CHAPTER 1

Reinventing Discovery

Tim Gowers is not your typical blogger. A mathematician at Cambridge University, Gowers is a recipient of the highest honor in mathematics, the Fields Medal, often called the Nobel Prize of mathematics. His blog radiates mathematical ideas and insight.

In January 2009, Gowers decided to use his blog to run a very unusual social experiment. He picked out an important and difficult unsolved mathematical problem, a problem he said he'd "love to solve." But instead of attacking the problem on his own, or with a few close colleagues, he decided to attack the problem completely in the open, using his blog to post ideas and partial progress. What's more, he issued an open invitation asking other people to help out. Anyone could follow along and, if they had an idea, explain it in the comments section of the blog. Gowers hoped that many minds would be more powerful than one, that they would stimulate each other with different expertise and perspectives, and collectively make easy work of his hard mathematical problem. He dubbed the experiment the Polymath Project.

The Polymath Project got off to a slow start. Seven hours after Gowers opened up his blog for mathematical discussion, not a single person had commented. Then a mathematician named Jozsef Solymosi from the University of British Columbia posted a comment suggesting a variation on Gowers's problem, a variation which was easier, but which Solymosi thought might throw light on the original problem. Fifteen minutes later, an Arizona high-school teacher named Jason Dyer chimed in with a thought of his own. And just three minutes after that, UCLA mathematician Terence Tao—like Gowers, a Fields medalist—added a comment. The comments

erupted: over the next 37 days, 27 people wrote 800 mathematical comments, containing more than 170,000 words. Reading through the comments you see ideas proposed, refined, and discarded, all with incredible speed. You see top mathematicians making mistakes, going down wrong paths, getting their hands dirty following up the most mundane of details, relentlessly pursuing a solution. And through all the false starts and wrong turns, you see a gradual dawning of insight. Gowers described the Polymath process as being "to normal research as driving is to pushing a car." Just 37 days after the project began Gowers announced that he was confident the polymaths had solved not just his original problem, but a harder problem that included the original as a special case. He described it as "one of the most exciting six weeks of my mathematical life." Months' more cleanup work remained to be done, but the core mathematical problem had been solved. (If you'd like to know the details of Gowers's problem, they're described in the appendix. If you just want to get on with reading this book, you can safely skip those details.)

The polymaths aren't standing still. Since Gowers's original project, nearly a dozen Polymath and Polymath-like projects have been launched, some attacking problems even more ambitious than Gowers's original problem. More than 100 mathematicians and other scientists have participated; mass collaboration is starting to spread through mathematics. Like the first Polymath Project, several of these projects have been great successes, really driving our understanding of mathematics forward. Others have been more qualified successes, falling short of achieving their (sometimes extremely ambitious) goals. Regardless, massively collaborative mathematics is a powerful new way of attacking hard mathematical problems.

Why is mass online collaboration useful in solving mathematical problems? Part of the answer is that even the best mathematicians can learn a great deal from people with complementary knowledge, and be stimulated to consider ideas in directions they wouldn't have considered on their own. Online tools create a shared space where this can happen, a short-term collective working memory where ideas can be rapidly improved by many minds. These tools enable us to scale up creative conversation, so connections that would ordinarily require fortuitous serendipity instead happen as a

matter of course. This speeds up the problem-solving process, and expands the range of problems that can be solved by the human mind.

The Polymath Project is a small part of a much bigger story, a story about how online tools are transforming the way scientists make discoveries. These tools are *cognitive tools*, actively amplifying our collective intelligence, making us smarter and so better able to solve the toughest scientific problems. To understand why all this matters, think back to the seventeenth century and the early days of modern science, the time of great discoveries such as Galileo's observation of the moons of Jupiter, and Newton's formulation of his laws of gravitation. The greatest legacy of Galileo, Newton, and their contemporaries wasn't those one-off breakthroughs. It was the method of scientific discovery itself, a way of understanding how nature works. At the beginning of the seventeenth century extraordinary genius was required to make even the tiniest of scientific advances. By developing the method of scientific discovery, early scientists ensured that by the end of the seventeenth century such scientific advances were run-of-the-mill, the likely outcome of any competent scientific investigation. What previously required genius became routine, and science exploded.

Such improvements to the way discoveries are made are more important than any single discovery. They extend the reach of the human mind into new realms of nature. Today, online tools offer us a fresh opportunity to improve the way discoveries are made, an opportunity on a scale not seen since the early days of modern science. I believe that the process of science—how discoveries are made—will change more in the next twenty years than it has in the past 300 years.

