
1001001

The Promise of BIG DATA

Petabytes of raw information could provide clues for everything from preventing TB to shrinking health care costs—if we can figure out how to use them.

HSPH microbiologist Sarah Fortune went to Camden, Maine in late 2010 to attend a small but widely revered conference on innovation called PopTech. Fortune had for more than a decade been trying to crack one of the tuberculosis bacterium's most infuriating characteristics: its rising resistance to antibiotic drugs.

Standing on the Camden Opera House stage, backlit by mammoth close-ups of fluorescent cells, Fortune shared with her fellow PopTech attendees TB's grim annual statistics: 2 billion people—nearly one-third of the world's population—are latent carriers. Every year, 15 million become sick and 1.4 million die.

Unlike most bacteria, TB cells do not replicate as carbon copies but in random patterns, she told the audience. TB cells behave more like snowflakes than Xeroxes. Fortune believes it is this variety that gives TB its extraordinary ability to defy conventional antibiotics.

Using silicon chips and a special camera, Fortune, the Melvin J. and Geraldine L. Glimcher Assistant Professor of Immunology and Infectious Diseases, and her fellow researchers had developed a way to capture 10,000 still images of this telltale growth every few days—exponentially more data than they had only a few years ago.

continued

The images are combined like old-fashioned flip books into what Fortune calls "movies." But only the human eye can assess the moving pictures, one by one—a method so laborious that it inhibits scientific progress.

The question troubling Fortune, and what had brought her to the conference, was the following: How could her lab swiftly analyze this unprecedented treasure trove? The new data could be a gold mine—one that could yield fundamental insights about potential diagnostic tools, treatments, even a vaccine—but not without ways to speed up analysis. Fortune needed help.

HSPH's Sarah Fortune crowdsourced an image processing project. In two days, volunteers with no scientific expertise measured cell growth in a 5,300-image "movie" of dividing TB cells. Without the volunteers' collective eyes, the task would have taken three months.

THE DILEMMA OF BIG DATA

What was happening in Sarah Fortune's lab is playing out in laboratories, businesses, and government agencies everywhere. Our ability to generate data has moved light-years ahead of where it was only a few years ago, and the amount of digital information now available to us is essentially unimaginable.

"In the last five years, more scientific data has been generated than in the entire history of mankind," says Winston Hide, associate professor of bioinformatics at HSPH. "You can imagine what's going to happen in the next five." And this data isn't simply linear; genetics and proteomics, to name just two fields of study, generate high-dimensional data, which is fundamentally different in scale.

