totality of his work that makes him a genius." He paused and continued: "The same goes for my mathematical work."

Putting aside some interesting questions raised by such self-assessment, I think it's actually an apt comparison. Though Gelfand did not prove any famous long-standing conjectures (such as Fermat's Last Theorem), the cumulative effect of his ideas on mathematics was staggering. Perhaps more importantly, Gelfand possessed an excellent taste for beautiful mathematics as well as an astute intuition about which areas of mathematics were the most interesting and promising. He was like an oracle who had the power to predict in which directions mathematics would move.

In the subject that was becoming increasingly fractured and specialized, he was one of the last remaining Renaissance men able to bridge different areas. He epitomized the unity of mathematics. Unlike most seminars, which focus on one area of math, if you came to Gelfand's seminar, you could see how all these different parts fit together. That's why all of us gathered every Monday night on the fourteenth floor of the main MGU building and eagerly awaited the word of the master.

And it was to this awe-inspiring man that Fuchs suggested I submit my first math paper. Gelfand's journal, *Functional Analysis and Applications*, was published in four slim issues a year, about a hundred pages each (a pitiful amount for a journal like this, but the publisher refused to give more, so one had to cope), and it was held in extremely high regard around the world. It was translated into English, and many science libraries around the world subscribed to it.

It was very difficult to get a paper published in this journal, partly because of the severe page limitations. There were in fact two types of papers that were published: research articles, each typically ten–fifteen pages long, containing detailed proofs, and short announcements, in which only the results were stated, without proofs. The announcement could not be longer than two pages. In theory, such a short paper was supposed to be followed eventually by a detailed article containing all proofs, but in reality quite often that did not happen because publishing a longer article was extremely difficult. Indeed, it was nearly impossible for a mathematician in the USSR to publish abroad (one needed to get all sorts of security clearances, which could easily take more than a year and a lot of effort). On the other hand, the number of math publications in the Soviet Union, considering the number of mathematicians there, was very small. Unfortunately, many of them were controlled by various groups, which would not allow outsiders to publish, and anti-Semitism was also prevalent in some of them.

Because of all this, a certain subculture of math papers emerged in the USSR, which came to be referred to as the "Russian tradition" of math papers: extremely terse writing, with few details provided. What many mathematicians outside of the Soviet Union did not realize was that this was largely done by necessity, not by choice.

It was this kind of short announcement that Fuchs was aiming at for my first article.

Each article submitted to *Functional Analysis and Applications*, including short announcements, had to be screened and approved by Gelfand. If he liked it, he would then let the article go through the standard refereeing process. This meant that for my article to be considered, I had to meet Israel Moiseevich in person. So before one of the first seminars of the fall semester of 1986, Fuchs introduced me to him.

Gelfand shook my hand, smiled, and said, "I am pleased to meet you. I've heard about you."

I was totally star-struck. I could swear that I saw a halo around Gelfand's head.

Then Gelfand turned to Fuchs and asked him to show my article, which Fuchs handed to him. Gelfand started to turn the pages. There were five of them, which I neatly typed (slowly, with two fingers) on a typewriter I borrowed at Kerosinka, and then inserted formulas by hand.

"Interesting," Gelfand said approvingly, and then turned to Fuchs: "But why is this important?"

Fuchs started to explain something about the discriminant of polynomials of degree *n* with distinct roots, and how my result could be used to describe the topology of the fiber of the discriminant, and… Gelfand interrupted him: "Mitya," he said, using the diminutive form of Fuchs' first name, "Do you know how many subscribers the journal has?"

"No, Israel Moiseevich, I don't."

"More than a thousand." That was a pretty large number given how specialized the journal was. "I cannot send you with every issue so that you would explain to each subscriber what this result is good for, now can I?"

Fuchs shook his head.

"This has to be written clearly in the paper, OK?" Gelfand made the point of saying all of this to Fuchs, as though it was all *his* fault. Then he said to both of us: "Otherwise, the paper looks good to me."

With that, he smiled again at me and went to talk to someone else.

Quite an exchange! Fuchs waited until Gelfand was out of the earshot and said to me, "Don't worry about this. He just wanted to impress you." (And he sure did!) "We'll just have to add a paragraph to this effect at the beginning of the paper, and after that he will probably publish it."

That was the best possible outcome. After adding a paragraph required by Gelfand, I officially submitted the article, and eventually it appeared in the journal.[7] With that, my first math project was complete. I crossed my first threshold and was at the beginning of a path that would lead me into the magical world of modern math.

This is the world I want to share with you.

[*] It is also worth noting that Gelfand was not elected as a full member of the Academy of Sciences of the USSR until the mid-1980s because the Mathematical Branch of the Academy was for decades controlled by the director of the Steklov Mathematical Institute in Moscow, Ivan Matveevich Vinogradov, nicknamed the "Anti-Semite-in-Chief of the USSR." Vinogradov had put in place draconian anti-Semitic policies at the Academy and the Steklov Institute, which was in his grip for almost fifty years.

Chapter 7

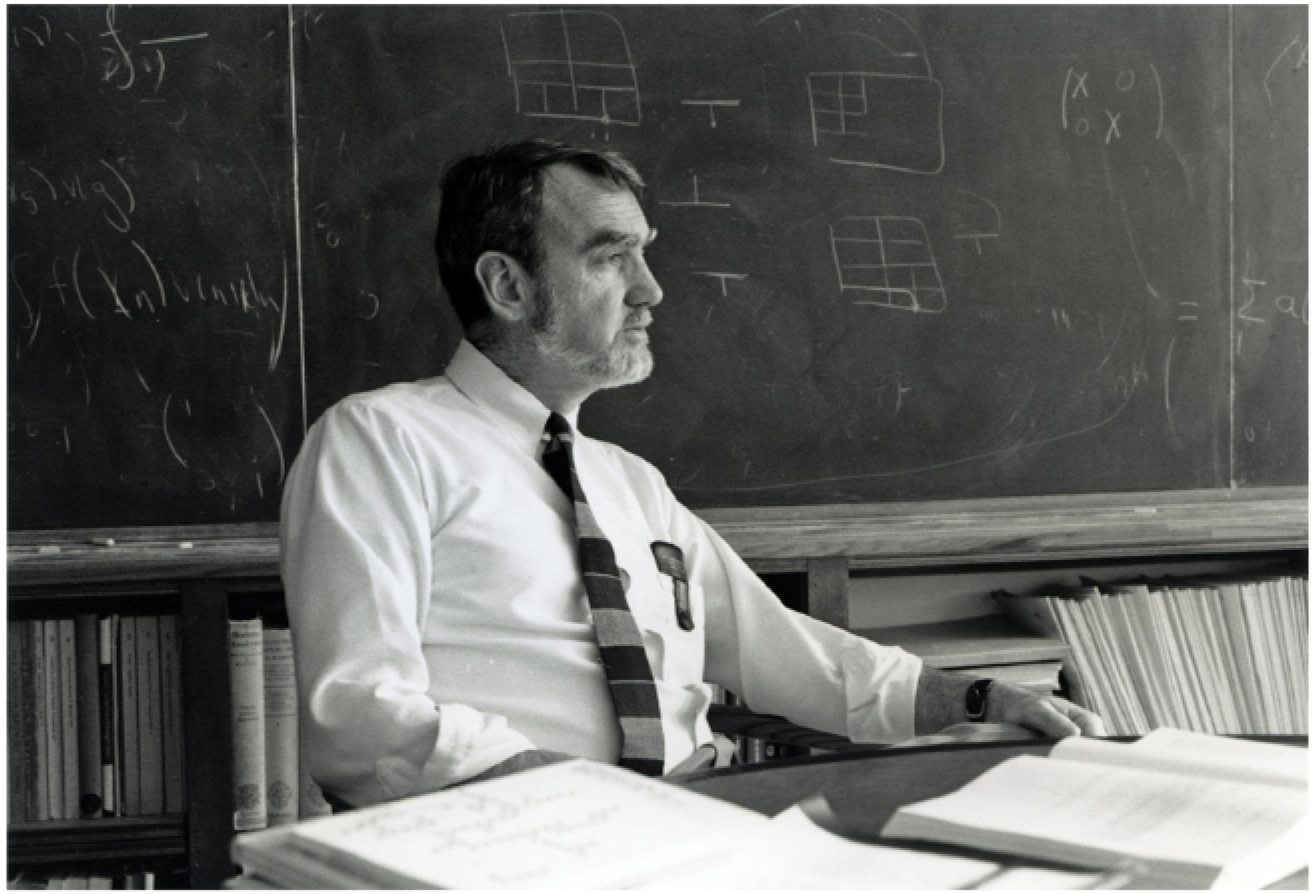
The Grand Unified Theory

The solution of the first problem was my initiation into the temple of mathematics. Somewhat serendipitously, the next mathematical project I did with Fuchs brought me into the midst of the Langlands Program, one of the deepest and most exciting mathematical theories to emerge in the past fifty years. I will tell you about my project below, but my goal in this book is to describe much more than my own experience. It is to give you a sense of modern math, to prove that it is really about originality, imagination, groundbreaking insights. And the Langlands Program is a great example. I like to think of it as a Grand Unified Theory of Mathematics because it uncovers and brings into focus mysterious patterns shared by different areas of math and thus points to deep, unexpected connections between them.

Mathematics consists of many subfields. They often feel like different continents, with mathematicians working in those subfields speaking different languages. That's why the idea of "unification," bringing together the theories coming from these diverse fields and realizing that they are all part of a single narrative, is so powerful. It's as if you suddenly realized that you could understand another language, one you had desperately tried to learn without much success.

It's useful to think about mathematics as a whole as a giant jigsaw puzzle, in which no one knows what the final image is going to look like. Solving this puzzle is a collective enterprise of thousands of people. They work in groups: here are the algebraists laboring over their part of the puzzle, here are the number theorists, here are the geometers, and so on. Each group has been able to create a small "island" of the big picture, but through most of the history of mathematics, it has been hard to see how these little islands will ever join up. As a result, most people work on expanding those islands of the puzzle. Every once in a while, however, someone will come who will see how to connect the islands. When this happens, important traits of the big picture emerge, and this gives a new meaning to the individual fields.

This is what Robert Langlands did, but his ambition went deeper than simply joining a few islands. Instead, the Langlands Program that he initiated in the late 1960s has become an attempt to find the mechanism by which we could build bridges between many islands, no matter how unrelated they may seem.



Robert Langlands at his office in Princeton, 1999. Photo by Jeff Mozzochi.

Langlands is now emeritus professor of mathematics at the Institute for Advanced Study in Princeton, where he occupies the office formerly held by Albert Einstein. A man of amazing talent and vision, he was born in 1936 and grew up in a small town near Vancouver; his parents had a millwork business. One of the striking things about Langlands is his fluency in many languages: English, French, German, Russian, and Turkish, even though he didn't speak any languages besides his native English before he entered college.[1]

I have had the opportunity to collaborate with Langlands closely in recent years, and we have often corresponded in Russian. At some point he sent me the list of Russian authors that he had read in the original. The list was so extensive that it seemed he may well have read more of my native Russian literature than I have. I often wonder whether Langlands' unusual language abilities have something to do with his power to bring together different mathematical cultures.

The key point of the Langlands Program is the concept of symmetry that is already familiar to us. We have talked about symmetry in geometry: for example, any rotation is a symmetry of a round table. Our study of these symmetries has led us to the notion of a group. We then saw that groups appear in mathematics in different guises: as groups of rotations, braid groups, and so on. We have also seen that groups were instrumental in classifying elementary particles and predicting the existence of quarks. The groups that are relevant to the Langlands Program appear in the study of numbers.

To explain this, we need to talk first about the numbers that we encounter in our everyday life. Each of us was born in a particular year, lives in a house that has a particular number on the street, has a phone number, a PIN to access a bank account at the ATM, and so forth. All of these numbers have something in common: each of them is obtained by adding number 1 to itself a certain number of times: 1 + 1 is 2, 1 + 1 + 1 is 3, and so on. These are called the natural numbers.

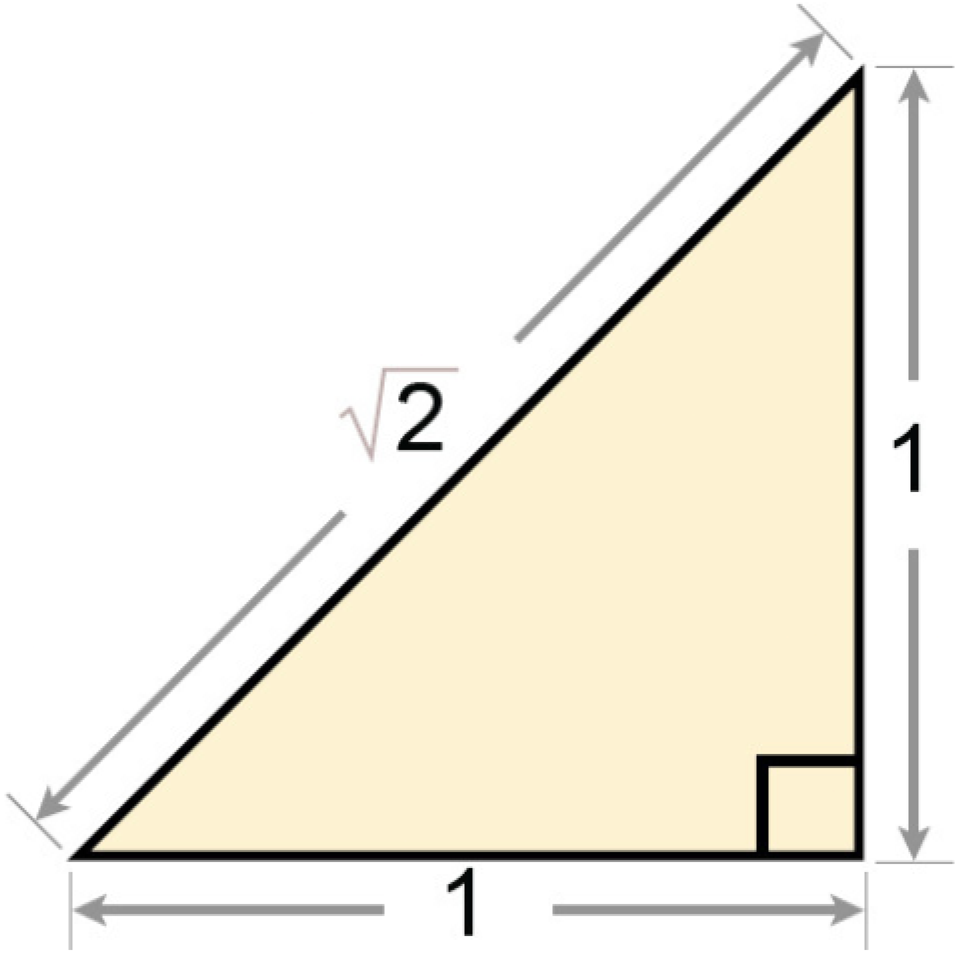
We also have the number 0, and the negative numbers: −1, −2, −3,... As we discussed in Chapter 5, these numbers go by the name "integers." So an integer is a natural number, or number 0, or the negative of a natural number.

We also encounter slightly more general numbers. A price, in dollars and cents, is often represented like this: $2.59, meaning two dollars and fifty-nine cents. This is the same as 2 plus the fraction 59/100, or 59 times 1/100. Here

1/100 means the quantity that being added to itself 100 times gives us 1. Numbers of this kind are called rational numbers, or fractions.

A good example of a rational number is a quarter; mathematically, it is represented by the fraction 1/4. More generally, for any two integers m and n we can form the fraction m/n. If m and n have a common divisor, say d (that is to say, $m = dm'$ and $n = dn'$), then we can cancel out d and write m'/n' instead of m/n. For example, 1/4 can also be represented as 25/100, and that's why Americans can say that a quarter is the same thing as 25 cents.

The vast majority of the numbers we encounter in our everyday life situations are these fractions, or rational numbers. But there are also numbers that are not rational. An example is the square root of 2, which we write as follows: $\sqrt{2}$. It is the number whose square is equal to 2. Geometrically, $\sqrt{2}$ is the length of the hypotenuse of the right triangle with legs of length 1.



It turns out we cannot represent it as m/n, where m and n are two natural numbers.[2] However, we can approximate it by rational numbers if we write the first few digits of its decimal form: 1.4142, then 1.41421, then 1.414213, and so on. But no matter how many decimal digits we retain, this will be an approximation – there will be more digits to follow. No finite decimal number will ever do justice to $\sqrt{2}$.

Since $\sqrt{2}$ is the length of the hypotenuse of the above triangle, we know that this number is out there. But it just does not fit the numerical system of rational numbers.

There are many other numbers like that, such as $\sqrt{3}$ or the cubic root of 2. We need to develop a systematic way to add these numbers to the rational numbers. Think of the rational numbers as a cup of tea. We can drink it by itself, but our experience will be enhanced if we mix in sugar, milk, honey, various spices – and these are like the numbers $\sqrt{2}$, $\sqrt{3}$, etc.

Let's try to mix in $\sqrt{2}$. This will be the equivalent of adding a cube of sugar to our cup of tea. So we drop $\sqrt{2}$ in the rationals and see what kind of numerical system we obtain. Surely, we want to be able to multiply the numbers within this new numerical system, so we have to include all numbers that are products of rational numbers and $\sqrt{2}$. These have the form $\frac{k}{l}\sqrt{2}$. So our numerical system must include all fractions $\frac{m}{n}$ (these are the rational numbers) and all numbers of the form $\frac{k}{l}\sqrt{2}$. But we also want to be able to add them to each other, so we also have to include the sums

$$\frac{m}{n} + \frac{k}{l}\sqrt{2}.$$

The collection of all numbers of this form is already "self-contained," in the sense that we can perform all the usual operations on them – addition, subtraction, multiplication, and division – and the result will also be a number of the same form.[3] This is our cup of tea with the cube of sugar fully mixed with the tea.

It turns out that this new numerical system has a hidden property that the rational numbers didn't have. This property will be our portal into the magical world of numbers. Namely, it turns out that this numerical system has symmetries.

By a "symmetry" I mean here a rule that assigns a new number to whatever number we begin with. In other words, a given symmetry transforms each number to another number from the same numerical system. We will say that a symmetry is a rule by which each number "goes" to some other number. This rule should be compatible with the operations of addition, subtraction, multiplication, and division. It is not clear yet why we should care about the symmetries of a numerical system. Please bear with me and you will see why momentarily.

Our numerical system has the identity symmetry, the rule by which every number goes to itself. This is like the rotation of a table by 0 degrees, under which every point of the table goes to itself.

It turns out that our numerical system also has a non-trivial symmetry. To explain what it is, let's observe that $\sqrt{2}$ is a solution of the equation $x^2 = 2$. Indeed, if we substitute $\sqrt{2}$ for x, we obtain an equality. But this equation actually has two solutions: one of them is $\sqrt{2}$ and the other is $-\sqrt{2}$. And we have in fact added both of them to the rational numbers when we constructed our new numerical system. Switching these two solutions, we obtain a symmetry of this numerical system.[*]

To illustrate this more fully in terms of our tea cup analogy, let's modify it slightly. Let's say that we drop a cube of white sugar and a cube of brown sugar in our cup and mix them with the tea. The former is like $\sqrt{2}$ and the latter is like $-\sqrt{2}$. Clearly, exchanging them will not change the resulting cup of tea. Likewise, exchanging $\sqrt{2}$ and $-\sqrt{2}$ will be a symmetry of our numerical system.