The Polymath Project illustrates just a single aspect of this change, a shift in how scientists work together to create knowledge. A second aspect of this change is a dramatic expansion in scientists' ability to find meaning in knowledge. Consider, for example, the studies you often see reported in the news saying "so-and-so genes cause such-and-such a disease." What makes these studies possible is a genetic map of human beings that's been assembled over the past twenty years. The best-known part of that map is the human genome, which scientists completed in 2003. Less well known, but

perhaps even more important, is the HapMap (short for haplotype map), completed in 2007, which charts how and where different human beings can *differ* in their genetic code. Those genetic variations determine much about our different susceptibilities to disease, and the HapMap says where those variations can occur—it's a genetic map not just of a single human being, but of the entire human race.

This human genetic map was the combined work of many, many biologists around the world. Each time they obtained a new chunk of genetic data in their laboratories, they uploaded that data to centralized online services such as GenBank, the amazing online repository of genetic information run by the US National Center for Biotechnology Information. GenBank integrates all this genetic information into a single, publicly accessible online database, a compilation of the work of thousands of biologists. It's information on a scale that's almost impossible to analyze by hand. Fortunately, anyone in the world may freely download the genetic map, and then use computer algorithms to analyze the map, perhaps discovering previously unsuspected facts about the human genome. You can, if you like, go to the GenBank site right now, and start browsing genetic information. (For links to GenBank and other resources, see the "Notes on Sources," starting on page 347.) This is, in fact, what makes those studies linking genes to disease possible: the scientists doing the studies start by finding a large group of people with the disease, and also a control group of people without the disease. They then use the human genetic map to find correlations between disease incidence and the genetic differences of the two groups.

A similar pattern of discovery is being used across science. Scientists in many fields are collaborating online to create enormous databases that map out the structure of the universe, the world's climate, the world's oceans, human languages, and even all the species of life. By integrating the work of hundreds or thousands of scientists, we are collectively mapping out the entire world. With these integrated maps anyone can use computer algorithms to discover connections that were never before suspected. Later in the book we'll see examples ranging from new ways of tracking influenza outbreaks to the discovery of orbiting pairs of supermassive black holes. We are, piece by piece, assembling all the world's knowledge into a single

giant edifice. That edifice is too vast to be comprehended by any individual working alone. But new computerized tools can help us find meaning hidden in all that knowledge.

If the Polymath Project illustrates a shift in how scientists collaborate to create knowledge, and GenBank and the genetic studies illustrate a shift in how scientists find meaning in knowledge, a third big shift is a change in the relationship between science and society. An example of this shift is the website Galaxy Zoo, which has recruited more than 200,000 online volunteers to help astronomers classify galaxy images. Those volunteers are shown photographs of galaxies, and asked to answer questions such as "Is this a spiral or an elliptical galaxy?" and "If this is a spiral, do the arms rotate clockwise or anticlockwise?" These are photographs that have been taken automatically by a robotic telescope, and have never before been seen by any human eye. You can think of Galaxy Zoo as a cosmological census, the largest ever undertaken, a census that has so far produced more than 150 million galaxy classifications.

The volunteer astronomers who participate in Galaxy Zoo are making astonishing discoveries. They have, for example, recently discovered an entirely new class of galaxy, the "green pea galaxies"—so named because the galaxies do, indeed, look like small green peas—where stars are forming faster than almost anywhere else in the universe. They've also discovered what is believed to be the first ever example of a quasar mirror, an enormous cloud of gas tens of thousands of light-years in diameter, which is glowing brightly as the gas is heated by light from a nearby quasar. In just three years, the work of the Galaxy Zoo volunteers has resulted in 22 scientific papers, and many more are in the works.

Galaxy Zoo is just one of many online citizen science projects that are recruiting volunteers, most of them without scientific training, to help solve scientific research problems. We'll see examples ranging across science, from volunteers who are using computer games to predict the shape of protein molecules, to volunteers who are helping understand how dinosaurs evolved. These are serious scientific projects, projects where large groups of volunteers with little scientific training can attack scientific problems beyond the reach of small groups of professionals. There's no way a team of professionals could do what Galaxy Zoo does—even working full

time, the pros don't have the time to classify hundreds of thousands (or more) of galaxies. You might suppose they'd use computers to classify the galaxy images, but in fact the human volunteers classify the galaxies more accurately than even the best computer programs. So the volunteers at projects such as Galaxy Zoo are expanding the boundary of what scientific problems can be solved, and in so doing, changing both who can be a scientist and what it means to be a scientist. How far can the boundary between professional and amateur scientist be blurred? Will we one day see Nobel Prizes won by huge collaborations dominated by amateurs?