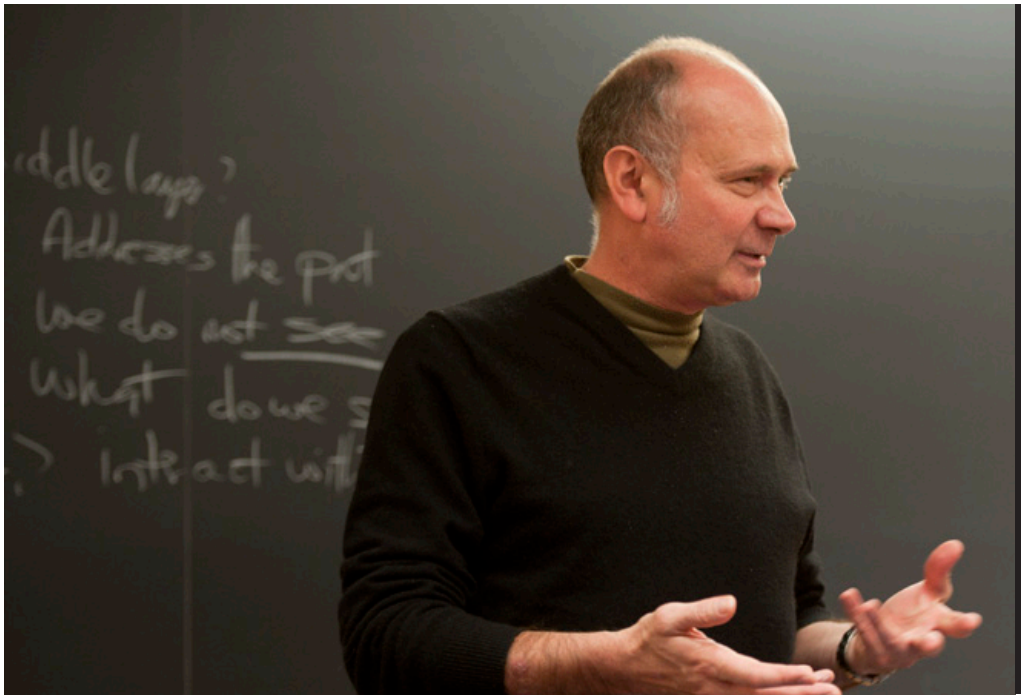
1,000,000,000,000,00

"Imagine a city made out of stacks of paper, each stack printed with sets of data," says Hide. He flings his arms in the air, drawing megaspace. "Imagine a whole planet that size. Imagine a million planets! Imagine a galaxy full

Sarah Fortune, Melvin J. and Geraldine L. Glimcher Assistant Professor of Immunology and Infectious Diseases

The crowdsourcing process Fortune used generated new, fundamental findings about TB cells that may yield clues to drug treatments.

Opposite and right, ©Tony Rinaldo



In just two weeks, Winston Hide, associate professor of biostatistics, joined a cancer database with a stem cell dataset—and got a big payoff. "We discovered a single gene that we think is responsible for the initiation of a whole class of leukemias."

00,000,000,000,000

a zettabyte equals 10^{21} bytes or one trillion gigabytes

of those, and we haven't even got there yet! That's high-dimensional data."

REVOLUTIONARY APPLICATIONS

In big data lies the potential for revolutionizing, well, everything. Police employing seismology-like data models can predict where crimes will occur and prevent them from happening. Astronomers using the Kepler telescope snag information on 200,000 stars every 30 seconds, which has led to the discovery of the first Earth-like planets outside our solar system. Businesses sifting social networking and supply-chain data dynamically tailor their products to fulfill desires we don't even know we have.

The same phenomena are at play in public health. For some time, DNA sequencing has held big data's starring role—after all, a single human genome consists of some 3 billion base pairs of DNA. Researchers at HSPH and across the campus at Harvard are sequencing and analyzing human genomes to ferret out clues to infections, cancer, and noncommunicable diseases.

But the potential public health uses of big data extend well beyond genomics. Environmental scientists are capturing huge quantities of air quality data from polluted areas and attempting to match it with equally bulky health care datasets for insights into respiratory disease. Epidemiologists are gathering information on social and sexual networks to better pinpoint the spread of disease and even create early warning systems. Comparative-

"In the last five years, more scientific data has been generated than in the entire history of mankind," says Winston Hide, associate professor of bioinformatics.

effectiveness researchers are combing government and clinical databases for proof of the best, most cost-effective treatments for hundreds of conditions—information that could transform health care policy. And disease researchers now have access to human genetic data and genomic databases of millions of bacteria—data they can combine to study treatment outcomes.

continued

1 PETABYTE
=1 quadrillion bytes

enough to store approximately:

2.8 million copies of the full text of the *Encyclopedia Britannica*;

or

1,903 years of music recorded at standard quality for an Apple iPod;

or

as much data as a stack of DVDs, each containing a two-hour standard definition video, roughly 1.8 times as high as the Empire State Building

1 TERABYTE
= 1 trillion bytes

enough to store approximately:

2,767 copies of the full text of the *Encyclopedia Britannica*;

or

16,667 hours of music recorded at standard quality for an Apple iPod;

or

1,333 hours of standard definition video

1 GIGABYTE
= 1 billion bytes

enough to store approximately:

212 copies of *War and Peace* or almost three copies of the full text (all 32 volumes) of the *Encyclopedia Britannica*;

or

250 songs recorded at Apple iTunes standard quality;

or

80 minutes of standard definition video

1 MEGABYTE
= 1 million bytes

enough to store a 500-page book in plain text

1 KILOBYTE
= 1 thousand bytes

enough to store a short paragraph's worth of plain text

1 BYTE

enough to store one letter of the alphabet

Sources: *Apple Computer, Amazon.com, New York Times, Perma-bound.com, the Official Website of the Empire State Building*

Younger scientists—raised in an era of social networking—may embrace an idea that previous generations of researchers have not: sharing data freely.