Under this exchange, rational numbers remain unchanged.[4] Therefore, the number of the form $\frac{m}{n}+\frac{k}{l}\sqrt{2}$ will go to the number $\frac{m}{n}-\frac{k}{l}\sqrt{2}$. In other words, in every number we simply change the sign in front of $\sqrt{2}$ and leave everything else the same.[5]

You see, our new numerical system is like a butterfly: the numbers $\frac{m}{n}+\frac{k}{l}\sqrt{2}$ are like the scales of a butterfly, and the symmetry of these numbers exchanging $\sqrt{2}$ and $-\sqrt{2}$ is like the symmetry of the butterfly exchanging its wings.

More generally, we can consider other equations in the variable x instead of $x^2 = 2$; for example, the cubic equation $x^3 - x + 1 = 0$. If the solutions of such an equation are not rational numbers (as is the case for the above equations), then we can adjoin them to the rational numbers. We can also adjoin to the rational numbers the solutions of several such equations at once. This way we obtain many different numerical systems, or, as mathematicians call them, *number fields*. The word "field" refers to the fact that this numerical system is closed under the operations of addition, subtraction, multiplication, and division.

Just like the number field obtained by adjoining $\sqrt{2}$, general number fields possess symmetries compatible with these operations. The symmetries of a given number field can be applied one after another (composed with each other), just like symmetries of a geometric object. It is not surprising then that these symmetries form a group. This group is called the *Galois group* of the number field,[6] in honor of the French mathematician Évariste Galois.

The story of Galois is one of the most romantic and fascinating stories about mathematicians ever told. A child prodigy, he made groundbreaking discoveries very young. And then he died in a duel at the age of twenty. There are different views on what was the reason for the duel, which happened on May 31, 1832: some say there was a woman involved, and some say it was because of his political activities. Certainly, Galois was uncompromising in expressing his political views, and he managed to upset many people during his short life.

It was literally on the eve of his death that, writing frantically in a candlelit room in the middle of the night, he completed his manuscript outlining his ideas about symmetries of numbers. It was in essence his love letter to humanity in which he shared with us the dazzling discoveries he had made. Indeed, the symmetry groups Galois discovered, which now carry his name, are the wonders of our world, like the Egyptian pyramids or the Hanging Gardens of Babylon. The difference is that we don't have to travel to another continent or through time to find them. They are right at our fingertips, wherever we are. And it's not just their beauty that is captivating; so is their high potency for real-world applications.

Alas, Galois was far ahead of his time. His ideas were so radical that his contemporaries could not understand them at first. His papers were twice rejected by the French Academy of Sciences, and it took almost fifty years for his work to be published and appreciated by other mathematicians. Nevertheless, it is now considered as one of the pillars of modern mathematics.

What Galois had done was bring the idea of symmetry, intuitively familiar to us in geometry, to the forefront of number theory. What's more, he showed symmetry's amazing power.

Before Galois, mathematicians focused on trying to discover explicit formulas for solutions of equations like $x^2 = 2$ and $x^3 - x + 1 = 0$, called polynomial equations. Sadly, this is how we are still taught at school, even though two centuries have passed since Galois' death. For example, we are required to memorize a formula for solutions of a general quadratic equation (that is, of degree 2)

$$ax^2 + bx + c = 0$$

in terms of its coefficients a, b, c. I won't write this formula here so as not to trigger any unpleasant memories. All we need to know about it now is that it involves taking the square root.

Likewise, there is a similar, but more complicated, formula for a general cubic equation (of degree 3)

$$ax^3 + bx^2 + cx + d = 0,$$

in terms of its coefficients a, b, c, d, which involves cubic roots. The task of solving a polynomial equation in terms of radicals (that is, square roots, cubic roots, and so forth) is quickly becoming more and more complicated as the degree of the equation grows.

The general formula for the solutions of the quadratic equations was already known to the Persian mathematician Al-Khwarizmi in the ninth century (the word "algebra" originated from the word "al-jabr," which appears in the title of his book). Formulas for solutions of the cubic and quartic (degree 4) equations were discovered in the first half of the sixteenth century. Naturally, the next target was a quintic equation (of degree 5). Prior to Galois, many mathematicians had been desperately trying to find a formula for its solutions for almost 300 years, to no avail. But Galois realized that they had been asking the wrong question. Instead, he said, we should focus on the group of symmetries of the number field obtained by adjoining the solutions of this equation to the rational numbers – this is what we now call the Galois group.

The question of describing the Galois group turns out to be much more tractable than the question of writing an explicit formula for the solutions. One can say something meaningful about this group even without knowing what the solutions are. And from this one can then infer important information about the solutions. In fact, Galois was able to show that a formula for solutions in terms of radicals (that is, square roots, cubic roots, and so on) exists if and only if the corresponding Galois group has a particularly simple structure: is what mathematicians now call a *solvable* group. For quadratic, cubic, and quartic equations, the Galois groups are always solvable. That's why solutions of these equations may be written in terms of the radicals. But Galois showed that the group of symmetries of a typical quintic equation (or an equation of a higher degree) is not solvable. This immediately implies that there is no formula for solutions of these equations in terms of radicals.[7]

I won't get into the details of this proof, but let's consider a couple of examples of Galois groups to give you an idea what these groups look like. We have already described the Galois group in the case of the equation $x^2 = 2$. This equation has two solutions, $\sqrt{2}$ and $-\sqrt{2}$, which we adjoin to the rational numbers. The Galois group of the resulting number field[8] then consists of two elements: the identity and the symmetry exchanging $\sqrt{2}$ and $-\sqrt{2}$.

As our next example, consider a cubic equation written above, and suppose that its coefficients are rational numbers, but all of its three solutions are irrational. We then construct a new number field by adjoining these solutions to the rational numbers. It's like adding three different ingredients to our cup of tea: say, a cube of sugar, a dash of milk, and a spoonful of honey. Under any symmetry of this number field (the cup of tea with these ingredients added), the cubic equation won't change because its coefficients are rational numbers, which are preserved by symmetries. Hence each solution of the cubic equation (one of the three ingredients) will necessarily go to another solution. This observation allows us to describe the Galois group of symmetries of this number field in terms of permutations of these three solutions. The main point is that we obtain this description without writing down any formulas for the solutions.[9]

Similarly, the Galois group of symmetries of the number field obtained by adjoining all solutions of an arbitrary polynomial equation to the rational numbers may also be described in terms of permutations of these solutions (there will be n solutions for a polynomial equation of degree n whose solutions are all distinct and not rational). This way we can infer a lot of information about the equation without expressing its solutions in terms of the coefficients.[10]

Galois' work is a great example of the power of a mathematical insight. Galois did not solve the problem of finding a formula for solutions of polynomial equations in the sense in which it was understood. He *hacked* the problem! He reformulated it, bent and warped it, looked at it in a totally different light. And his brilliant insight has forever changed the way people think about numbers and equations.

And then, 150 years later, Langlands took these ideas much farther. In 1967, he came up with revolutionary insights tying together the theory of Galois groups and another area of mathematics called harmonic analysis. These two areas, which seem light years apart, turned out to be closely related. Langlands, then in his early thirties, summarized his ideas in a letter to the eminent mathematician André Weil. Copies were widely circulated among mathematicians at the time.[11] The letter's cover note is remarkable for its understatement:[12]

> Professor Weil: In response to your invitation to come and talk, I wrote the enclosed letter. After I wrote it I realized there was hardly a statement in it of which I was certain. If you are willing to read it as pure speculation I would appreciate that; if not – I am sure you have a waste basket handy.

What followed was the beginning of a groundbreaking theory that forever changed the way we think about mathematics. Thus, the Langlands Program was born.

Several generations of mathematicians have dedicated their lives to solving the problems put forward by Langlands. What was it that so inspired them? The answer is coming up in the next chapter.

[*] Note that here and below I use a minus sign (a dash) to represent negative numbers, rather than a hyphen. This conforms to the standard mathematical notation. In fact, there isn't really any difference between the two because $-N = 0 - N$.

Chapter 8

Magic Numbers

When we first talked about symmetries in Chapter 2, we saw that representations of a group named $SU(3)$ govern the behavior of elementary particles. The focus of the Langlands Program is also on representations of a group, but this time it is the Galois group of symmetries of a number field of the kind discussed in the previous chapter. It turns out that these representations form the "source code" of a number field, carrying all essential information about numbers.

Langlands' marvelous idea was that we can extract this information from objects of an entirely different nature: the so-called automorphic functions, which come from another field of mathematics called harmonic analysis. The roots of harmonic analysis are in the study of harmonics, which are the basic sound waves whose frequencies are multiples of each other. The idea is that a general sound wave is a superposition of harmonics, the way the sound of a symphony is a superposition of the harmonics corresponding to the notes played by various instruments. Mathematically, this means expressing a given function as a superposition of the functions describing harmonics, such as the familiar trigonometric functions sine and cosine. Automorphic functions are more sophisticated versions of these familiar harmonics. There are powerful analytic methods for doing calculations with these automorphic functions. And Langlands' surprising insight was that we can use these functions to learn about much more difficult questions in number theory. This way we find a hidden harmony of numbers.

I wrote in the Preface that one of the principal functions of mathematics is the ordering of information, or, as Langlands himself put it, "creating order from seeming chaos."[1] Langlands' idea is so powerful precisely because it helps organize seemingly chaotic data from number theory into regular patterns full of symmetry and harmony.

If we think of different fields of mathematics as continents, then number theory would be like North America and harmonic analysis like Europe. Over the years, it's been taking us less and less time to travel from one continent to the other. It used to take days by boat, and now only hours by plane. But imagine that a new technology was invented that would allow you to be instantly transported from anywhere in North America to someplace in Europe. That would be an equivalent of the connections discovered by Langlands.

I will now describe one of these breathtaking connections, which is closely related to Fermat's Last Theorem that we talked about in Chapter 6.

Fermat's Last Theorem is deceptively simple to state. It says that there are no natural numbers x, y, and z solving the equation

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

if n is greater than 2.

As I wrote, this result was guessed by the French mathematician Pierre Fermat more than 350 years ago, in 1637. He wrote about it on the margin of an old book he was reading, saying that he had found a “truly marvelous” proof of this statement, but “the margin is too small to contain it.” Call it a seventeenth-century Twitter-style proof: “I have found a marvelous proof of this theorem, but unfortunately I can’t write it here because it’s longer than one hundred and forty chara” – sorry, ran out of space.

There is little doubt that Fermat was mistaken. It took more than 350 years to find the real proof, and it is incredibly complicated. There are two main steps: first, in 1986, Ken Ribet showed that Fermat’s Last Theorem follows from the so-called Shimura–Taniyama–Weil conjecture.

(Perhaps, I should note that a mathematical conjecture is a statement that one expects to be true, but for which one does not yet know a proof. Once the proof is found, the conjecture becomes a theorem.[2])

What Ken Ribet showed was that if there exist natural numbers *x*, *y*, *z* solving Fermat’s equation, then, using these numbers, one can construct a certain cubic equation, which has a property precluded by the Shimura–Taniyama–Weil conjecture (I will explain below what this equation and this property are). If we know that the Shimura–Taniyama–Weil conjecture is true, then this equation cannot exist. But then the numbers *x*, *y*, *z* solving Fermat’s equation cannot exist either.[3]

Let’s pause for a minute and go over the logic of this argument one more time. In order to prove Fermat’s Last Theorem, we assume that it is false; that is, we suppose that there exist natural numbers *x*, *y*, *z* such that Fermat’s equation is satisfied. Then we associate to these numbers a cubic equation, which turns out to have a certain undesirable property. The Shimura–Taniyama–Weil conjecture tells us that such an equation *cannot exist*. But then these numbers *x*, *y*, *z* cannot exist either. Hence there can be no solutions to Fermat’s equation. Therefore Fermat’s Last Theorem is true! Schematically, the flow chart of this argument looks as follows (we abbreviate “Fermat’s Last Theorem” as FLT and “Shimura–Taniyama–Weil conjecture” as STWC):

$$\boxed{\text{FLT is false}} \Longrightarrow \boxed{\begin{matrix}\text{there exist } x,y,z \\ \text{solving the equation}\end{matrix}} \Longrightarrow \boxed{\begin{matrix}\text{a cubic} \\ \text{equation}\end{matrix}}$$

$$\overset{\text{STWC}}{\Longrightarrow} \boxed{\begin{matrix}\text{this cubic equation} \\ \text{cannot exist}\end{matrix}} \Longrightarrow \boxed{\begin{matrix}x,y,z \text{ solving the} \\ \text{equation cannot exist}\end{matrix}}$$

This kind of argument is called *proof by contradiction*. We start with the statement that is opposite to what we are trying to prove (in our case, it is the statement that there exist natural numbers *x*, *y*, *z* solving Fermat’s equation, which is opposite to what we want to prove). If, through a chain of implications, we then arrive at a statement that is demonstrably false (in our case, the existence of a cubic equation that is prohibited by the Shimura–Taniyama–Weil conjecture), then we conclude that the statement we started with is false. Hence the statement we wanted to prove (Fermat’s Last Theorem) is true.

What remains then to establish Fermat’s Last Theorem is to prove that the Shimura–Taniyama–Weil conjecture is true. Once this was understood (in 1986, after Ribet’s work), the search was on for a proof of the Shimura–Taniyama–Weil conjecture.

Several proofs had been announced over the years, but subsequent analysis showed that these proofs contained mistakes or gaps. In 1993, Andrew Wiles claimed that he had proved the conjecture, but a few months later it was found that there was a gap in his proof. For a while, it looked like his proof would be remembered alongside many other famous "non-proofs," in which gaps were found, but never closed.

Luckily, Wiles was able to close the gap within a year, with the help of another mathematician, Richard Taylor. Together, they completed the proof.[4] In a wonderful documentary film about Fermat's Last Theorem, Wiles gets emotional when he recounts this moment, and one can only imagine what a gut-wrenching experience it must have been for him.

Thus, the Shimura–Taniyama–Weil conjecture is a key result in proving Fermat's Last Theorem. It may also be viewed as a special case of the Langlands Program, and hence it provides an excellent illustration of the unexpected connections predicted by the Langlands Program.

The Shimura–Taniyama–Weil conjecture is a statement about certain equations. A large part of mathematics is in fact about solving equations. We want to know whether a given equation has a solution in a given domain; if so, can we find one? If there are several solutions, then how many? Why do some equations have solutions and some don't?

In the previous chapter we talked about polynomial equations on one variable, such as $x^2 = 2$. Fermat's Last Theorem is about an equation on three variables: $x^n + y^n = z^n$. And the Shimura–Taniyama–Weil conjecture is about a class of algebraic equations on two variables, such as this one:

$$y^2 + y = x^3 - x^2.$$

A solution of this equation is a pair of numbers x, y such that the left-hand side is equal to the right-hand side.

But what kind of numbers do we want x and y to be? There are several choices: one possibility is to say that x and y are natural numbers or integers. Another possibility is to say rational numbers. We can also look for solutions x, y that are real numbers, or even complex numbers – we will discuss this option in more detail in the next chapter.

It turns out that there is one more choice, which is less obvious but equally important: to consider solutions x, y "modulo N," for some fixed natural number N. This means that we look for integers x and y such that the left-hand side is equal to the right-hand side up to a number that is divisible by N.

For example, let's look for solutions modulo $N = 5$. There is one obvious solution $x = 0, y = 0$. And there are three other, slightly less obvious, solutions: $x = 0, y = 4$ is a solution modulo 5 because the left-hand side is then 20 and the right-hand side is 0. The difference between the left- and right-hand sides is 20, which is divisible by 5. So this is indeed a solution of the equation modulo 5. By a similar argument, $x = 1, y = 0$ and $x = 1, y = 4$ are also solutions modulo 5.

We have already discussed this kind of arithmetic in Chapter 2 when we talked about the group of rotations of a round table. We saw then that the addition of angles was done "modulo 360." That is to say, if the result of addition of two angles is greater than 360 degrees, we subtract from it 360 to bring it into the range from 0 to 360. For example, rotation by 450 degrees is the same as rotation by 90 degrees, because $450-360 = 90$.

We also encounter this arithmetic when we use the clock. If we start working at 10 o'clock in the morning and work for 8 hours, when do we finish? Well, $10 + 8 = 18$, so a natural thing to say would be: "We finish at 18 o'clock." This would be perfectly fine to say in France where they record hours as numbers from 0 to 24 (actually, not so fine, because a working day in France is usually limited to seven hours). But in the U.S. we say: "We finish at 6 pm." How do we get 6 out of 18? We subtract 12 from it: $18-12 = 6$.

So we use the same idea with hours as we do with angles. In the first case, we do addition "modulo 360." In the second case, we do addition "modulo 12."

Likewise, we can do addition modulo any natural number N. Consider the set of all consecutive whole numbers between 0 and $N - 1$,

$$\{0,1,2,\ldots,N-2,N-1\}.$$

If $N = 12$, this is the set of possible hours. In general, the role of 12 is played by number N, so that it's not 12 that takes us back to 0, but N.

We define addition on the set of these numbers in the same way as for the hours. Given any two numbers from this set, we add them up, and if the result is greater than N, we subtract N from it to get a number from the same set. This operation makes this set into a group. The identity element is the number 0: adding it to any other number does not change it. Indeed, we have $n + 0 = n$. And for any number n from our set, its "additive inverse" is $N - n$, because $n + (N - n) = N$, which is the same as 0 according to our rules.

For example, let's take $N = 3$. Then we have the set {0, 1, 2} and addition modulo 3. For example, we have

$$2+2=1 \qquad \text{modulo} \quad 3$$

in this system, because 2 + 2 = 4, but since 4 = 3 + 1, the number 4 is equal to 1 modulo 3.

So if someone says to you: "2 plus 2 equals 4" to indicate a well-established fact, you can now say (with a condescending smile if you like): "Well, actually, that's not always true." And if they ask you to explain what you mean, you can tell them, "If you do addition modulo 3, then 2 plus 2 is equal to 1."

Given any two numbers from the above set, we can also multiply them. The result may not be between 0 and $N - 1$, but there will be a unique number in this range that will differ from the result of multiplication by something divisible by N. However, in general, the set $\{1, 2,\ldots, N - 1\}$ is not a group with respect to multiplication. We do have the identity element: number 1. But not every element has the multiplicative inverse modulo N. This happens if and only if N is a *prime number*, that is, a number that is not divisible by any other natural number other than 1 and itself.[5]

The first few primes are 2, 3, 5, 7, 11, 13,... (It is customary to exclude 1 from this list.) Even natural numbers, except for 2, are not prime, because they are divisible by 2, and 9 isn't prime, because it is divisible by 3. There are in fact infinitely many primes – no matter how large a prime number is, there is another prime number that is even larger.[6] Primes, because they are indivisible, are the elementary particles of the world of natural numbers; every other natural number, in fact, can be written, in a unique way, as the product of prime numbers. For example, $60 = 2 \cdot 2 \cdot 3 \cdot 5$.

Let us fix a prime number. As is customary, we will denote it by p. Then we consider the set of all consecutive whole numbers between 0 to $p - 1$; that is,

$$\{0,1,2,3,4,\ldots,p-2,p-1\}.$$

And we consider two operations on them: addition and multiplication modulo p.