Citizen science is part of a larger shift in the relationship between science and society. Galaxy Zoo and similar projects are examples of institutions that are bridging the scientific community and the rest of society in new ways. We'll see that online tools enable many other new bridging institutions, including open access publishing, which gives the public direct access to the results of science, and science blogging, which is helping create a more open and more transparent scientific community. What other new ways can we find to build bridges between science and the rest of society? And what will be the long-run impact of these new bridging institutions?

The story so far is an optimistic story of possibility, of new tools that are changing the world. But there's a problem with this story, some major obstacles that prevent scientists from taking full advantage of online tools. To understand the obstacles, consider the studies linking genes to disease that we discussed earlier. There's a crucial part of that story which I glossed over, but which is actually quite puzzling: *why* is it that biologists share genetic data in GenBank in the first place? When you think about it, it's a peculiar choice: if you're a professional biologist it's to your advantage to keep data secret as long as possible. Why share your data online before you get a chance to publish a paper or take out a patent on your work? In the scientific world it's papers and, in some fields, patents that are rewarded by jobs and promotions. Publicly releasing data typically does nothing for your career, and might even damage it, by helping your scientific competitors.

In part for these reasons, GenBank took off slowly after it was launched in 1982. While many biologists were happy to access others' data in GenBank, they had little interest in contributing

their own data. But that has changed over time. Part of the reason for the change was a historic conference held in Bermuda in 1996, and attended by many of the world's leading biologists, including several of the leaders of the government-sponsored Human Genome Project. Also present was Craig Venter, who would later lead a private effort to sequence the human genome. Although many attendees weren't willing to unilaterally make the first move to share all their genetic data in advance of publication, everyone could see that science as a whole would benefit enormously if open sharing of data became common practice. So they sat and talked the issue over for days, eventually coming to a joint agreement—now known as the Bermuda Agreement—that all human genetic data should be immediately shared online. The agreement wasn't just empty rhetoric. The biologists in the room had enough clout that they convinced several major scientific grant agencies to make immediate data sharing a mandatory requirement of working on the human genome. Scientists who refused to share data would get no grant money to do research. This changed the game, and immediate sharing of human genetic data became the norm. The Bermuda agreement eventually made its way to the highest levels of government: on March 14, 2000, US President Bill Clinton and UK Prime Minister Tony Blair issued a joint statement praising the principles described in the Bermuda Agreement, and urging scientists in every country to adopt similar principles. It's because of the Bermuda Agreement and similar subsequent agreements that the human genome and the HapMap are publicly available.

This is a happy story, but it has an unhappy coda. The Bermuda Agreement originally only applied to human genetic data. There have since been many attempts to extend the spirit of the agreement, so that more genetic data is shared. But despite these attempts, there are still many forms of life for which genetic data remains secret. For example, as of 2010 there is no worldwide agreement to share data about the influenza virus. Steps toward such an agreement remain bogged down in wrangling among the leading parties. To give you the flavor of how many scientists think about sharing non-human genetic data, one scientist recently told me that he'd been "sitting on a genome" for an entire species (!) for more than a year. Without any incentive to share, and with many reasons

not to, scientists hoard their data. As a result, there's an emerging data divide between our understanding of life-forms such as human beings, where nearly all genetic data are available online, and life-forms such as influenza, where important data remain locked up.

This story makes it sound as though the scientists involved are greedy and destructive. After all, this research is typically paid for using public funds. Shouldn't scientists make their results available as soon as possible? There's truth to these ideas, but the situation is complex. To understand what's going on, you need to understand the incredible competitive pressures on ambitious young scientists. On the rare occasion a good long-term job at a major university opens up, there are often hundreds of superbly-qualified applicants. Competition for jobs is so fierce that eighty-hour-plus workweeks are common among young scientists. As much of that time as possible is spent working on the one thing that will get such a job: amassing an impressive record of scientific papers. Those papers will bring in the research grants and letters of recommendation necessary to find long-term employment. The pace relaxes after tenure, but continued grant support still requires a strong work ethic. The result is that while many scientists agree in principle that they'd love to share their data in advance of publication, they worry that doing so will give their competitors an unfair advantage. Those competitors could exploit that knowledge to rush their results into print first, or, worse, even steal the data outright and present the results as their own. It's only practical to share data if everyone is protected by a collective agreement such as the Bermuda agreement.