According to McKinsey & Company, with the right tools, big data could be worth $9 billion to U.S. public health surveillance alone and $300 billion to American health care in general, the former by improving detection of and response to infectious disease outbreaks, and the latter largely through reductions in expenditures.

A CRITICAL BOTTLENECK

It's hardly a given, though, that we'll get to this nirvana any time soon. Our ability to generate data far outstrips our ability to analyze it. "If we really start trying to exploit all these databases, we will need more trained staff and more resources to do it," says Victor De Gruttola, who chairs HSPH's biostatistics department.

Most researchers agree that lives are lost every day that data sit in storage, untouched. The problems are vast and urgent. Consider just one example—recent news that a dozen Indian patients had contracted totally drug-resistant tuberculosis. "Even just a few people in Mumbai is a terrible danger sign," says Fortune, because it could portend the rapid spread of a highly transmissible and untreatable infection.

To counter these trends, some scientists are venturing into crowdsourcing. Others are developing sophisticated algorithms to parse data in a keystroke. And still more are inventing ways to share massive, disparate datasets to yield surprising insights.

WISDOM OF THE CROWD

At PopTech, frustrated with the slow pace of her research, Sarah Fortune took a risk that most scientists wouldn't. She asked the audience for advice on how to analyze her images. "We would like to engage lots of eyes in that process," she said.

When Fortune walked off the stage, Josh Nesbit, a young entrepreneur in the audience, resolved to meet her. Nesbit had launched a company, Medic Mobile, that had built an emergency response system after the Haiti earthquake, calling on 2,500 Creole speakers to translate text messages. When the system was overwhelmed by victims texting for help, Nesbit turned to a Silicon Valley crowdsourcing company called CrowdFlower, which has signed up more than 2 million people to perform micro-tasks, often for pennies a task. The volunteers used CrowdFlower's website to translate, map, and organize nearly 100,000 messages, imploring rescuers for food, water, and help escaping from fallen buildings.

The evening after Fortune's talk, at a glitzy reception, Nesbit shared his story. Fortune instantly saw the possibilities: She could crowdsource the image processing of her growing TB cultures. In May 2011, CrowdFlower put one of Fortune's laboratory "movies" online. Some 1,000 interested people, with no scientific expertise, signed on to help. They measured and labeled the distance between cells as one cell split into two and two split into four, shooting off in patterns too random for computer programs to track. In two days, they'd measured cell growth in a 5,300-image movie. Without their collective eyes, it would have taken three months.

More important, their analysis generated new, fundamental findings about TB cells, which are shaped like cough drops. "We discovered that mycobacterial cell growth is not even," Fortune says. "One end of the cell is different from the other end, and in fact, it only grows from one end." She calls the nongrowing ends "privileged"—that is, not terribly vulnerable to antibiotics. That crowd-enabled insight, she says, may yield clues to pathogenesis and drug treatment.

Pardis Sabeti, assistant professor in the Department of Immunology and Infectious Diseases at HSPH and computational biologist at the Broad Institute.

©Tony Rinaldo

FINDING ALL THE NEEDLES IN A HAYSTACK

Around the time that Fortune was wondering how to quickly analyze thousands of images, David Reshef was pondering an even larger problem: He wanted to parse millions of relationships buried in big data. An MD/PhD candidate at the Harvard-MIT Division of Health Sciences and Technology, Reshef and his brother, Yakir, spent their childhoods in Kenya with their physician parents, planting in David a lifelong fascination with global health. *continued on page 42*

THE SCALE

On a computer, data is translated into 0s and 1s called bits. Eight bits make up one byte—enough information to represent one letter, number, or symbol.