As have seen above, this set is a group with respect to addition modulo p. What is even more remarkable is that if we remove number 0 and consider the set of consecutive whole numbers between 1 and $p - 1$, that is $\{1,2,\ldots,p-1\}$, we obtain a group with respect to multiplication modulo p. The element 1 is the multiplicative identity (this is clear), and I claim that any natural number between 1 and p − 1 has a multiplicative inverse.[7]

For example, if p = 5, we find that

$$2 \cdot 3 = 1 \qquad \text{modulo} \quad 5,$$

and

$$4 \cdot 4 = 1 \qquad \text{modulo} \quad 5,$$

so that the multiplicative inverse of 2 modulo 5 is 3, and 4 is its own inverse modulo 5. It turns out that this is true in general.[8]

In our everyday life, we are used to numbers that are integers or fractions. Sometimes, we use numbers like $\sqrt{2}$. But we have now discovered a numerical system of an entirely different nature: the finite set of numbers {0,1,2,..., p−1}, where p is a prime, on which we have the operations of addition and multiplication modulo p. It is called the *finite field* with p elements. These finite fields form an important archipelago in the world of numbers – one that, unfortunately, most of us are never told exists.

Even though these numerical systems look very different from the numerical systems we are used to, such as rational numbers, they have the same salient properties: they are closed under the operations of addition, subtraction, multiplication, and division.[9] Therefore, everything we can do with the rational numbers may also be done with these, more esoteric looking, finite fields.

Actually, they are not so esoteric any more, having found important applications – most notably, in cryptography. When we make a purchase online and enter our credit card number, this number gets encrypted using the arithmetic modulo primes, which is dictated by the equations very much like the ones we have looked at above (see the description of the RSA encryption algorithm in endnote 7 to Chapter 14).

Let's go back to the cubic equation

$$y^2 + y = x^3 - x^2$$

that we considered above. Let us look for solutions of this equation modulo p, for various primes p. For example, we have seen above that there are 4 solutions modulo 5. But note that the solutions modulo p = 5 are not necessarily solutions modulo other primes (say, p = 7 or p = 11). So these solutions do depend on the prime p modulo which we do the arithmetic.

The question we are going to ask now is the following: how does the number of solutions of this equation, taken modulo p, depend on p? For small p, we can count them explicitly (perhaps, with the aid of a computer), so we can actually compile a small table.

Mathematicians have known for some time that the number of solutions of an equation of this type modulo p is roughly equal to p. Let's denote the "deficit," the number by which the actual number of solutions differs from the expected number of solutions (namely, p), by a_p. This means that the number of solutions of the above equation modulo p is equal to $p - a_p$. The numbers a_p could be positive or negative for a given p. For example, we found above that for p = 5 there are 4 solutions. Since 4 = 5−1, we obtain that a_5 = 1.

We can find the numbers a_p for small primes on a computer. They seem to be random. There does not appear to be any natural formula or rule that would enable us to compute them. What's worse, the computation very quickly becomes immensely complicated.

But what if I told you that there was in fact a simple rule that generated the numbers a_p all at once?

In case you are wondering what exactly I mean here by a "rule" generating these numbers, let's consider a more familiar sequence, the so-called Fibonacci numbers:

$$1, 1, 2, 3, 5, 8, 13, 21, 34, \ldots$$

Named after an Italian mathematician who introduced them in his book published in 1202 (in the context of a problem of mating rabbits, no less), Fibonacci numbers are ubiquitous in nature: from petal arrangements in flowers to the patterns on the surface of a pineapple. They also have many applications, such as the "Fibonacci retracement" in the technical analysis of stock trading.

The Fibonacci numbers are defined as follows: the first two of them are equal to 1. Each number after that is equal to the sum of the preceding two Fibonacci numbers. For example, $2 = 1 + 1, 3 = 2 + 1, 5 = 3 + 2$, and so on. If we denote the nth Fibonacci number by F_n, then we have $F_1 = 1, F_2 = 1$ and

$$F_n = F_{n-1} + F_{n-2}, \qquad n > 2.$$

In principle, this rule enables us to find the nth Fibonacci number for any n. But in order to do this, we have to first find all Fibonacci numbers F_i for i between 1 and $n - 1$.

However, it turns out that these numbers could also be generated in the following way. Consider the series

$$q + q(q + q^2) + q(q + q^2)^2 + q(q + q^2)^3 + q(q + q^2)^4 + \ldots.$$

In words, we multiply an auxiliary variable q by the sum of all powers of the expression $(q + q^2)$. If we open the brackets, we obtain an infinite series, whose first terms are

$$q + q^2 + 2q^3 + 3q^4 + 5q^5 + 8q^6 + 13q^7 + \ldots$$

For example, let's compute the term with q^3. It can only occur in q, $q(q + q^2)$, and $q(q + q^2)^2$. (Indeed, all other expressions that appear in the defining sum, such as $q(q + q^2)^3$, will only contain powers of q greater than 3.) The first of these does not contain q^3, and each of the other two contains q^3 once. Their sum yields $2q^3$. We obtain in a similar way other terms of the series.

Analyzing the first terms of this series, we find that for n between 1 and 7, the coefficient in front of q^n is the nth Fibonacci number F_n. For example, we have the term $13q^7$ and $F_7 = 13$. It turns out that this is true for all n. For this reason, mathematicians call this infinite series the *generating function* of the Fibonacci numbers.

This remarkable function can be used to give an effective formula for calculating the nth Fibonacci number without any reference to the preceding Fibonacci numbers.[10] But even putting the computational aspects aside, we can appreciate the value added by this generating function: instead of giving a self-referential recursive procedure, the generating function beholds all Fibonacci numbers at once.

Let's go back to the numbers a_p counting the solutions of the cubic equation modulo primes. Think of these numbers as analogues of the Fibonacci numbers (let's ignore the fact that the numbers a_p are labeled by the prime numbers p, whereas the Fibonacci numbers F_n are labeled by all natural numbers n).

It seems nearly unbelievable that there would be a rule generating these numbers. And yet, German mathematician Martin Eichler discovered one in 1954.[11] Namely, consider the following generating function:

$$q(1-q)^2(1-q^{11})^2(1-q^2)^2(1-q^{22})^2(1-q^3)^2(1-q^{33})^2(1-q^4)^2(1-q^{44})^2 \dots$$

In words, this is q times the product of factors of the form $(1-q^a)^2$, with a going over the list of numbers of the form n and $11n$, where $n = 1,2,3,\dots$. Let's open the brackets, using the standard rules:

$$(1-q)^2 = 1-2q+q^2, \qquad (1-q^{11})^2 = 1-2q^{11}+q^{22}, \qquad \dots$$

and then multiply all the factors. Collecting the terms, we obtain an infinite sum, which begins like this:

$$q - 2q^2 - q^3 + 2q^4 + q^5 + 2q^6 - 2q^7 - 2q^9 - 2q^{10} + q^{11} - 2q^{12} + 4q^{13} + \dots$$

and the ellipses stand for the terms with the powers of q greater than 13. Though this series is infinite, each coefficient is well-defined because it is determined by finitely many factors in the above product. Let us denote the coefficient in front of q^m by b_m. So we have $b_1 = 1$, $b_2 = -2$, $b_3 = -1$, $b_4 = 2$, $b_5 = 1$, etc. It is easy to compute them by hand or on a computer.

An astounding insight of Eichler was that for all prime numbers p, the coefficient b_p is equal to a_p. In other words, $a_2 = b_2, a_3 = b_3, a_5 = b_5, a_7 = b_7$, and so on.

Let's check, for example, that this is true for $p = 5$. In this case, looking at the generating function we find that the coefficient in front of q^5 is $b_5 = 1$. On the other hand, we have seen that our cubic equation has 4 solutions modulo $p = 5$. Therefore $a_5 = 5 - 4 = 1$, so indeed $a_5 = b_5$.

We started out with what looked like a problem of infinite complexity: counting solutions of the cubic equation

$$y^2 + y = x^3 - x^2$$

modulo p, for all primes p. And yet, all information about this problem is contained in a single line:

$$q(1-q)^2(1-q^{11})^2(1-q^2)^2(1-q^{22})^2(1-q^3)^2(1-q^{33})^2(1-q^4)^2(1-q^{44})^2 \dots$$

This one line is a secret code containing all information about the numbers of solutions of the cubic equation modulo all primes.

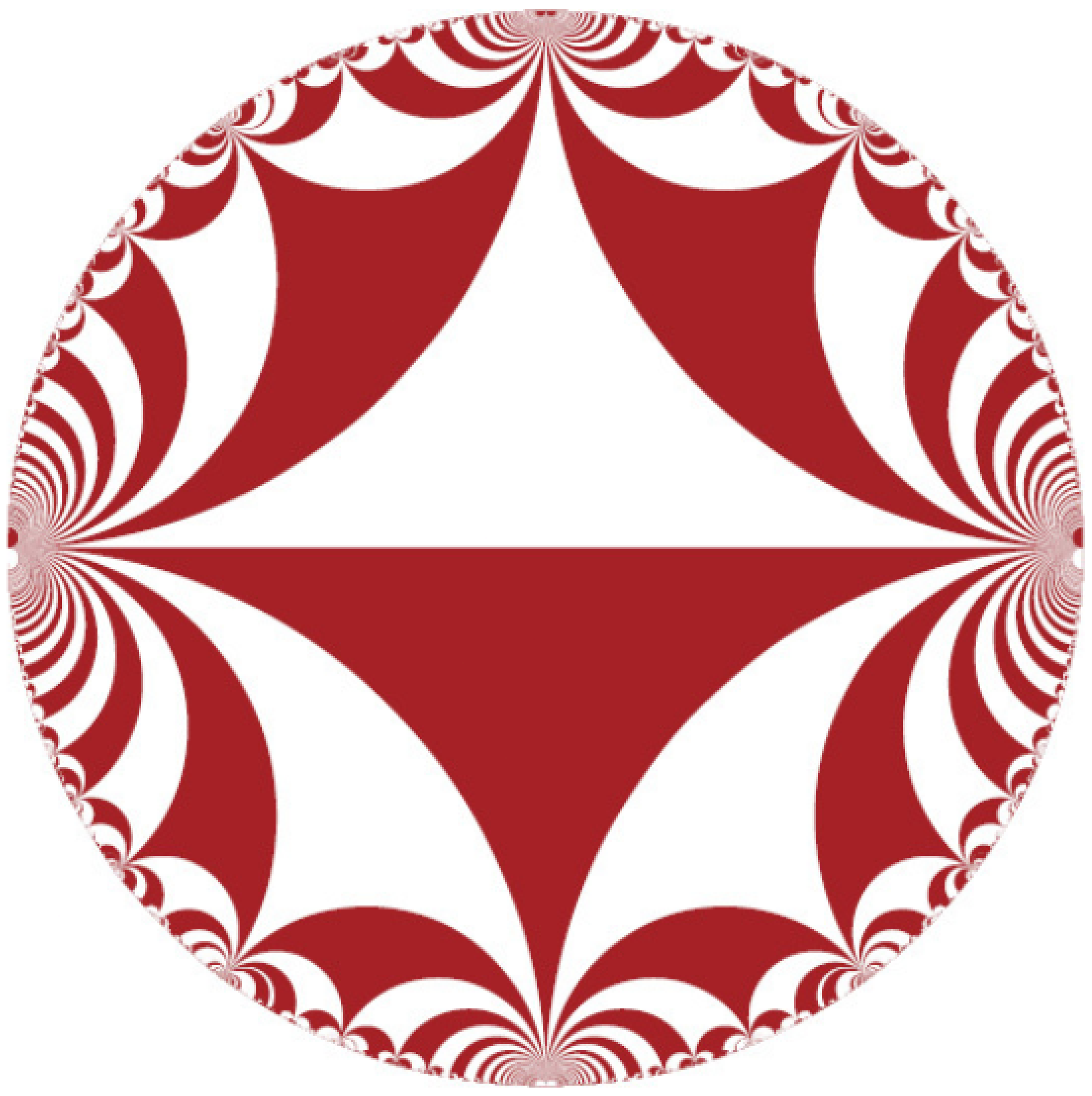
A useful analogy would be to think of the cubic equation like a sophisticated biological organism, and its solutions as various traits of this organism. We know that all of these traits are encoded in the DNA molecule. Likewise, all the complexity of our cubic equation turns out to be encoded in a generating function, which is like the DNA of this equation. Furthermore, this function is defined by a simple rule.

What's even more fascinating is that if q is a number whose absolute value is less than 1, then the above infinite sum converges to a well-defined number. So we obtain a function in q, and this function turns out to have a very special property that is similar to the periodicity of the familiar trigonometric functions, sine and cosine.

The sine function $\sin(x)$ is periodic with the period 2π, that is to say, $\sin(x + 2\pi) = \sin(x)$. But then also $\sin(x+4\pi) = \sin(x)$, and more generally $\sin(x+2\pi n) = \sin(x)$ for any integer n. Think about it this way: each integer n gives rise to a symmetry of the line: every point x on the line is shifted to $x + 2\pi n$. Therefore, the group of all integers is realized as a group of symmetries of the line. The periodicity of the sine function means that this function is invariant under this group.

Likewise, the Eichler generating function of the variable q written above turns out to be invariant under a certain symmetry group. Here we should take q to be not a real, but rather a complex number (we will discuss this topic in the next chapter). Then we can view q not as a point on the line, as in the case of the sine function, but as a point inside a unit disc on the complex plane. The symmetry property is similar: on this disc there is a group of symmetries, and our function is invariant under this group.[12] A function with this kind of invariance property is called a *modular form*.

This symmetry group of the disc is very rich. To get an idea of what it is, let's look at this picture, on which the disc is broken into infinitely many triangles.[13]



The symmetries act on the disc by exchanging these triangles. In fact, for any two triangles, there is a symmetry exchanging them. Though these symmetries of the disc are quite sophisticated, this is analogous to how, when the group of integers acts on the line, its symmetries move around the intervals $[2\pi m, 2\pi(m + 1)]$. The sine function is invariant under those symmetries, whereas the Eichler generating function is invariant under symmetries of the disc.

As I mentioned at the beginning of this chapter, the sine function is the simplest example of a "harmonic" (basic wave) that is used in the harmonic analysis on the line. Likewise, the Eichler function, together with other modular forms, are the harmonics that appear in the harmonic analysis on the unit disc.

The magnificent insight of Eichler was that the seemingly random numbers of solutions of a cubic equation modulo primes come from a single generating function, which obeys an exquisite symmetry – revealing a hidden harmony and order in those numbers. Similarly, as if in a stroke of black magic, the Langlands Program organizes previously inaccessible information into regular patterns, weaving a delicate tapestry of numbers, symmetries, and equations.

When I first talked about mathematics at the beginning of this book, you may have wondered what I meant by a mathematical result being "beautiful" or "elegant." This is what I meant. The fact that these highly abstract notions coalesce in such refined harmony is absolutely mind-boggling. It points to something rich and mysterious lurking beneath the surface, as if the curtain had been lifted and we caught glimpses of the reality that had been carefully hidden from us. These are the wonders of modern math, and of the modern world.

One might also ask whether, in addition to possessing innate beauty and establishing a surprising link between areas of mathematics that seem to be far removed from each other, this result has any practical applications. This is a fair question. At present, I am not aware of any. But cubic equations over finite fields of p elements of the kind we have considered above (which give rise to the so-called elliptic curves) are widely used in cryptography.[14] So I would not be surprised if the analogues of Eichler's result will also one day find applications as powerful and ubiquitous as encryption algorithms.

The Shimura–Taniyama–Weil conjecture is a generalization of Eichler's result. It says that for *any* cubic equation like the one above (subject to some mild conditions), the numbers of solutions modulo primes are the coefficients of a modular form. Moreover, there is a one-to-one correspondence between the cubic equations and the modular forms of a certain kind.

What do I mean here by a one-to-one correspondence? Suppose that we have five pens and five pencils. We can assign a pencil to each pen in such a way that each pencil is assigned to one and only pen. This is called a one-to-one correspondence.

There are many different ways to do it. But suppose that under our one-to-one correspondence each pen has exactly the same length as the pencil assigned to it. We will then call the length an "invariant" and say that our correspondence preserves this invariant. If all pens have different lengths, the one-to-one correspondence will be uniquely determined by this property.

Now, in the case of the Shimura–Taniyama–Weil conjecture, the objects on one side are the cubic equations such as the one above. These will be our pens, and for each of them the numbers a_p will be the invariants attached to it. (It's like the length of a pen, except that now there isn't just one invariant, but a whole collection labeled by primes p.)

The objects on the other side of the correspondence are modular forms. These will be our pencils, and for each of them, the coefficients b_p will be the invariants attached to it (like the length of a pencil).

The Shimura–Taniyama–Weil conjecture says that there is a one-to-one correspondence between these objects preserving these invariants:

$$\boxed{\begin{array}{c}\text{cubic equations}\\ \{\text{numbers of solutions } a_p\}\end{array}} \longleftrightarrow \boxed{\begin{array}{c}\text{modular forms}\\ \{\text{coefficients } b_p\}\end{array}}$$

That is to say, for any cubic equation there exists a modular form such that $a_p = b_p$ for all primes p, and vice versa.[15]

Now I can explain the link between the Shimura–Taniyama–Weil conjecture and Fermat's Last Theorem: starting from a solution of the Fermat equation, we can construct a certain cubic equation.[16] However, Ken Ribet showed that the numbers of solutions of this cubic equation modulo primes cannot be the coefficients of a modular form whose existence is stipulated by the Shimura–Taniyama–Weil conjecture. Once this conjecture is proved, we conclude that such a cubic equation cannot exist. Therefore, there are no solutions to the Fermat equation.

The Shimura–Taniyama–Weil conjecture is a stunning result because the numbers a_p come from the study of solutions of an equation modulo primes – they are from the world of number theory – and the numbers b_p are the coefficients of a modular form, from the world of harmonic analysis. These two worlds seem to be light years apart, and yet it turns out that they describe one and the same thing!

The Shimura–Taniyama–Weil conjecture may be recast as a special case of the Langlands Program. In order to do that, we replace each of the cubic equations appearing in the Shimura–Taniyama–Weil conjecture by a certain two-dimensional representation of the Galois group. This representation is naturally obtained from the cubic equation, and the numbers a_p can be attached directly to this representation (rather than the cubic equation). Therefore the conjecture may be expressed as a relation between two-dimensional representations of the Galois group and modular forms.

(I recall from Chapter 2 that a two-dimensional representation of a group is a rule that assigns a symmetry of a two-dimensional space (that is, a plane) to each element of this group. For example, in Chapter 2 we talked about a two-dimensional representation of the circle group.)

Even more generally, conjectures of the Langlands Program relate, in unexpected and profound ways, n-dimensional representations of the Galois group (which generalize the two-dimensional representations corresponding to the cubic equations in the Shimura–Taniyama–Weil conjecture) and the so-called *automorphic functions* (which generalize the modular forms in the Shimura–Taniyama–Weil conjecture):

$$\boxed{\begin{array}{c}\text{representations of}\\ \text{the Galois group}\end{array}} \longleftrightarrow \boxed{\begin{array}{c}\text{automorphic}\\ \text{functions}\end{array}}$$

Though there is little doubt that these conjectures are true, most of them are still unproved to this day, despite an enormous effort by several generations of mathematicians in the past forty-five years.

You may be wondering: how could one come up with these kinds of conjectures in the first place?

This is really a question about the nature of mathematical insight. The ability to see patterns and connections that no one had seen before does not come easily. It is usually the product of months, if not years, of hard work. Little by little, the inkling of a new phenomenon or a theory emerges, and at first you don't believe it yourself. But then you say: "what if it's true?" You try to test the idea by doing sample calculations. Sometimes these calculations are hard, and you have to navigate through mountains of heavy formulas. The probability of making a mistake is very high, and if it does not work at first, you try to redo it, over and over again.