A similar pattern has seen scientists resist contributing to many other online projects. Consider Wikipedia, the online encyclopedia. Wikipedia has a vision statement to warm a scientist's heart: *"Imagine a world in which every single human being can freely share in the sum of all knowledge. That's our commitment."* You might think Wikipedia was started by scientists eager to share all the world's knowledge, but you'd be wrong. In fact, it was started by Jimmy "Jimbo" Wales, who at the time was cofounder of an online company mostly specializing in adult content, and Larry Sanger, a philosopher who left academia to work with Wales on online encyclopedias. In the early days of Wikipedia there was little involvement from

scientists. This was despite the fact that anyone in the world can edit Wikipedia, and, in fact, it's written entirely by its users. So here's this incredibly exciting project, which anyone can get involved in, which is taking off rapidly, and which expresses core scientific values. Why weren't scientists lining up to be involved? The problem is the same as with the genetic data: why would scientists take the time to contribute to Wikipedia when they could be doing something more respectable among their peers, like writing a paper? That's the kind of activity that leads to jobs, grants, and promotions. It doesn't matter that contributing to Wikipedia might be more intrinsically valuable. In the early days work on Wikipedia was seen by scientists as frivolous, a waste of time, as not being serious science. I'm happy to say that this has changed over the years, and today Wikipedia's success has to some extent legitimized work on it by scientists. But isn't it strange that the modern-day Library of Alexandria came from outside academia?

There's a puzzle here. Scientists helped create the internet and the world wide web. They've taken enthusiastically to online tools such as email, and pioneered striking projects such as the Polymath Project and Galaxy Zoo. Why is it that they've only reluctantly adopted tools such as GenBank and Wikipedia? The reason is that, despite their radical appearance, the Polymath Project, Galaxy Zoo, and similar undertakings have an inherent underlying conservatism: they're ultimately projects in service of the conventional goal of writing scientific papers. That conservatism helps them attract contributors who are willing to use unconventional means such as blogs to more effectively achieve a conventional end (writing a scientific paper). But when the goal isn't simply to produce a scientific paper —as with GenBank, Wikipedia, and many other tools—there's no direct motivation for scientists to contribute. And that's a problem, because some of the best ideas for improving the way scientists work involve a break away from the scientific paper as the ultimate goal of scientific research. There are opportunities being missed that dwarf GenBank and Wikipedia in their potential impact. In this book, we'll delve into the history and culture of science, and see how this situation arose, in which scientists are often reluctant to share their ideas and data in ways that speed up the advancement of science. The good news is that we'll find leverage

points where small changes today will lead to a future where scientists do take full advantage of online tools, greatly increasing our capacity for scientific discovery.

Revolutions are sometimes marked by a single, spectacular event: the storming of the Bastille during the French Revolution, or the signing of the US Declaration of Independence. But often the most important revolutions aren't announced with the blare of trumpets. They occur quietly, too slowly to make the news, but fast enough that if you aren't alert, the revolution is over before you're aware it's happening. The change described in this book is like this. It's not a single event, nor is it a change that's happening quickly. It's a slow revolution that has quietly been gathering steam for years. Indeed, it's a change that many scientists have missed or underestimated, being so focused on their own specialty that they don't appreciate just how broad-ranging the impact of the new online tools is. They're like surfers at the beach who are so intent on watching the waves crash and recede that they're missing the rise of the tide. But you shouldn't let the slow, quiet nature of the current changes in how science is done fool you. We are in the midst of a great change in how knowledge is constructed. Imagine you were alive in the seventeenth century, at the dawn of modern science. Most people alive at that time had no idea of the great transformation that was going on, a transformation in how we know. Even if you were not a scientist, wouldn't you have wanted to at least be aware of the remarkable transformation that was going on in how we understood the world? A change of similar magnitude is going on today: we are reinventing discovery.

I wrote this book because I believe the reinvention of discovery is one of the great changes of our time. To historians looking back a hundred years from now, there will be two eras of science: pre-network science, and networked science. We are living in the time of transition to the second era of science. But it's going to be a bumpy transition, and there is a possibility it will fail or fall short of its potential. And so I also wrote the book to help create a widely shared public understanding of the opportunity now before us, an understanding that a more open approach to science isn't just a nice idea, but that it must be demanded of our scientists and our scientific institutions.

This change is important. Improving the way science is done means speeding up the rate of all scientific discovery. It means speeding up things such as curing cancer, solving the climate-change problem, launching humanity permanently into space. It means fundamental insights into the human condition, into how the universe works and what it is made of. It means discoveries we've not yet dreamt of. Over the next few years we have an astonishing opportunity to change and improve the way science is done. This book is the story of this change, what it means for us, and what we need to do to make it happen.