2.5 PETABYTES
Memory capacity of the human brain

13 PETABYTES
Amount that could be downloaded from the Internet in two minutes if every American got on a computer at the same time

98 PETABYTES
Websites indexed by Google

4.75 EXABYTES
Total genome sequences of all people on Earth

422 EXABYTES
Total digital data created in 2008

1 ZETTABYTE
World's current digital storage capacity

1.8 ZETTABYTES
Total digital data created in 2011

Source: *Popular Science and Katie Peek*

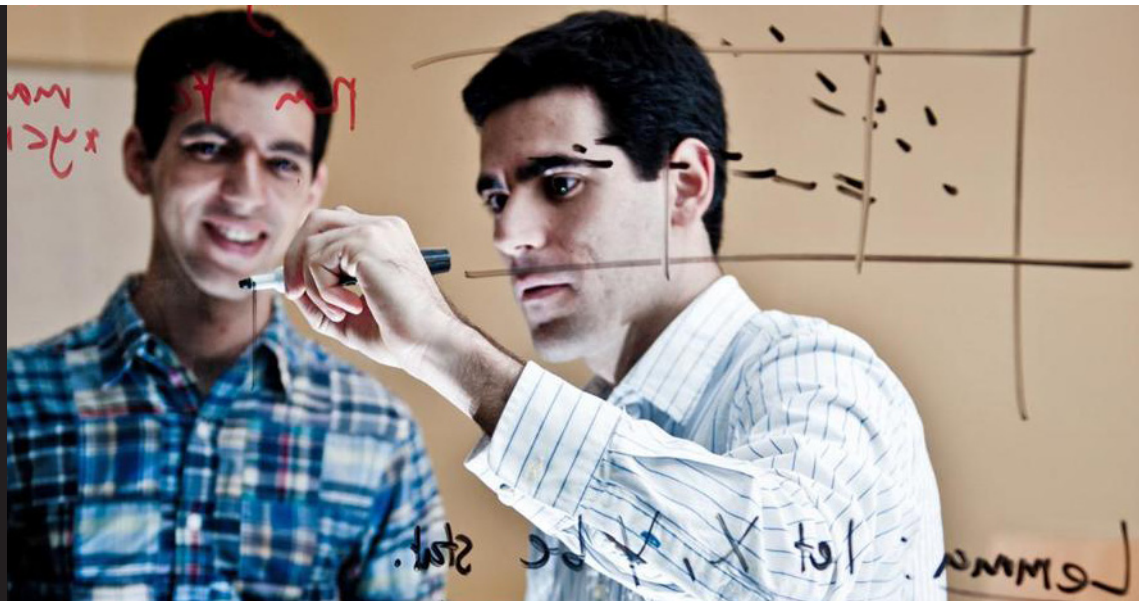
BIG DATA *continued from page 19*

In 2007, Reshef met Pardis Sabeti, an assistant professor in HSPH's Department of Immunology and Infectious Diseases and a computational biologist at the Broad Institute. Reshef talked excitedly about his desire to apply computational methods to public health problems. Sabeti, a geneticist who has made discoveries about malaria and the lethal African Lassa virus by mining big data, found Reshef remarkably like-minded. "You should come work with me," she told him.

They began developing tools for visualizing relationships in huge databases (including a World Health Organization database containing more than 60,000 relationships among data from 200 countries). But visualization tools work best when scientists have an idea of what to visualize in a pile of data. Reshef wasn't seeking the proverbial needle in a haystack; he wanted to find all the needles.

Scientists are excited by the potential of hypothesis-generating, rather than hypothesis-driven, science that big data mining offers.

To discover hidden relationships in the data, he needed a treasure-seeking tool, the computational equivalent of a metal detector. Reshef and his brother, Yakir, who was just graduating from Harvard with a math degree, started to spend every spare minute together, scribbling equations on the glass walls of the Broad and consulting with Sabeti and Michael Mitzenmacher, professor of computer science at Harvard. One hot night, running the latest version of their algorithm on a PC, they realized their program finally worked—and fast. (The algorithm now produces results in minutes or hours, depending on the size of the dataset. Without it, the data could take months to analyze.) "We were so excited, we called Pardis," Reshef says. It was 3 a.m.



David Reshef, right, and his brother Yakir teamed with HSPH's Pardis Sabeti to create tools for visualizing relationships in huge databases.