More often than not, at the end of the day (or a month, or a year), you realize that your initial idea was wrong, and you have to try something else. These are the moments of frustration and despair. You feel that you have wasted an enormous amount of time, with nothing to show for it. This is hard to stomach. But you can never give up. You go back to the drawing board, you analyze more data, you learn from your previous mistakes, you try to come up with a better idea. And every once in a while, suddenly, your idea starts to work. It's as if you had spent a fruitless day surfing, when you finally catch a wave: you try to hold on to it and ride it for as long as possible. At moments like this, you have to free your imagination and let the wave take you as far as it can. Even if the idea sounds totally crazy at first.

The statement of the Shimura–Taniyama–Weil conjecture must have sounded crazy to its creators. How could it not? Yes, the conjecture had its roots in earlier results, such as those by Eichler that we discussed above (which were subsequently generalized by Shimura), which showed that for *some* cubic equations, the numbers of solutions modulo p were recorded in the coefficients of a modular form. But the idea that this was true for *any* cubic equation must have sounded totally outrageous at the time. This was a leap of faith, first made by the Japanese mathematician Yutaka Taniyama, in the form of a question that he posed at the International Symposium on Algebraic Number Theory held in Tokyo in September 1955.

I've always wondered: what did it take for him to come to *believe* that this wasn't crazy, but real? To have the courage to say it publicly?

We'll never know. Unfortunately, not long after his great discovery, in November 1958, Taniyama committed suicide. He was only thirty-one. To add to the tragedy, shortly afterward the woman whom he was planning to marry also took her life, leaving the following note:[17]

> We promised each other that no matter where we went, we would never be separated. Now that he is gone, I must go too in order to join him.

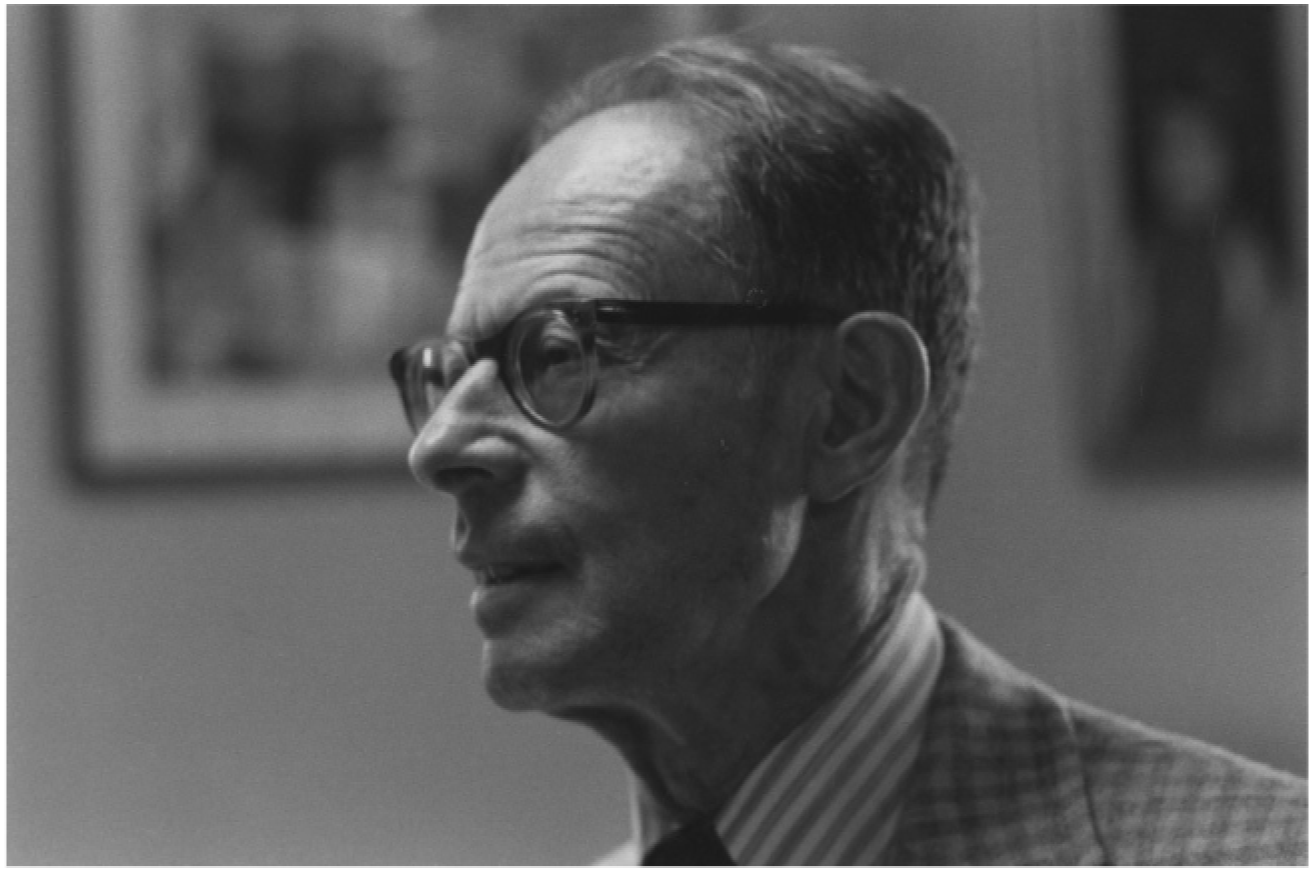
The conjecture was subsequently made more precise by Taniyama's friend and colleague Goro Shimura, another Japanese mathematician. Shimura has worked most of his life at Princeton University and is currently an emeritus professor there. He has made major contributions to mathematics, many pertinent to the Langlands Program, and several fundamental concepts in this area carry his name (such as the "Eichler–Shimura congruence relations" and "Shimura varieties").

In his thoughtful essay about Taniyama, Shimura made this striking comment:[18]

> Though he was by no means a sloppy type, he was gifted with the special capability of making many mistakes, mostly in the right direction. I envied him for this, and tried in vain to imitate him, but found it quite difficult to make good mistakes.

In the words of Shimura, Taniyama "wasn't very careful when he stated his problem" at the Symposium in Tokyo in September 1955.[19] Some corrections had to be made. And yet, this was a revolutionary insight, which led to one of the most significant achievements in mathematics of the twentieth century.

The third person whose name is attached to the conjecture is André Weil, whom I have mentioned earlier. He is one of the giants of mathematics in the twentieth century. Known for his brilliance as well as his temper, he was born in France and came to the United States during World War II. After holding academic appointments at various American universities, he settled at the Institute for Advanced Study in Princeton in 1958 and stayed there until his death in 1998, at age 92.



André Weil, 1981. Photo by Herman Landshoff. From the Shelby White and Leon Levy Archives Center, Institute for Advanced Study, Princeton.

Weil is particularly relevant to the Langlands Program, and not just because the famous letter in which Robert Langlands first formulated his ideas was addressed to him, or because of the Shimura–Taniyama–Weil conjecture. The Langlands Program is best seen through the prism of the "big picture" of mathematics that André Weil outlined in a letter to his sister. We will talk about it in the next chapter. This will be our stepping stone toward bringing the Langlands Program into the realm of geometry.

Chapter 9

Rosetta Stone

In 1940, during the war, André Weil was imprisoned in France for refusing to serve in the army. As the obituary published in *The Economist* put it,[1]

> [Weil] had been deeply struck.... by the damage wreaked upon mathematics in France by the first world war, when "a misguided notion of equality in the face of sacrifice" led to the slaughter of the country's young scientific elite. In the light of this, he believed he had a duty, not just to himself but also to civilization, to devote his life to mathematics. Indeed, he argued, to let himself be diverted from the subject would be a sin. When others raised the objection "but if everybody were to behave like you...", he replied that this possibility seemed to him so implausible that he did not feel obliged to take it into account.

While in prison, Weil wrote a letter to his sister Simone Weil, a famous philosopher and humanist. This letter is a remarkable document; in it, he tries to explain in fairly elementary terms (accessible even to a philosopher – just kidding!) the "big picture" of mathematics as he saw it. Doing so, he set a great example to follow for all mathematicians. I sometimes joke that perhaps we should jail some of the leading mathematicians to force them to express their ideas in accessible terms, the way Weil did.

Weil writes in the letter about the role of analogy in mathematics, and he illustrates it by the analogy that interested him the most: between number theory and geometry.

This analogy proved to be extremely important for the development of the Langlands Program. As we discussed earlier, the roots of the Langlands Program are in number theory. Langlands conjectured that hard questions of number theory, such as the counting of solutions of equations modulo primes, can be solved by using methods of harmonic analysis – more specifically, the study of automorphic functions. This is exciting: first of all, it gives us a new way to solve what previously looked like intractable problems. And second, it points to deep and fundamental connections between different areas of mathematics. So naturally, we want to know what is really going on here: why might these hidden connections exist? And we still don't fully understand it. Even the Shimura–Taniyama–Weil conjecture took a very long time to be resolved. And it's only a special case of the general Langlands conjectures. There are hundreds and thousands of similar statements that are still not proved.

So how should we approach these difficult conjectures? One way is just to keep working hard and try to come up with new ideas and insights. This has been happening, and significant progress has been made. Another possibility is to try to expand the scope of the Langlands Program. Since it points to some essential structures in number theory and harmonic analysis and connections between them, chances are that similar structures and connections can also be found between other fields of mathematics.

This has indeed turned out to be the case. It was gradually realized that the same mysterious patterns may be observed in other areas of mathematics, such as geometry, and even in quantum physics. When we learn something about these patterns in one area, we get hints about their meaning in other areas. I have written earlier that the Langlands Program is a Grand Unified Theory of mathematics. What I mean by this is that the Langlands Program points to some universal phenomena and connections between these phenomena across different fields of mathematics. And I believe that it holds the keys to understanding what mathematics is really about, far beyond the original Langlands conjectures.

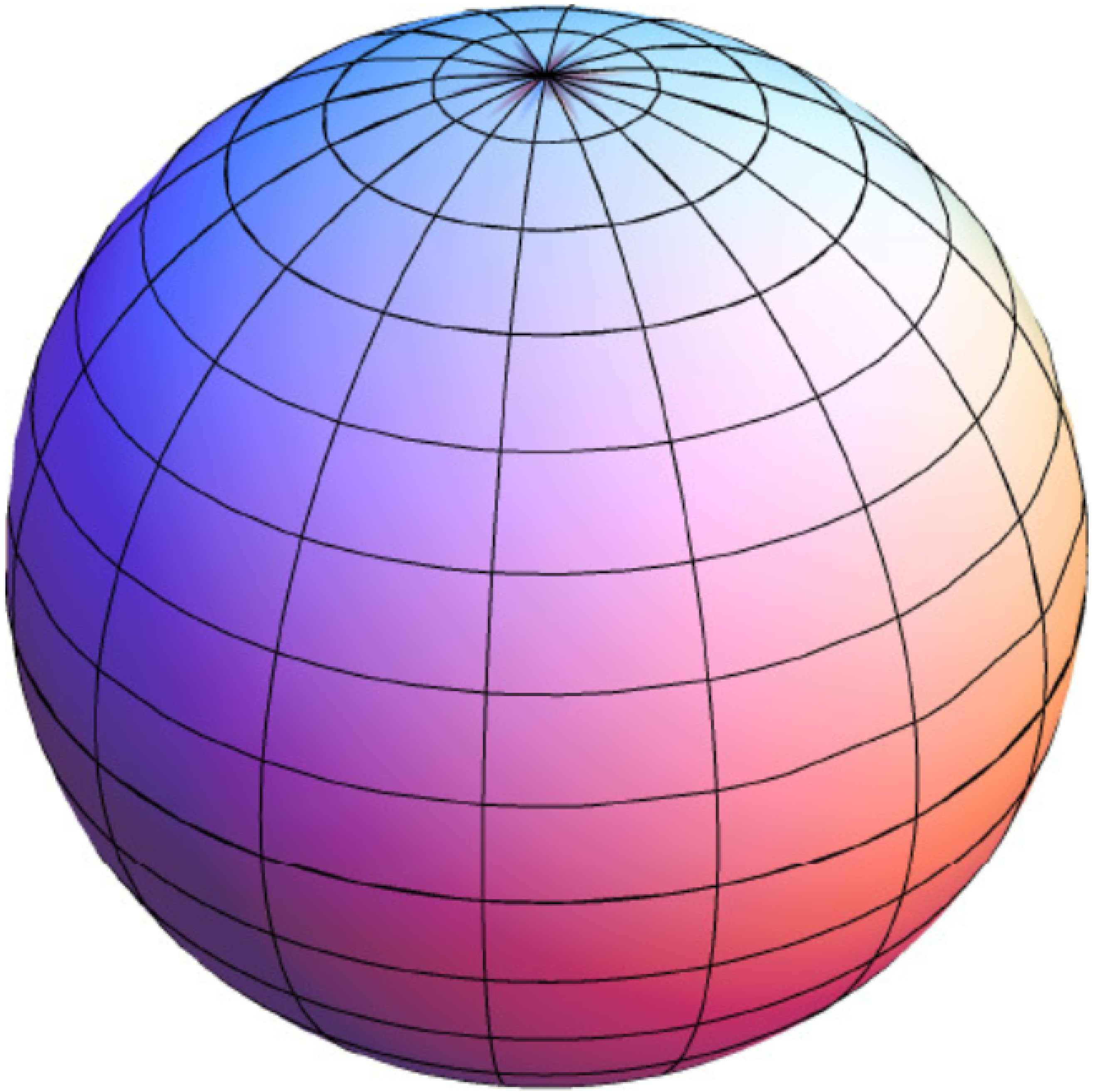
The Langlands Program is now a vast subject. There is a large community of people working on it in different fields: number theory, harmonic analysis, geometry, representation theory, mathematical physics. Although they work with very different objects, they are all observing similar phenomena. And these phenomena give us clues to understanding how these diverse domains are interconnected, like parts of a giant jigsaw puzzle.

My entry point to the Langlands Program was through my work on Kac–Moody algebras, which I will describe in detail in the next few chapters. But the more I learned about the Langlands Program, the more I got excited by how ubiquitous it is in mathematics.

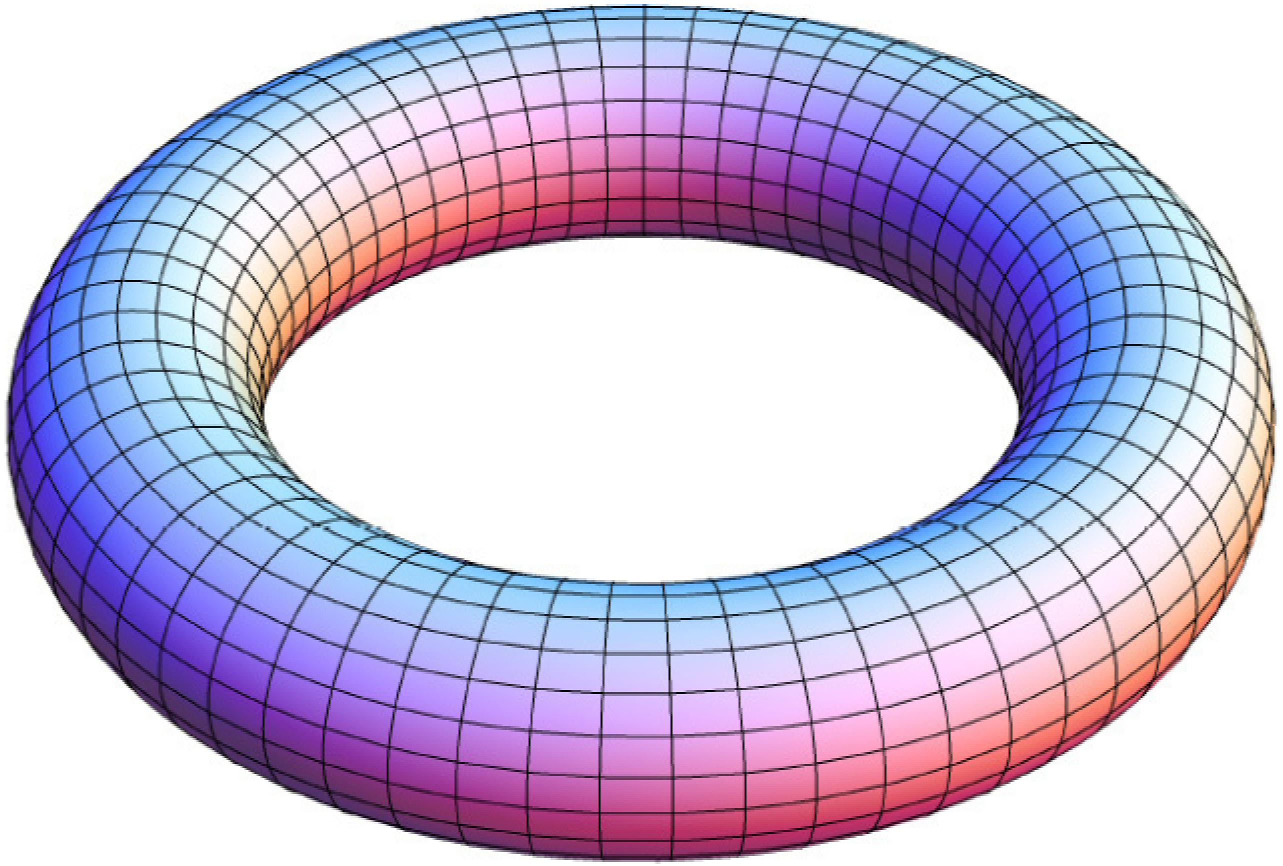
Think of different areas of modern math as languages. We have sentences from these languages that we think mean the same thing. We put them next to each other, and little by little we start developing a dictionary that allows us to translate between different areas of mathematics. André Weil gave us a suitable framework for understanding connections between number theory and geometry, a kind of "Rosetta stone" of modern math.

On one side, we have objects of number theory: rational numbers and other number fields that we discussed in the previous chapter, such as the one obtained by adjoining $\sqrt{2}$, and their Galois groups.

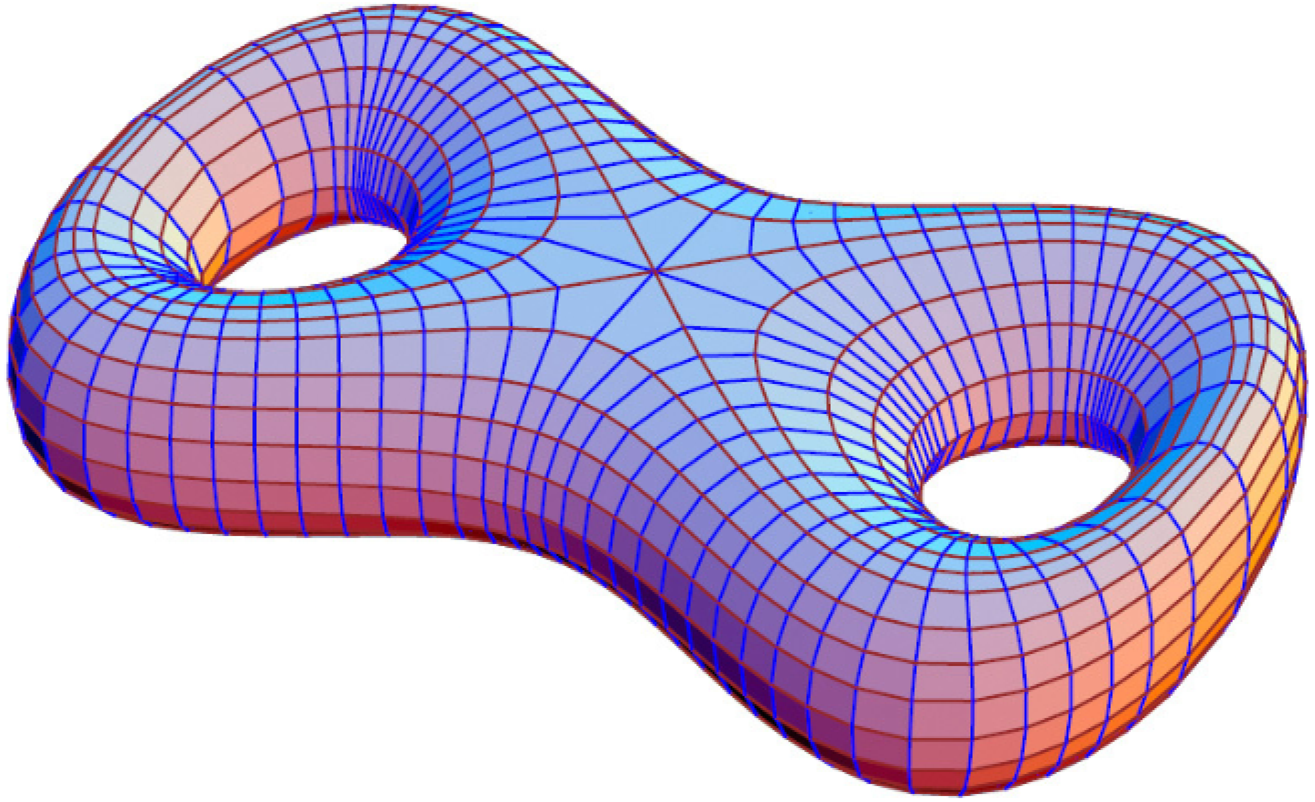
On the other side, we have the so-called Riemann surfaces. The simplest example is the sphere.[2]



The next example is the torus, the surface in the shape of a donut. I want to emphasize that we are considering here the *surface* of the donut, not its interior.



The next example is the surface of a Danish pastry, shown on the next picture (or you can think of it as the surface of a pretzel).



The torus has one "hole," and the Danish has two "holes." There are also surfaces having n holes for any $n = 3, 4, 5,...$ Mathematicians call the number of holes the *genus* of the Riemann surface.[*] They are named after the German mathematician Bernhard Riemann, who lived in the nineteenth century. His work opened up several important directions in mathematics. Riemann's theory of curved spaces, which we now call Riemannian geometry, is the cornerstone of Einstein's general relativity theory. Einstein's equations describe the force of gravity in terms of the so-called Riemann tensor expressing the curvature of space-time.

At first glance, number theory has nothing in common with the Riemann surfaces. However, it turns out that there are many analogies between them. The key point is that there is another class of objects between these two.

To see this, we have to realize that a Riemann surface may be described by an algebraic equation. For example, consider again a cubic equation such as

$$y^2 + y = x^3 - x^2.$$

As we noted earlier, when we talk about solutions of such an equation, it is important to specify to what numerical system they belong. There are many choices, and different choices give rise to different mathematical theories.

In the previous chapter, we discussed solutions modulo prime numbers, and that's one theory. But we can also look for solutions in *complex numbers*. That's another theory, which yields Riemann surfaces.

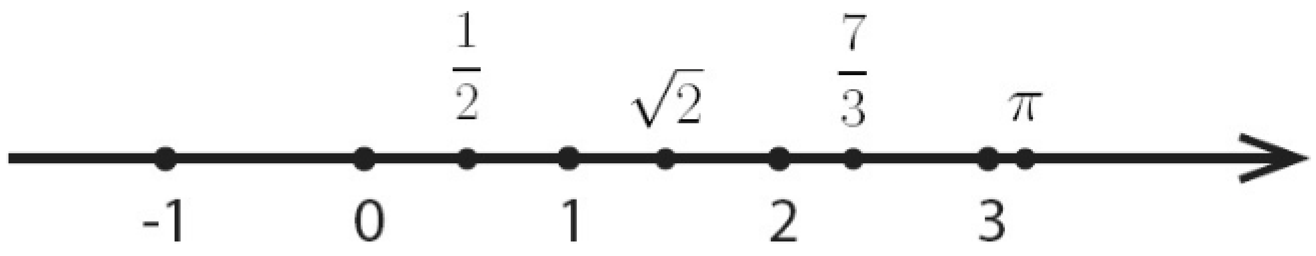
People often ascribe almost mystical qualities to complex numbers, as if these are some incredibly complicated objects. The truth is that they are no more complicated than the numbers we discussed in the previous chapter when trying to make sense of the square root of 2.

Let me explain. In the previous chapter, we adjoined to the rational numbers two solutions of the equation $x^2 = 2$, which we denoted by $\sqrt{2}$ and $-\sqrt{2}$. Now, instead of looking at the equation $x^2 = 2$, we look at the equation $x^2 = -1$. Does it look much more complicated than the previous one? No. It has no solutions among rational numbers, but we are not afraid of this. Let's adjoin the two solutions of this equation to the rational numbers. Denote them by $\sqrt{-1}$ and $-\sqrt{-1}$. They solve the equation $x^2 = -1$, that is,

$$\sqrt{-1}^2 = -1, \qquad (-\sqrt{-1})^2 = -1.$$

There is only a minor difference with the previous case. The number $\sqrt{2}$ is not rational, but it is a *real* number, so by adjoining it to the rational numbers, we don't leave the realm of real numbers.

We can think of real numbers geometrically as follows. Draw a line and mark two points on it, which will represent numbers 0 and 1. Then mark the point to the right of 1 whose distance to 1 is equal to the distance between 0 and 1. This point will represent number 2. We represent all other integers in a similar fashion. Next, we mark rational numbers by subdividing the intervals between the points representing the integers. For example, the number $\frac{1}{2}$ is exactly halfway between 0 and 1; the number $\frac{7}{3}$ is one-third of the way from 2 to 3, and so on. Now, the real numbers are, intuitively, in one-to-one correspondence with all points of this line.[3]



Recall that we encounter number $\sqrt{2}$ as the length of the hypotenuse of the right triangle with legs of length 1. So we mark $\sqrt{2}$ on the line of real numbers by finding a point to the right of 0 whose distance to 0 is equal to the length of this hypotenuse. Likewise, we can mark[4] on this line the number π, which is the circumference of a circle of diameter 1.

On the other hand, the equation $x^2 = -1$ has no solutions among rational numbers, and it also has no solutions among real numbers. Indeed, the square of any real number must be positive or 0, so it cannot be equal to −1. So unlike $\sqrt{2}$ and $-\sqrt{2}$, the numbers $\sqrt{-1}$ and $-\sqrt{-1}$ are not real numbers. But so what? We follow the same procedure and introduce them in exactly the same way as we introduced the numbers $\sqrt{2}$ and $-\sqrt{2}$. And we use the same rules to do arithmetic with these new numbers.

Let's recall how we argued before: we noticed that the equation $x^2 = 2$ had no solutions among the rational numbers. So we created two solutions of this equation, called them $\sqrt{2}$ and $-\sqrt{2}$, and adjoined them to the rational numbers, creating a new numerical system (which we then called a number field). Likewise, now we take the equation $x^2 = -1$ and notice that it also has no solutions among rational numbers. So we *create* two solutions of this equation, denote them by $\sqrt{-1}$ and $-\sqrt{-1}$, and adjoin them to the rational numbers. It's exactly the same procedure! Why should we think of this new numerical system as anything more complicated than our old numerical system, the one with $\sqrt{2}$?

The reason is purely psychological: whereas we can represent $\sqrt{2}$ as the length of a side of a right triangle, we don't have such an obvious geometric representation of $\sqrt{-1}$. But we can manipulate $\sqrt{-1}$ algebraically as effectively as $\sqrt{2}$.

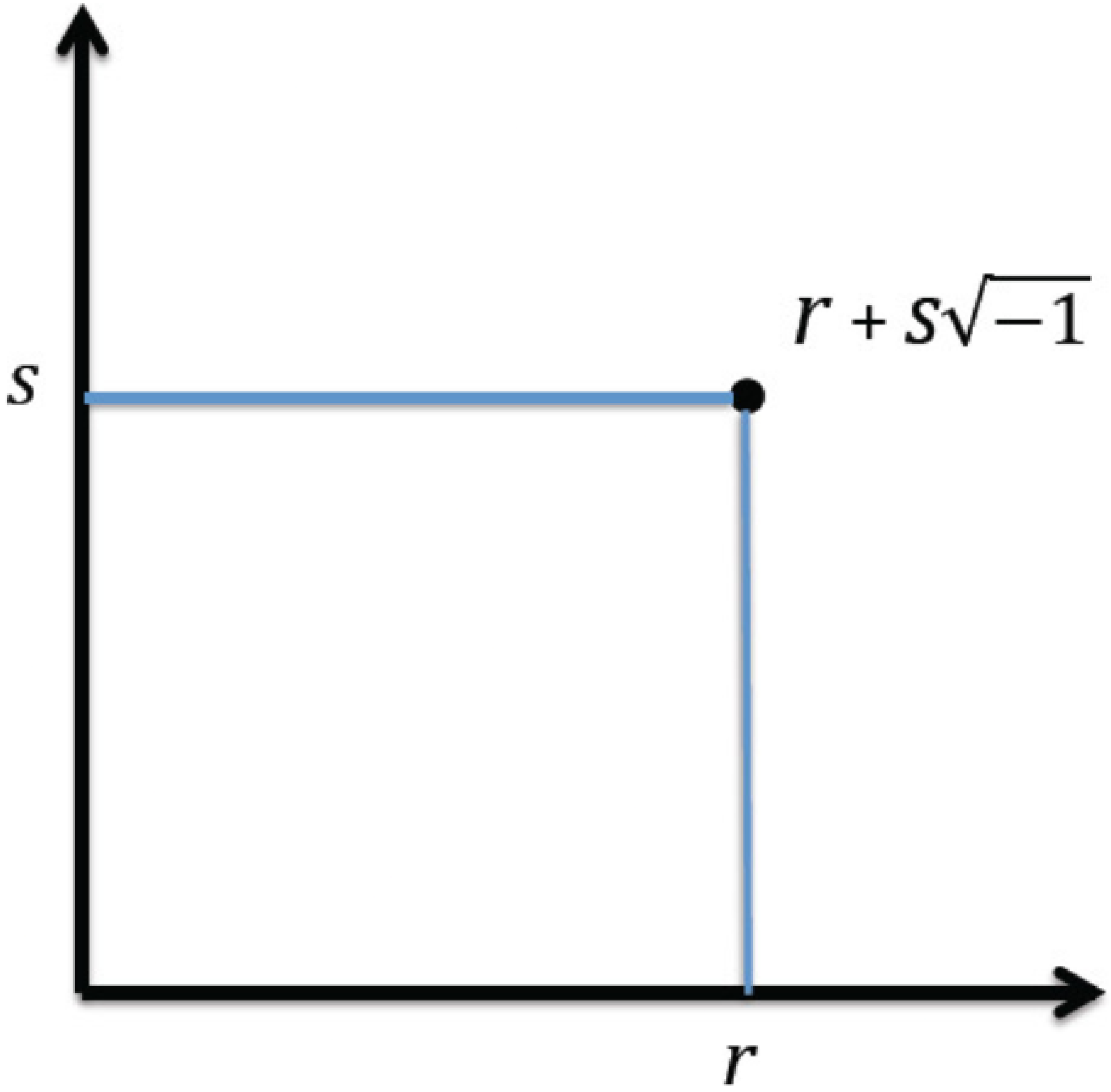
Elements of the new numerical system we obtain by adjoining $\sqrt{-1}$ to the rational numbers are called complex numbers. Each of them may be written as follows:

$$r + s\sqrt{-1},$$

where r and s are rational numbers. Compare this to the formula above expressing general elements of the numerical system obtained by adjoining $\sqrt{2}$. We can add any two numbers of this form by adding separately their r-parts and s-parts. We can also multiply any two such numbers by opening the brackets and using the fact that $\sqrt{-1}\cdot\sqrt{-1}=-1$. In a similar way, we can also subtract and divide these numbers.

Finally, we extend the definition of complex numbers by allowing r and s in the above formula to be arbitrary real numbers (not just the rational numbers). Then we obtain the most general complex numbers. Note that it is customary to denote $\sqrt{-1}$ by i (for "imaginary"), but I chose not to do this to emphasize the algebraic meaning of this number: it really is just a square root of −1, nothing more and nothing less. It is just as concrete as the square root of 2. There is nothing mysterious about it.

We can get a feel for how concrete these numbers are by representing them geometrically. Just as the real numbers may be represented geometrically as points of a line, complex numbers may be represented as points of a plane. Namely, we represent the complex number $r+s\sqrt{-1}$ as a point on the plane with coordinates r and s:[5]



Let's go back to our cubic equation

$$y^2 + y = x^3 - x^2$$

and let's look for solutions x and y that are complex numbers.

A remarkable fact is that the set of all such solutions turns out to be exactly the set of points of a torus depicted earlier. In other words, each point of the torus can be assigned to one and only one pair of complex numbers x, y solving the above cubic equation, and vice versa.[6]

If you've never thought about complex numbers before, your head might be starting to hurt just about now. This is completely natural. Wrapping one's mind around a single complex number is challenging enough, let alone pairs of complex numbers solving some equation. It's not obvious at all that these pairs are in one-to-one correspondence with points on the surface of a donut, so don't be alarmed if you don't see why this is so. In fact, many professional mathematicians would be hard-pressed to prove this surprising and non-trivial result.[7]

In order to convince ourselves that solutions of algebraic equations give rise to geometric shapes, let's look at a simpler situation: solutions over the real numbers instead of complex numbers. For example, consider the equation

$$x^2 + y^2 = 1$$

and let's mark its solutions as points on the plane with coordinates x and y. The set of all such solutions is a circle of radius one, centered at the origin. Likewise, solutions of any other algebraic equation in two *real*-valued variables x and y form a curve on this plane.[8]

Now, complex numbers are in some sense doubles of real numbers (indeed, each complex number is determined by a pair of real numbers), so it's not surprising that solutions of an algebraic equation in *complex*-valued variables x and y form a Riemann surface. (A curve is one-dimensional, and a Riemann surface is two-dimensional, in the sense explained in Chapter 10.)

In addition to real and complex solutions, we may also look for solutions x, y of these equations that take values in a finite field $\{0, 1, 2, \ldots, p-2, p-1\}$, where p is a prime number. This means that when we substitute x, y in the above cubic equation, say, the left- and right-hand sides become integers that are equal to each other up to an integer multiple of p. This gives us an object that mathematicians call a "curve over a finite field." Of course, these are not really curves. The terminology is due to the fact that when we look for solutions in real numbers, we obtain curves on the plane.[9]

A deep insight of Weil was that the most fundamental object here is an algebraic equation, like the cubic one above. Depending on the choice of the domain where we look for solutions, the same equation gives rise to a surface, a curve, or a bunch of points. But those are nothing but *avatars* of an ineffable being, which is the equation itself, the way Vishnu has ten avatars, or incarnations, in Hinduism. Somewhat serendipitously, in the letter to his sister, André Weil invoked the *Bhagavad-Gita*,[10] a sacred text of Hinduism, in which the doctrine of avatars of Vishnu is believed to appear for the first time.[11] Weil wrote poetically about what happens when the inkling of an analogy between two theories is turned into concrete knowledge:[12]

> Gone are the two theories, gone their troubles and delicious reflections in one another, their furtive caresses, their inexplicable quarrels; alas, we have but one theory, whose majestic beauty can no longer excite us. Nothing is more fertile than these illicit liaisons; nothing gives more pleasure to the connoisseur.... The pleasure comes from the illusion and the kindling of the senses; once the illusion disappears and knowledge is acquired,

> we attain indifference; in the *Gita* there are some lucid verses to that effect. But let's go back to algebraic functions.

The connection between Riemann surfaces and curves over finite fields should now be clear: both come from the same kind of equations, but we look for solutions in different domains, either finite fields or complex numbers. On the other hand, "any argument or result in number theory can be translated, word for word," to curves over finite fields, as Weil put it in his letter.[13] Weil's idea was therefore that curves over finite fields are the objects that intermediate between number theory and Riemann surfaces.

Thus, we find a bridge, or a "turntable" – as Weil called it – between number theory and Riemann surfaces, and that is the theory of algebraic curves over finite fields. In other words, we have three parallel tracks, or columns:

Number Theory *Curves over Finite Fields* *Riemann Surfaces*

Weil wanted to exploit this in the following way: take a statement in one of the three columns and translate it into statements in the other columns. He wrote to his sister:[14]

> My work consists in deciphering a trilingual text; of each of the three columns I have only disparate fragments; I have some ideas about each of the three languages: but I know as well there are great differences in meaning from one column to another, for which nothing has prepared me in advance. In the several years I have worked at it, I have found little pieces of the dictionary.

Weil went on to find one of the most spectacular applications of his Rosetta stone: what we now call the Weil conjectures. The proof of these conjectures[15] greatly stimulated the development of mathematics in the second half of the twentieth century.

Let's go back to the Langlands Program. Langlands' original ideas concerned the left column of Weil's Rosetta stone; that is, number theory. Langlands related representations of the Galois groups of number fields, which are objects studied in number theory, to automorphic functions, which are objects in harmonic analysis – an area of mathematics that is far removed from number theory (and also far away from other columns of the Rosetta stone). Now we can ask whether this kind of relation may also be found if we replace the Galois groups by some objects in the the middle and the right columns of Weil's Rosetta stone.

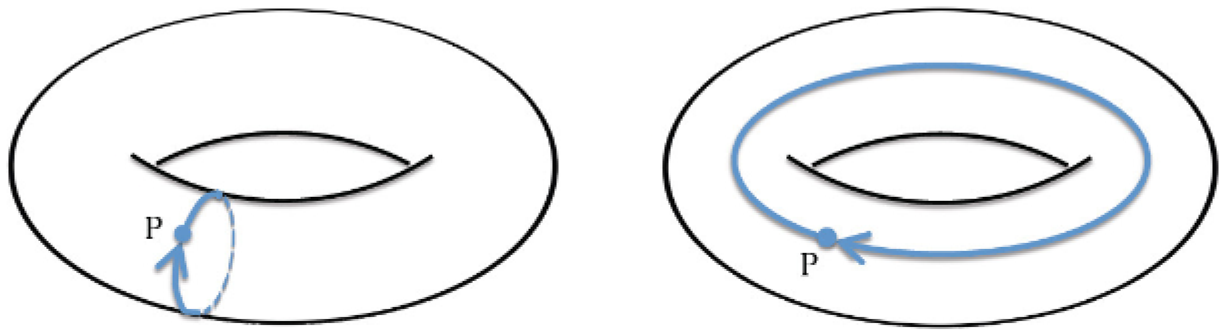
It is fairly straightforward to translate Langlands' relation to the middle column because all the necessary ingredients are readily available. Galois groups of number fields should be replaced here by the Galois groups relevant to curves over finite fields. There also exists a branch of harmonic analysis that studies suitable automorphic functions. Already in his original work, Langlands related representations of the Galois groups and automorphic functions relevant to the middle column.