Over the next year, they tested the tool, called MINE, on several giant datasets, including the WHO data and a 6,700-variable database of the human gut microbiome that generated 22 million possible paired relationships. Last December, the Reshef brothers were the lead authors of a paper in *Science* that showed the tool's range. The algorithm has helped pinpoint interesting associations between gut bacteria, demonstrating that both diet and gender influence gut bacteria. The tool also identified nonintuitive associations between female obesity and income. In just a few weeks, more than 50,000 visitors tapped the MINE website, including, says Sabeti, visitors from "every imaginable field: genomics to finance to pharma to education and beyond."

ChieYu Lin, courtesy of Pardis Sabeti

HARMONIZING INCOMPATIBLE DATA

To analyze data, whether through crowdsourcing or algorithms, you have to start with a decent database—or several. Sharing massive datasets offers huge potential for improving public health. Biostatistics chair Victor De Gruttola is working on an Institute of Medicine project identifying indicators and methods for monitoring HIV care in the U.S. "There are many tremendous sources of information, but none are sufficient in themselves to gauge the prevalence of HIV care, as well as access to mental health and substance abuse treatment and support services," he says. For example, the U.S. Centers for Disease Control and Prevention captures diagnostic, demographic, and medical information, but no data on the use of antiretroviral drugs. Medicaid and Medicare track service use through claims data, but not clinical measurements such as immune function at diagnosis. De Gruttola posits that if researchers could join these datasets, they'd learn which vulnerable groups of patients aren't getting the treatments they need.

Easier said than done. That's in part because scientists employ a mélange of incompatible structures to create their data. Winston Hide, the biostatistics associate professor, has taken a step toward fixing that problem. He and researchers at 30 organizations, including Oxford University, have invented a common language and tools for sharing data across disciplines, called Investigation-Study-Assay (ISA). (For information, visit isacommons.org.) The technology is intended to be simple for researchers to use—a sort of scientific lingua franca.

In just two weeks, Hide joined a cancer database with a stem cell dataset—and got a big payoff. "We discovered a single gene that we think is responsible for the initiation of a whole class of leukemias," he says. "Not until we could combine the information coherently could we discover things about the underlying molecular biology."

A NEW WAY OF DOING SCIENCE

These innovative methods for mining big data are transforming the way science is done. Sabeti and Reshef are excited by the potential of hypothesis-generating (rather than hypothesis-driven) science, providing researchers with important new questions to answer. Analyzing genetic data for natural selection, for example, Sabeti had stumbled on clues to the virulence of Lassa fever, a deadly infection endemic in West Africa. She and Reshef believe that the hypothesis-generating power of big data will ultimately help researchers gain insights into the most pressing public health problems, such as the emergence and spread of resistant strains of malaria.

Meanwhile, Winston Hide believes younger scientists—raised in an era of social networking—will embrace an idea that previous generations of researchers have not: sharing data freely. It's an option that makes intuitive sense, he says, to generations raised with social networking.

Crowdsourcing is also breaking down the walls between the academy and the rest of the world. For many scientists, though, it's a tough transition: Academics have typically held their data close, because tenure, promotions, and reputation rest on being the first to publish. Sharing research takes a leap of faith in a cutthroat academic world that has yet to embrace the notion of a public commons of data. The change also comes with ethical questions, including privacy dilemmas. And employing crowds to analyze one's data begs the question of quality: How can you trust the results?

Most researchers agree that lives are lost every day that data sit in storage, untouched.

Surprisingly, Fortune says she trusts the results more than those that would have come from her lab. "I think the power of crowdsourcing is that they're going to give us better data than we can generate ourselves." That's because CrowdFlower uses redundancy to ensure quality (five people may analyze the same image).

Indeed, she's become a huge fan of speeding scientific progress through crowdsourcing. "I love the idea of citizen science," she says. "We're asking people to do some not very sophisticated tasks. You could stand in line at the bank and measure bacteria for me."

To a single citizen scientist labeling a batch of images, the work may feel tedious. In fact, it's transformative—a small contribution to what may be public health's data-driven revolution. "It's just the beginning," says Winston Hide. "You should watch this space."

Elaine Grant is assistant director of development communications and marketing at HSPH.