However, it is not at all clear how to translate this relation to the right column of the Rosetta stone. In order to do this, we have to find geometric analogues of the Galois groups and automorphic functions in the theory of Riemann surfaces. When Langlands first formulated his ideas, the former was known, but the latter was a big mystery. It wasn't until the 1980s that the appropriate notion was found, starting with the pioneering work by a brilliant Russian mathematician Vladimir Drinfeld. This enabled the translation of the Langlands relation to the third column of the Rosetta stone.

Let's discuss first the geometric analogue of the Galois group. It is the so-called *fundamental group* of a Riemann surface.

The fundamental group is one of the most important concepts in the mathematical field of topology, which focuses on the most salient features of geometric shapes (such as the number of "holes" in a Riemann surface).

Consider, for example, a torus. We pick a point on it – call it P – and look at the closed paths starting and ending at this point. Two such paths are shown on the picture.



Likewise, the fundamental group of any given Riemann surface consists of such closed paths on this Riemann surface starting and ending at the same fixed point P.[16]

Given two paths starting and ending at the point P, we construct another path as follows: we move along the first path and then move along the second path. This way we obtain a new path, which will also start and end at the point P. It turns out that this "addition" of closed paths satisfies all properties of a group listed in Chapter 2. Thus, we find that these paths indeed form a group.[17]

You may have noticed that the rule of addition of paths in the fundamental group is similar to the rule of addition of braids in the braid groups, as defined in Chapter 5. This is not accidental. As explained in Chapter 5, braids with n threads may be viewed as paths on the space of collections of n distinct points on the plane. In fact, the braid group B_n is precisely the fundamental group of this space.[18]

It turns out that the two paths on the torus shown on the above picture commute with each other; that is, adding them in two possible orders gives us the same element of the fundamental group.[19] The most general element of the fundamental group of the torus is therefore obtained by following the first path M times and then following the second path N times, where M and N are two integers (if M is negative, then we follow the first path $-M$ times in the opposite direction, and similarly for negative N). Since the two basic paths commute with each other, the order in which we follow these paths does not matter; the result will be the same.

For other Riemann surfaces, the structure of the fundamental group is more complicated.[20] Different paths do not necessarily commute with each other. This is similar to braids with more than two threads not commuting with each other, as we discussed in Chapter 5.

It has been known for some time that there is a deep analogy between the Galois groups and the fundamental groups.[21] This provides the answer to our first question: what is the analogue of the Galois group in the right column of Weil's Rosetta stone? It is the fundamental group of the Riemann surface.

Our next question is to find suitable analogues of the automorphic functions, the objects that appear on the other side of the Langlands relation. And here we have to make a quantum leap. The good old functions turn out to be inadequate. They need to be replaced by more sophisticated objects of modern mathematics called *sheaves*, which will be described in Chapter 14.

This was proposed by Vladimir Drinfeld in the 1980s. He gave a new formulation of the Langlands Program that applies to the middle and the right columns, which concern curves over finite fields and Riemann surfaces, respectively. This formulation became known as the geometric Langlands Program. In particular, Drinfeld found the analogues of the automorphic functions suitable for the right column of Weil's Rosetta stone.

I met Drinfeld at Harvard University in the spring of 1990. Not only did he get me excited about the Langlands Program, he also told me that I had a role to play in its development. That's because Drinfeld saw a connection between the geometric Langlands Program and the work I did as a student in Moscow. The results of this work were essential in Drinfeld's new approach, and this in turn shaped my mathematical life: the Langlands Program has played a dominant role in my research ever since.

So let us return to Moscow, and see where I went after finishing my first paper, on braid groups.

*My editor tells me that the pretzels at the German bar near his house are genus–3 (and delicious).

Chapter 10

Being in the Loop

In Moscow in the fall of 1986, I was in the third year of my studies at Kerosinka. With the braid group paper finished and submitted, Fuchs had a question for me: "What do you want to do next?"

I wanted another problem to solve. It turned out that for several years Fuchs had been working with his former student Boris Feigin on representations of "Lie algebras." Fuchs said it was an active area with many unsolved problems and with close ties to quantum physics.

That sure caught my attention. Even though Evgeny Evgenievich had "converted" me to math, and even though I was enchanted by mathematics, I had never lost my childhood fascination with physics. That the worlds of math and quantum physics might come together was exciting for me.

Fuchs handed me an eighty-page research paper he and Feigin had written.

"I first thought of giving you a textbook on Lie algebras," he said. "But then I thought, why not just give you this paper?"

I put the paper carefully in my backpack. It was still unpublished at the time, and, thanks to the tight controls that Soviet authorities (afraid that people would make copies of banned literature, like books of Solzhenitsyn or *Doctor Zhivago*) placed on photocopiers, there were only a handful of copies available in the entire world. Very few people had ever gotten to see this paper – Feigin later joked that I may have been the only one who had read it from beginning to end.

It was written in English and was supposed to appear in a collection of papers published in the U.S. But the publisher badly mismanaged the book, and its publication was delayed for some fifteen years. By then, most of the results were reproduced elsewhere, so it wasn't much read after it came out either. Nevertheless, the article became famous, and Feigin and Fuchs eventually got their due credit. Their paper has been widely cited in the literature (as a "Moscow Preprint"), and even a new term was coined, "Feigin–Fuchs representations," to refer to the new representations of Lie algebras they studied in this paper.

As I started reading the paper, my first question was: what are these objects that carry such a strange name, "Lie algebras"? The paper that Fuchs gave me assumed quite a bit of knowledge about topics I'd never studied, so I went to a bookstore and bought all the textbooks on Lie algebras I could find. Whatever I could not find, I borrowed from the library at Kerosinka. I was reading all these books in parallel with the Feigin–Fuchs article. This experience shaped my learning style. Since then, I've never been satisfied with one source; I try to find all available sources and devour them.

To explain what Lie algebras are, I first need to tell you about "Lie groups." Both are named after a Norwegian mathematician Sophus Lie (pronounced LEE) who invented them.

Mathematical concepts populate the Kingdom of Mathematics, just like species of animals populating the Animal Kingdom: they are linked to each other, form families and subfamilies, and often two different concepts mate and produce an offspring.

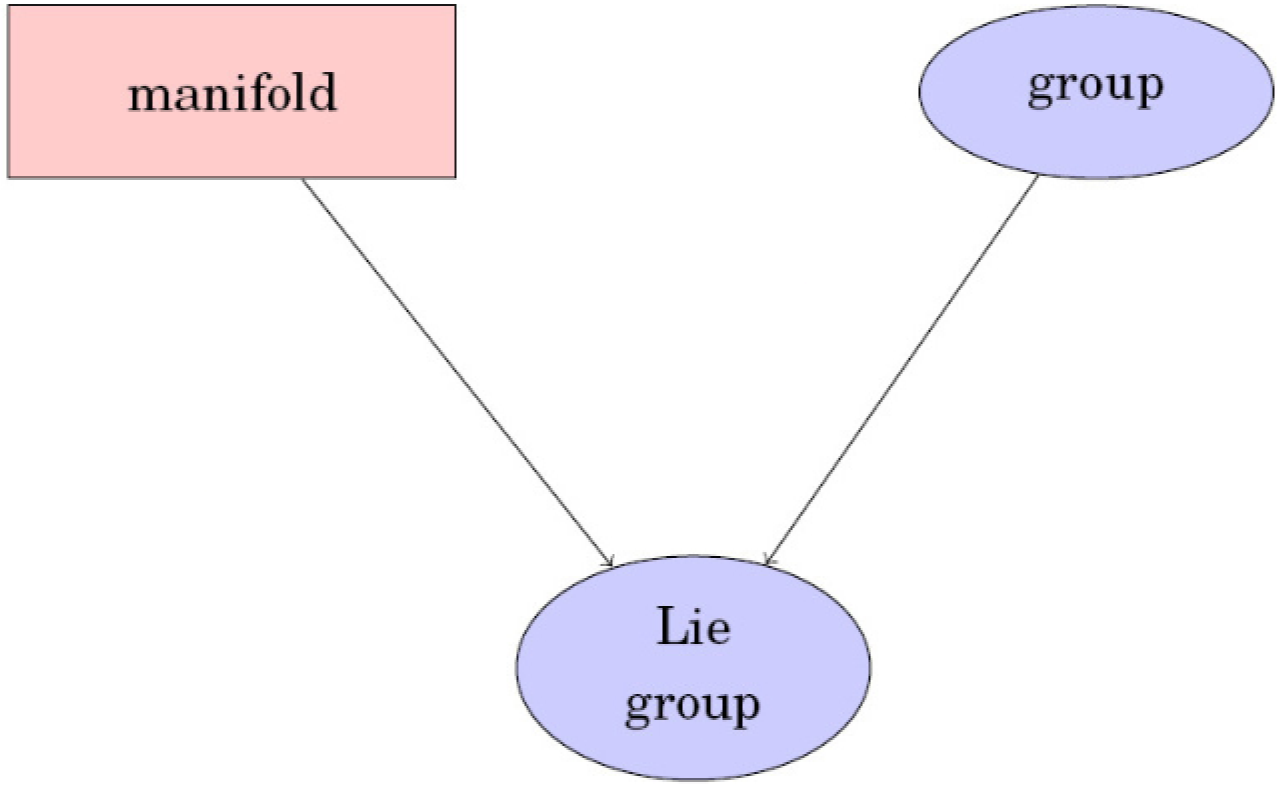
The concept of a group is a good example. Think of groups as analogues of birds, which form a class in the Animal Kingdom, or Animalia (called class Aves). That class is split into twenty-three orders; each order in turn splits into families, and each of those splits into genera. For example, the African fish eagle belongs to the order of Accipitriformes, the family of Accipitridae, and the genus of *Haliaeetus* (compared to these names, "Lie group" doesn't sound so exotic!). Likewise, groups form a large class of mathematical concepts, and within this class there are different "orders," "families," and "genera."

For example, there is an order of finite groups that includes all groups with finitely many elements. The group of symmetries of a square table, which we discussed in Chapter 2, consists of four elements, so it is a finite group. Likewise, the Galois group of a number field obtained by adjoining the solutions of a polynomial equation to the rational numbers is a finite group (for example, in the case of a quadratic equation it has two elements). The class of finite groups is further subdivided into families, such as the family of Galois groups. Another family consists of the crystallographic groups, which are the groups of symmetries of various crystals.

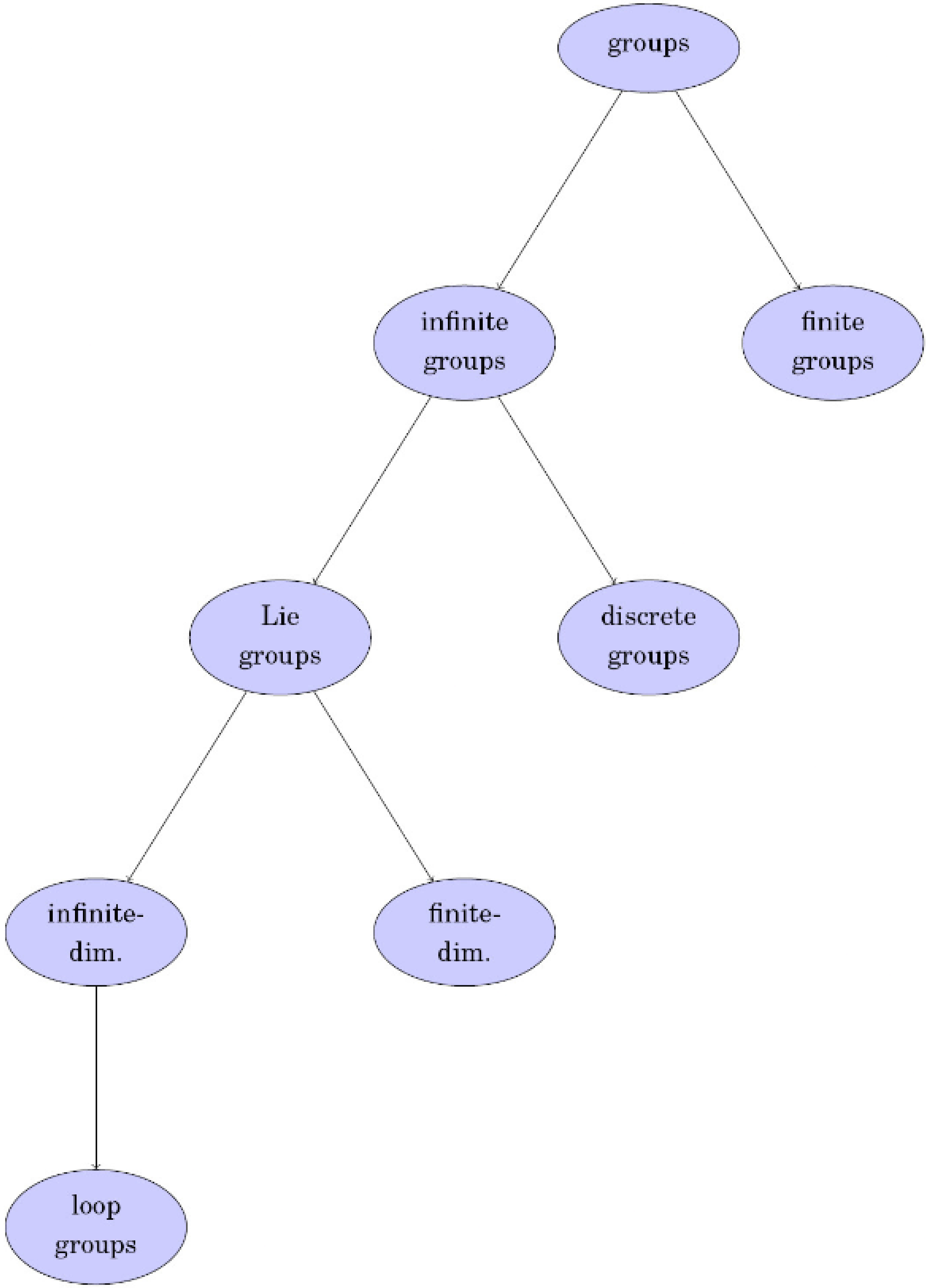
There is also another order, of infinite groups. For example, the group of integers is infinite, and so is the braid group B_n, which we discussed in Chapter 5, for each fixed $n = 2, 3, 4,\ldots$ (B_n consists of braids with n threads; there are infinitely many such braids). The group of rotations of a round table, which consists of all points on a circle, is also an infinite group.

But there is an important difference between the group of integers and the circle group. The group of integers is discrete; that is to say, its elements do not combine into a continuous geometric shape in any natural sense. We can't move continuously from one integer to the next; we jump from one to another. In contrast, we can change the angle of rotation continuously between 0 and 360 degrees. And together, these angles combine into a geometric shape: namely, the circle. Mathematicians call such shapes *manifolds*.

The group of integers and the braid groups belong to the family of discrete infinite groups in the Kingdom of Mathematics. And the circle group belongs to another family, that of Lie groups. Put simply, a Lie group is a group whose elements are points of a manifold. So this concept is the offspring resulting from the marriage of two mathematical concepts: group and manifold.

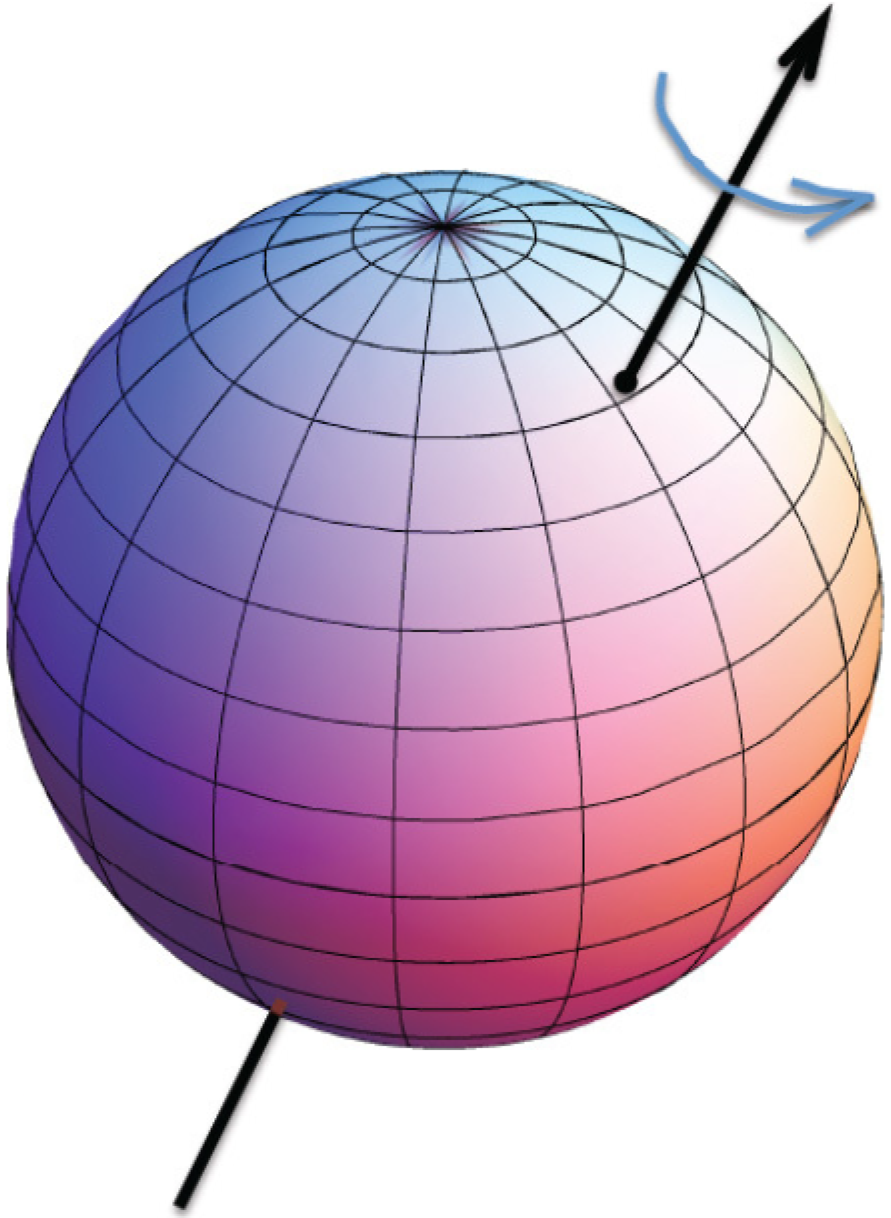


Here is the tree of group-related concepts that we will discuss in this chapter (some of these concepts have not yet been introduced, but will be later in the chapter).



Many symmetries arising in nature are described by Lie groups, and that's why they are so important to study. For example, the group $SU(3)$ we talked about in Chapter 2, which is used to classify elementary particles, is a Lie group.

Here is another example of a Lie group: the group of rotations of a sphere. A rotation of a round table is determined by its angle. But in the case of a sphere, there is more freedom: we have to specify the axis as well as the angle of rotation, as shown on the picture. The axis can be any line passing through the center of the sphere.



The group of rotations of the sphere has a name in math: the special orthogonal group of the 3-dimensional space, or, as it is commonly abbreviated, $SO(3)$. We can think of the symmetries of the sphere as transformations of the 3-dimensional space in which the sphere is embedded. These transformations are orthogonal, meaning that they preserve all distances.[1] Incidentally, this gives us a 3-dimensional representation of the group $SO(3)$, a concept we introduced in Chapter 2.

Likewise, the group of rotations of the round table, which we have discussed above, is called $SO(2)$; these rotations are special orthogonal transformations of the plane, which is 2-dimensional. Thus, we have a 2-dimensional representation of the group $SO(2)$.

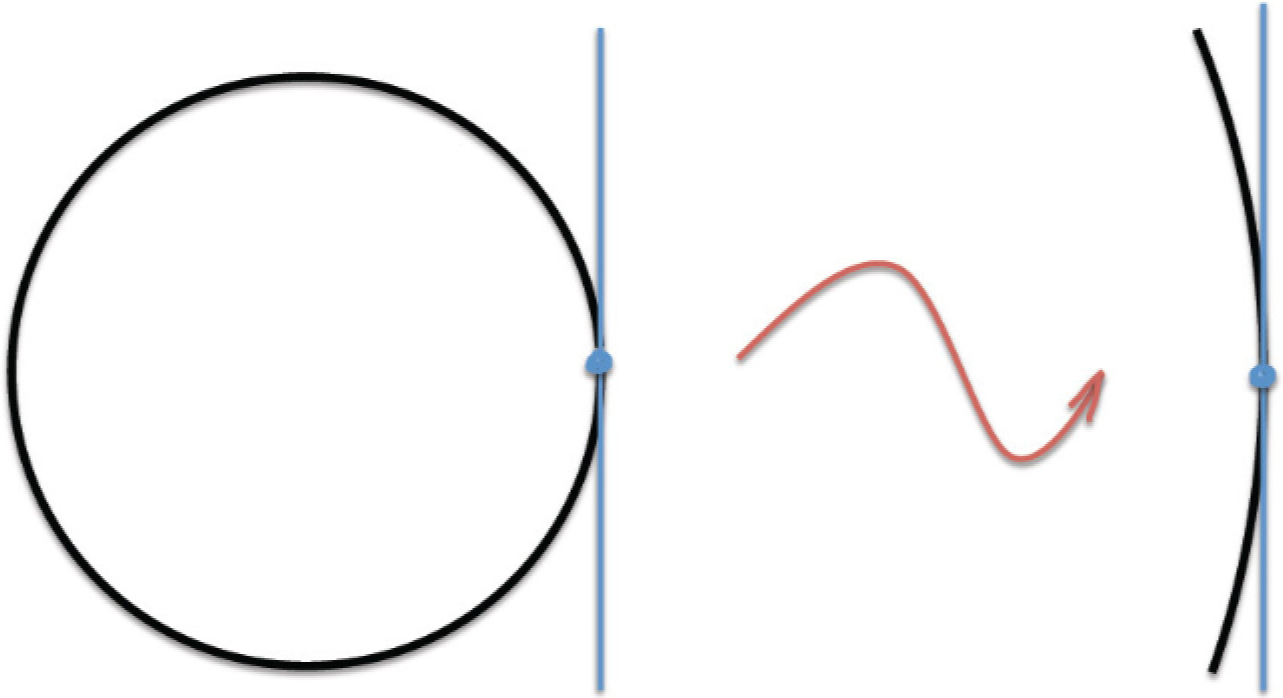
The groups $SO(2)$ and $SO(3)$ are not only groups but also manifolds (that is, geometric shapes). The group $SO(2)$ is the circle, which is a manifold. So $SO(2)$ is a group and a manifold. That's why we say that it is a Lie group. Likewise, elements of the group $SO(3)$ are points of another manifold, but it is more tricky to visualize it. (Note that this manifold is *not* a sphere.) Recall that each rotation of the sphere is determined by the axis and the angle of rotation. Now observe that each point of the sphere gives rise to an axis of rotation: the line connecting this point and the center of the sphere. And the angle of rotation is the same as a point of a circle. So an element of the group $SO(3)$ is determined by a point of the sphere (it defines the axis of rotation) together with a point of a circle (it defines the angle of rotation).

Perhaps we should start with a simpler question: what is the dimension of $SO(3)$? To answer this question, we need to discuss the meaning of dimension more systematically. We have already mentioned in Chapter 2 that the world around us is three-dimensional. That is to say, in order to specify a position of a point in space, we need to specify three numbers, or coordinates, (x, y, z). A plane, on the other hand, is two-dimensional: a position on the plane is specified by two coordinates, (x, y). And a line is one-dimensional: there is only one coordinate, x.

But what is the dimension of a circle? It is tempting to say that the circle is two-dimensional because we usually draw the circle on a plane, which is two-dimensional. Each point of the circle, when viewed as a point of the plane, is described by two coordinates. But the mathematical definition of the dimension of a given geometric object (such as a circle) is the number of independent coordinates we need *on this object* to pinpoint any location on it. This number has nothing to do with the dimension of the landscape into which the object is embedded (such as a plane). Indeed, a circle can also be embedded into a three-dimensional space (think of a ring on one's finger), or a space of even larger dimension. What matters is that for a particular circle, the position of any point on it can be described by one number, namely, the angle. This is the sole coordinate on the circle. That's why we say that the circle is one-dimensional.

Of course, in order to speak of the angle, we need to pick a reference point on the circle corresponding to angle 0. Likewise, in order to assign a coordinate x to each point on the line we need to pick a reference point on it corresponding to $x = 0$. We can set up a coordinate system on a given object in many different ways. But each of these coordinate systems will have the same *number* of coordinates, and it's this number that is called the *dimension* of this object.

Note that as we zoom in and look at a smaller and smaller neighborhood of a point of the circle, the curvature of the circle all but disappears. There is practically no difference between a small neighborhood of a point on the circle and a small neighborhood of the same point on the tangent line to the circle; this is the line that is the closest approximation to the circle near this point. This shows that the circle and the line have the same dimension.[2]



As we zoom in on a point, the circle and the tangent line appear closer and closer to each other.

Likewise, the sphere is embedded into a three-dimensional space, but its intrinsic dimension is two. Indeed, there are two independent coordinates on the sphere: latitude and longitude. We know them well because we use them to determine the position on the surface of the Earth, which is close to the shape of a sphere. The mesh on the sphere that we see on the above picture is made of the "parallels" and "meridians," which correspond to fixed values of latitude and longitude. The fact that there are two coordinates on the sphere tells us that it is two-dimensional.

What about the Lie group $SO(3)$? Every point of $SO(3)$ is a rotation of the sphere, so we have three coordinates: the axis of rotation (which may be specified by a point at which the axis pierces the sphere) is described by two coordinates, and the angle of rotation gives rise to the third coordinate. Hence the dimension of the group $SO(3)$ is equal to three.

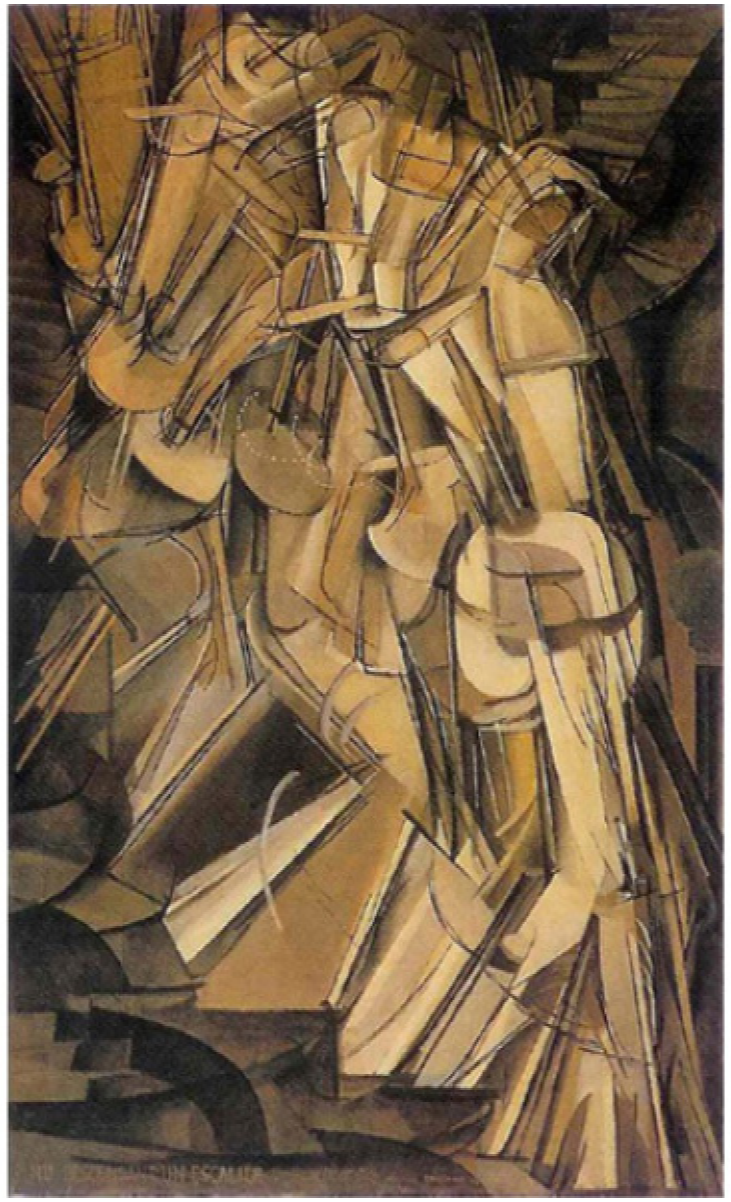
Thinking about a Lie group, or any manifold, of more than three dimensions can be very challenging. Our brain is wired in such a way that we can only imagine geometric shapes, or manifolds, in dimensions up to three. Even imagining the four-dimensional combination of space and time is a strenuous task: we just don't perceive the time (which constitutes the fourth dimension) as an equivalent of a spatial dimension. What about higher dimensions? How can we analyze five- or six- or hundred-dimensional manifolds?

Think about this in terms of the following analogy: works of art give us two-dimensional renderings of three-dimensional objects. Artists paint two-dimensional projections of those objects on the canvas and use the technique of perspective to create the illusion of depth, the third dimension, in their paintings. Likewise, we can imagine four-dimensional objects by analyzing their three-dimensional projections.

Another, more efficient way to imagine a fourth dimension is to think of a four-dimensional object as a collection of its three-dimensional "slices." This would be similar to slicing a loaf of bread, which is three-dimensional, into slices so thin that we could think of them as being two-dimensional.

If the fourth dimension represents time, then this four-dimensional "slicing" is known as photography. Indeed, snapping a picture of a moving person gives us a three-dimensional slice of a four-dimensional object representing that person in the four-dimensional space-time (this slice is then projected onto a plane). Taking several pictures in succession, we obtain a collection of such slices. If we run these pictures quickly in front of our eyes, we can see that movement. This is of course the basic idea of cinema.

We can also convey the impression of the person’s movement by juxtaposing the pictures. At the beginning of the twentieth century, artists got interested in this idea and used it as a way to include the fourth dimension into their paintings, to render them dynamic. A milestone in this direction was Marcel Duchamp’s 1912 painting *Nude Descending a Staircase, No. 2*.



It is interesting to note that Einstein’s relativity theory, which demonstrated that space and time cannot be separated from each other, appeared around the same time. This brought the notion of the four-dimensional space-time continuum to the forefront of physics. In parallel, mathematicians such as Henri Poincaré were delving deeper into the mysteries of higher-dimensional geometry and transcending the Euclidean paradigm.

Duchamp was fascinated with the idea of the fourth dimension as well as non-Euclidean geometry. Reading E.P. Jouffret’s book *Elementary Treatise on Four-Dimensional Geometry and Introduction to the Geometry of n Dimensions*, which in particular presented the groundbreaking ideas of Poincaré, Duchamp left the following note:[3]

> The shadow cast by a 4-dimensional figure on our space is a 3-dimensional shadow (see Jouffret – Geom. of 4-dim., page 186, last 3 lines).... by analogy with the method by which architects depict a plan of each story of a house, a 4-dimensional figure can be represented (in each one of its stories) by three-dimensional sections. These different stories will be bound to one another by the 4th dim.

According to art historian Linda Dalrymple Henderson,[4] “Duchamp found something deliciously subversive about the new geometries with their challenge to so many long-standing ‘truths.’ ” The interest of Duchamp and other artists of that era in the fourth dimension, she writes, was one of the elements that led to the birth of abstract art.

Thus, mathematics informed art; it allowed artists to see hidden dimensions and inspired them to expose, in a tantalizing aesthetic form, some profound truths about our world. The works of modern art they created helped elevate our perception of reality, affecting our collective consciousness. This in turn influenced the next generations of mathematicians. Philosopher of science Gerald Holton put this eloquently:[5]

> Indeed, a culture is kept alive by the interaction of all its parts. Its progress is an alchemical process, in which all its varied ingredients can combine to form new jewels. On this point, I imagine that Poincaré and Duchamp are in agreement with me and with each other, both having by now undoubtedly met somewhere in that hyperspace which, in their different ways, they loved so well.

Mathematics enables us to perceive geometry in all of its incarnations, shapes, and forms. It is a universal language that applies equally well in all dimensions, whether we can visualize the corresponding objects or not, and allows us to go far beyond our limited visual imagination. In fact, Charles Darwin wrote that mathematics endows us with "an extra sense."[6]

For example, though we cannot imagine a four-dimensional space, we can actualize it mathematically. We simply represent points of this space as quadruples of numbers (x, y, z, t), just like we represent points of the three-dimensional space by triples of numbers (x, y, z). In the same way, we can view points of an n-dimensional flat space, for any natural number n, as n-tuples of numbers (we can analyze these in the same way as the rows of a spreadsheet, as we discussed in Chapter 2).

Perhaps I need to explain why I refer to these spaces as being flat. A line is clearly flat and so is a plane. But it's not as obvious that we should think of the three-dimensional space as flat. (Note that I am not talking here about various curved manifolds embedded into the three-dimensional space, such as a sphere or a torus. I am talking about the three-dimensional space itself.) The reason is that it has no curvature. The precise mathematical definition of curvature is subtle (it was given by Bernhard Riemann, the creator of Riemann surfaces), and we won't go into the details now as this is tangential to our immediate goals. A good way to think about the flatness of the three-dimensional space is to realize that it has three infinite coordinate axes that are perpendicular to each other, just as a plane has two perpendicular coordinate axes. Likewise, an n-dimensional space, with n perpendicular coordinate axes, has no curvature and hence is flat.

Physicists have thought for centuries that we inhabit a flat three-dimensional space, but, as we discussed in the Preface, Einstein has shown in his general relativity theory that gravity causes space to curve (the curvature is small, so that we don't notice it in our everyday life, but it is non-zero). Therefore our space is in fact an example of a curved three-dimensional manifold.

This brings up the question of how a curved space could possibly exist by itself, without being embedded into a flat space of higher dimension, the way a sphere is embedded into a flat three-dimensional space. We are used to thinking that the space we live in is flat, and so in our everyday experience curved shapes seem to appear only within the confines of that flat space. But this is a misunderstanding, an artifact of our narrow perception of reality. And the irony is that the space we live in isn't flat to begin with! Mathematics gives us a way out of this trap: as Riemann showed, curved spaces do exist intrinsically, as objects of their own making, without a flat space containing them. What we need to define such a space is a rule of measuring distances between any two points of this space (this rule must satisfy certain natural properties); this is what mathematicians call a *metric*. The mathematical concepts of metric and the curvature tensor, introduced by Riemann, are the cornerstones of Einstein's general relativity theory.[7]

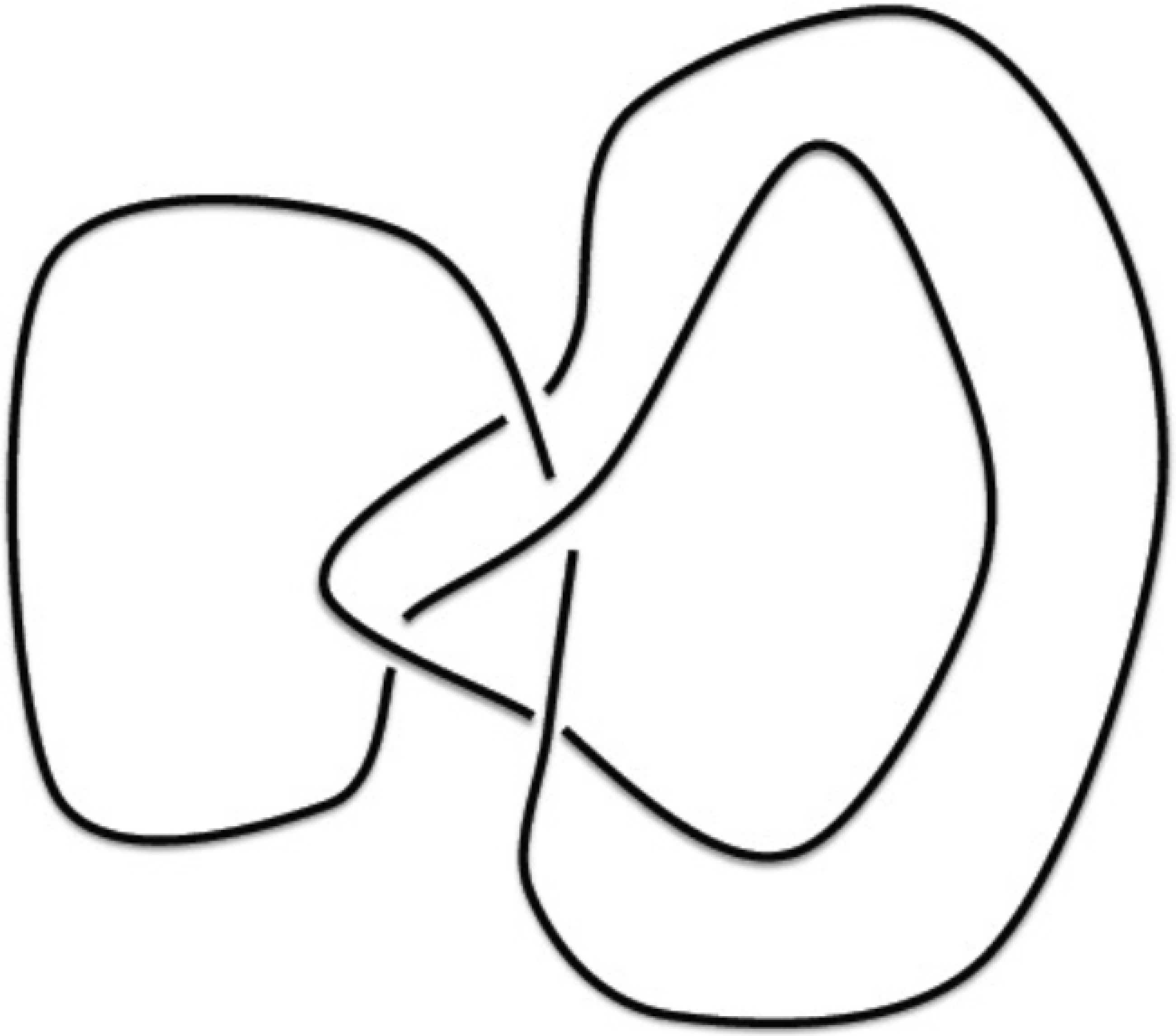
Curved shapes, or manifolds, can have arbitrarily high dimensions. Recall that the circle is defined as the set of points on a plane equidistant from a given point (or, as my examiner at MGU insisted, the set of *all* such points!). Likewise, a sphere is the set of all points in the three-dimensional space equidistant from a given point. Now, define a higher-dimensional analogue of a sphere – some call it a hypersphere – as the set of points equidistant from a given point in the n-dimensional space. This condition gives us one constraint on the n coordinates. Therefore the dimension of the hypersphere inside the n-dimensional space is $(n - 1)$. Further, we can study the Lie group of rotations of this hypersphere.[8] It is denoted by $SO(n)$.

From the point of view of the taxonomy of groups in the Kingdom of Mathematics, the family of Lie groups is subdivided into two genera: that of finite-dimensional Lie groups (such as the circle group and the group $SO(3)$) and that of infinite-dimensional Lie groups. Note that any finite-dimensional Lie group is already infinite, in the sense that it has infinitely many elements. For example, the circle group has infinitely many elements (these are the points of the circle). But it is one-dimensional because all of its elements may be described by one coordinate (the angle). For an infinite-dimensional Lie group, we need infinitely many coordinates in order to describe its elements. This kind of "double infinity" is really hard to imagine. Yet, such groups do arise in nature, so we need to study them as well. I will now describe an example of an infinite-dimensional Lie group known as a *loop group*.

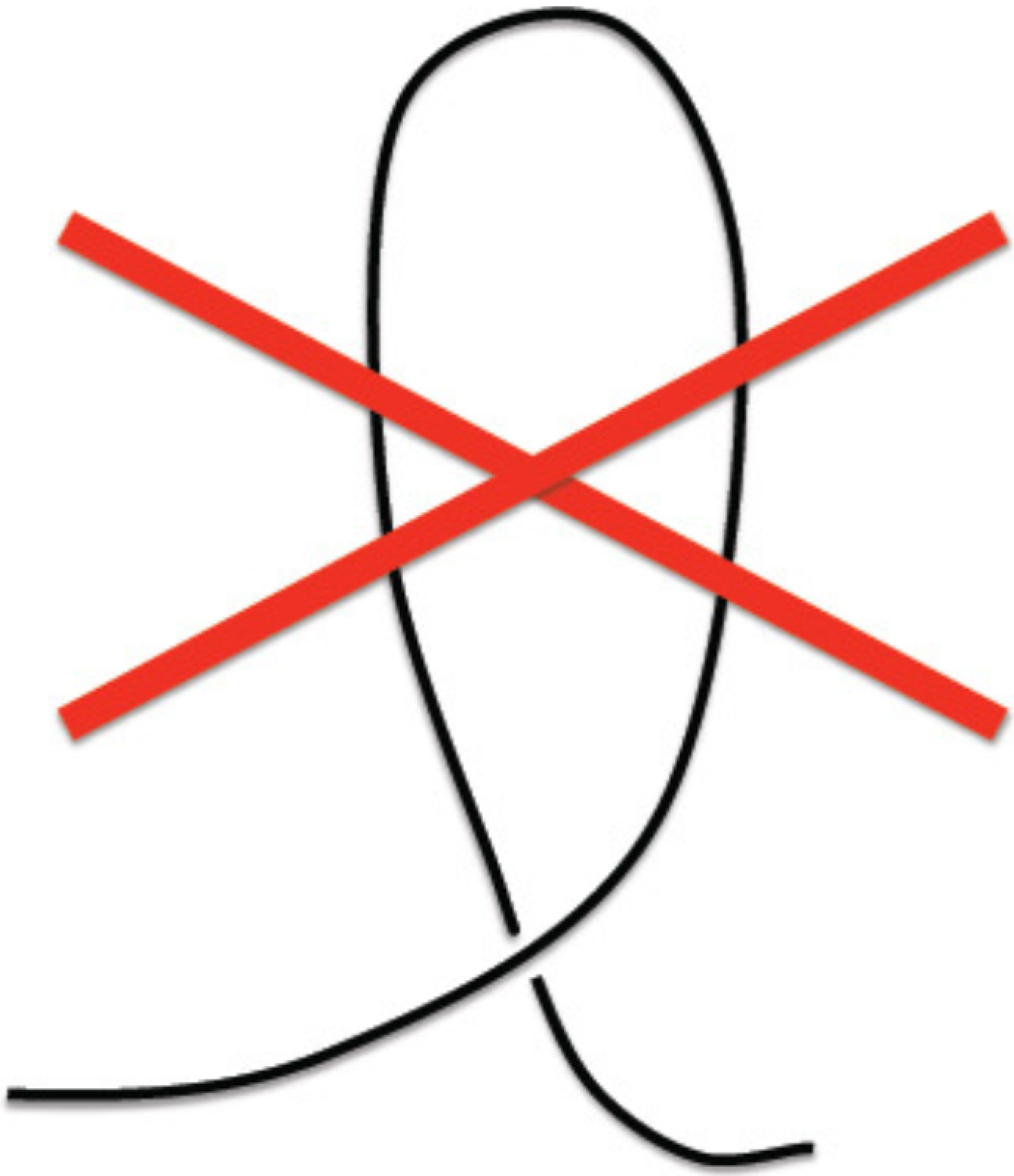
To explain what it is, let's first consider loops in the three-dimensional space. Simply put, a loop is a closed curve, such as the one shown on the left-hand side of the picture below. We have already seen them when we talked about braid groups (we called them "knots").[9] I want to stress that non-closed curves, such as the one shown on the right-hand side of the picture, are *not* considered as loops.

Similarly, we can also consider loops (that is, closed curves) inside any manifold M. The space of all these loops is called the loop space of M.

As we will discuss in more detail in Chapter 17, these loops play a big role in string theory. In conventional quantum physics, the fundamental objects are elementary particles, such as electrons or quarks. They are point-like objects, with no internal structure; that is, zero-dimensional. In string theory it is postulated that fundamental objects of nature are one-dimensional strings.[10] A closed string is nothing but a loop embedded in a manifold M (the space-time). That's why loop spaces are the bread and butter of string theory.



This is a loop



This is *not* a loop

Now let's consider the loop space of the Lie group $SO(3)$. Its elements are loops in $SO(3)$. Let's look at one of these loops closely. First of all, it is similar to the loop pictured above. Indeed, $SO(3)$ is three-dimensional, so on a small scale it looks like the three-dimensional flat space. Second, each point on this loop is an element of $SO(3)$, that is, a rotation of the sphere. Hence our loop is a sophisticated object: it is a one-parameter collection of rotations of the sphere. Given two such loops, we can produce a third by composing the corresponding rotations of the sphere. Thus, the loop space of $SO(3)$ becomes a group. We call it the loop group of $SO(3)$.[11] It's a good example of an infinite-dimensional Lie group: we really cannot describe its elements by using a finite number of coordinates.[12]

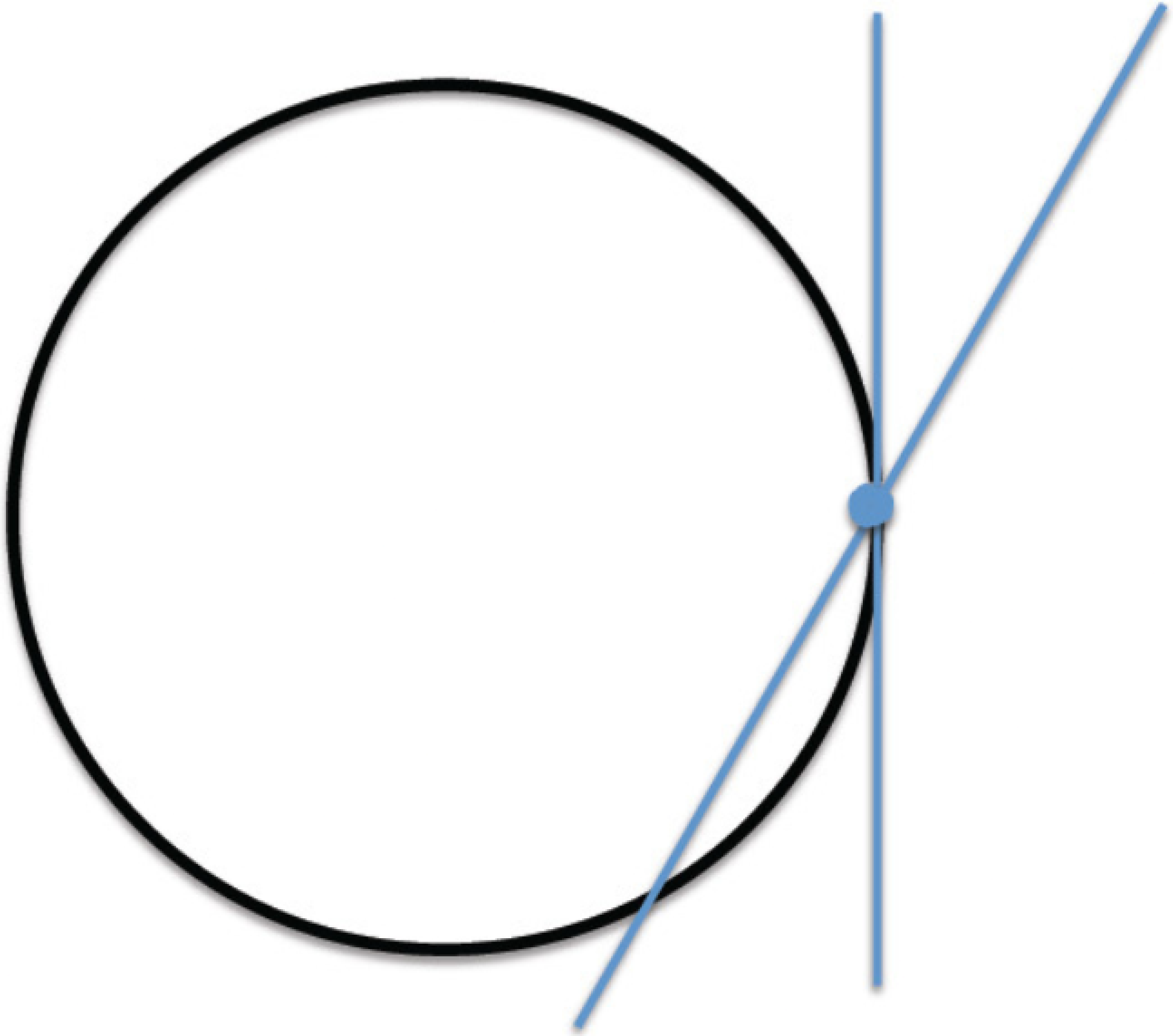
The loop group of any other Lie group (for example, the group $SO(n)$ of rotations of a hypersphere) is also an infinite-dimensional Lie group. These loop groups arise as symmetry groups in string theory.

The second concept relevant to the paper by Feigin and Fuchs that I was studying was the concept of a Lie algebra. Each Lie algebra is in some sense a simplified version of a given Lie group.

The term "Lie algebra" is bound to create some confusion. When we hear the word "algebra," we think of the stuff we studied in high school, such as solving quadratic equations. However, now the word "algebra" is used in a different connotation: as part of the indivisible term "Lie algebra" referring to mathematical objects with specific properties. Despite what the name suggests, these objects do not form a family in the class of all algebras, the way Lie groups form a family in the class of all groups. But nevermind, we'll just have to live with this inconsistency of terminology.

To explain what a Lie algebra is, I first have to tell you about the concept of the *tangent space*. Don't worry, we are not going off on a tangent; we follow one of the key ideas of calculus called “linearization,” that is, approximation of curved shapes by linear, or flat, ones.

For example, the tangent space to a circle at a given point is the line that passes through this point and is the line closest to the circle among all lines passing through this point. We have already encountered it above when we talked about the dimension of the circle. The tangent line just touches the circle at this particular point, ever so slightly, whereas all other lines passing through this point cross the circle at another point as well, as shown on the picture.



Likewise, any curve (that is, a one-dimensional manifold) can be approximated near a given point by a tangent line. René Descartes, who described an efficient method for computing these tangent lines in his *Géométrie*, published in 1637, wrote:[13] “I dare say that this is not only the most useful and most general problem in geometry that I know, but even that I have ever desired to know.” Similarly, a sphere can be approximated at a given point by a tangent plane. Think of a basketball: when we put it on the floor, it touches the floor at one point, and the floor becomes its tangent plane at that point.[14] And an n-dimensional manifold may be approximated at a given point by a flat n-dimensional space.

Now, on any Lie group we have a special point, which is the identity element of this group. We take the tangent space to the Lie group at this point – and *voilà*, that’s the Lie algebra of this Lie group. So each Lie group has its own Lie algebra, which is like a younger sister of the Lie group.[15]

For example, the circle group is a Lie group, and the identity element of this group is a particular point on this circle[16] corresponding to the angle 0. The tangent line at this point is therefore the Lie algebra of the circle group. Alas, we cannot draw a picture of the group $SO(3)$ and its tangent space because they are both three-dimensional. But the mathematical theory describing tangent spaces is set up in such a way that it works equally well in all dimensions. If we want to imagine how things work, we can model them on one- or two-dimensional examples (like a circle or a sphere). In doing so, we use lower-dimensional manifolds as metaphors for more complicated, higher-dimensional manifolds. But we don’t have to do this; the language of mathematics enables us to transcend our limited visual intuition. Mathematically, the Lie algebra of an n-dimensional Lie group is an n-dimensional flat space, also known as a vector space.[17]

There is more. The operation of multiplication on a Lie group gives rise to an operation on its Lie algebra: given any two elements of the Lie algebra, we can construct a third. The properties of this operation are more difficult to describe than the properties of multiplication in a Lie group, and they are not essential to us at the moment.[18] An example, which would be familiar to those readers who have studied vector calculus, is the operation of *cross-product* in the three-dimensional space.[19] If you have seen this operation, you may have wondered about its weird-looking properties. And guess what, this operation actually makes the three-dimensional space into a Lie algebra!

It turns out that this is in fact the Lie algebra of the Lie group *SO*(3). So the esoteric-looking operation of cross-product is inherited from the rule of composition of rotations of the sphere.

You may be wondering why we care about Lie algebras if the operation on them is so weird-looking. Why not stick with Lie groups? The main reason is that, unlike a Lie group, which is usually curved (like a circle), a Lie algebra is a flat space (like a line, a plane, and so on). This makes the study of Lie algebras much simpler than the study of Lie groups.

For example, we can talk about the Lie algebras of loop groups.[20] These Lie algebras, which we should think of as simplified versions of the loop groups, are called *Kac–Moody algebras*, after two mathematicians: Victor Kac (Russian-born, emigrated to the U.S., now Professor at MIT) and Robert Moody (British-born, emigrated to Canada, now Professor at University of Alberta). They started investigating these Lie algebras independently in 1968. Since then, the theory of Kac–Moody algebras has been one of the hottest and fastest growing areas of mathematics.[21]

It was these Kac–Moody algebras that Fuchs had suggested as the topic of my next research project. When I started learning all this, I saw that I had to do a lot of studying before I could get to the point where I could do something on my own. But I was fascinated with the subject.

Fuchs lived in the northeastern part of Moscow, not far from a train station where I could catch a train to my hometown. I used to go home every Friday for the weekend, so Fuchs suggested that I come to his place every Friday at 5 pm and then take the train home after our meeting. I would usually work with him for about three hours (during which he would also feed me dinner), and then I would catch the last train, arriving home around midnight. Those meetings played a big role in my mathematical education. We had them week after week, the entire fall semester of 1986, and then the spring semester of 1987 as well.

It wasn't until January 1987 that I finished reading the long paper by Feigin and Fuchs and felt that I could start working on my research project. By that time, I was able to get a pass to the Moscow Science Library, a huge repository of books and journals, not only in Russian (many of which Kerosinka's library also had), but in other languages as well. I started going there regularly to pore over dozens of math journals, looking for articles about Kac–Moody algebras and related subjects.

I was also eager to learn about their applications to quantum physics, which was of course a huge draw for me. As I mentioned above, Kac–Moody algebras play an important role in string theory, but they also appear as symmetries of models of two-dimensional quantum physics. We live in a three-dimensional space, so realistic models describing our world should be three-dimensional. If we include time, we get four dimensions. But mathematically, nothing precludes us from building and analyzing models describing worlds of other dimensions. The models in dimensions less than three are simpler, and we have a better chance of solving them. We can then use what we learn to tackle the more sophisticated three- and four-dimensional models.

This is in fact one of the main ideas of the subject called "mathematical physics" – study models of different dimensions that may not be directly applicable to our physical world, but share some of the salient features of the realistic models.

Some of these low-dimensional models also have real-world applications. For example, a very thin metal layer may be viewed as a two-dimensional system and hence may be effectively described by a two-dimensional model. A famous example is the so-called Ising model of interacting particles occupying the nodes of a two-dimensional lattice. The exact solution of the Ising model by Lars Onsager provided valuable insights into the phenomenon of spontaneous magnetization, or ferromagnetism. At the core of Onsager's calculation was a hidden symmetry of this model, underscoring once again symmetry's paramount role in understanding physical systems. It was subsequently understood that this symmetry is described by the so-called Virasoro algebra, a close cousin of Kac–Moody algebras.[22] (In fact, it was the Virasoro algebra that was the main subject of the paper by Feigin and Fuchs which I was studying.) There is also a large class of models of this type in which symmetries are described by the Kac–Moody algebras proper. The mathematical theory of Kac–Moody algebras is essential for understanding these models.[23]

Kerosinka's library subscribed to a publication called *Referativny Zhurnal*, the Journal of References. This journal, published monthly, had short reviews of all new articles, in all languages, organized by subject, with a short summary of each. I started reading it regularly, and what a valuable source it turned out to be! Every month a new volume about math papers would come, and I would fish through the relevant sections trying to find something of interest. If I found something that sounded exciting, I would write down the reference and get it on my next visit to the Moscow Science Library. This way, I discovered a lot of interesting stuff.

One day, while turning the pages of the *Referativny Zhurnal*, I stumbled upon a review of a paper by a Japanese mathematician Minoru Wakimoto, which was published in one of the journals I was paying close attention to, *Communications in Mathematical Physics*. The review did not say much, but the title referred to the Kac–Moody algebra associated to the group of rotations of the sphere, $SO(3)$, so I took down the reference and on my next visit to the Science Library I read the article.

In it, the author constructed novel realizations of the Kac–Moody algebra associated to $SO(3)$. To give the gist of what they are, I will use the language of quantum physics (which is relevant here because Kac–Moody algebras describe symmetries of models of quantum physics). Realistic quantum models, like those describing the interaction of elementary particles, are quite complicated. But we can construct much simpler, idealized, "free-field models," in which there is no or almost no interaction. The quantum fields in these models are "free" from each other, hence the name.[24] It is often possible to realize a complicated, and hence more interesting, quantum model inside one of these free-field models. This allows us to dissect and deconstruct the complicated models, and perform computations that are not accessible otherwise. Such realizations are very useful as the result. However, for quantum models with Kac–Moody algebras as symmetries, the known examples of such free-field realizations had been rather narrow in scope.

As I was reading Wakimoto's paper, I saw right away that the result could be interpreted as giving the broadest possible free-field realization in the case of the simplest Kac–Moody algebra, the one associated to *SO*(3). I understood the importance of this result, and it made me wonder: where did this realization come from? Is there a way to generalize it to other Kac–Moody algebras? I felt that I was ready to tackle these questions.

How to describe the excitement I felt when I saw this beautiful work and realized its potential? I guess it's like when, after a long journey, suddenly a mountain peak comes in full view. You catch your breath, take in its majestic beauty, and all you can say is "Wow!" It's the moment of revelation. You have not yet reached the summit, you don't even know yet what obstacles lie ahead, but its allure is irresistible, and you already imagine yourself at the top. It's yours to conquer now. But do you have the strength and stamina to do it?