

What's Big Data Got to Do with It?

Advances in computing power
and a wealth of digital
information are changing
scientific research.

By Kathleen McGarvey

In June—and in digital

culture, that's already a good while ago—the CEO of the social networking service Twitter, Dick Costolo, announced that users were posting 400 million tweets a day. And that was up 60 million tweets per day from the figure just three months before. It all adds up to a billion tweets every two and a half days.

As a microblogging service that allows people to post messages of no more than 140 characters, Twitter is an immense but transitory compendium of observations, insights, outbursts, and mundanities. What value could it have for scientific researchers?

A lot, as it happens. Henry Kautz, chair of the computer science department, and colleagues Adam Sadilek and Vincent Silenzio have shown that Twitter messages can be harnessed to predict the spread of infectious diseases, such as influenza. This year, they have published two papers explaining how, by using the geo-tags embedded in tweets, scientists can use social networking data to model the transmission of disease—and even to forecast when and if a specific individual will fall ill.

Kautz, Sadilek, a postdoctoral fellow in computer science, and Silenzio, associate professor of psychiatry and a member of the Department of Community and Preventive Medicine, have programmed computers to identify tweets in which people talk about feeling sick—disregarding messages where people use the term figuratively.

“Once you have that, you can start to map where people are sick,” Kautz says, because GPS in cell phones indicate where a tweet was made. “And you can actually start to create a visualization of the spread of disease through cities and across time.”

“These results provide a foundation for research on fundamental questions of public health,” the team writes, “including the identification of non-cooperative disease carriers (‘Typhoid Marys’), adaptive vaccine policies, and our understanding of the emergence of global epidemics from day-to-day interpersonal interactions.”

And they note that the approach has applicability far beyond infectious diseases, for modeling and predicting political ideas, purchasing preferences, or nearly anything else rooted in behavior.

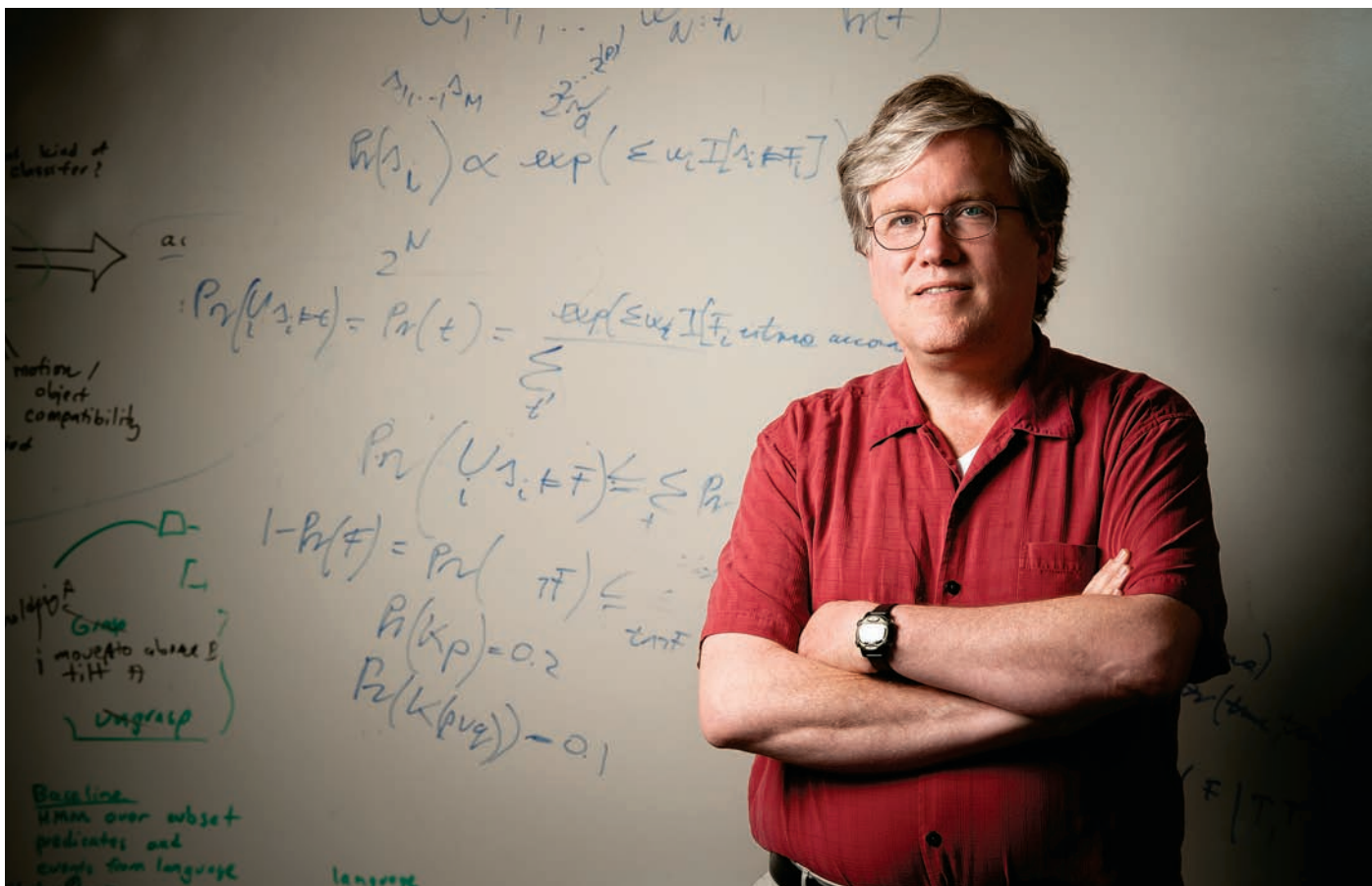
“It's actually pretty neat,” Kautz says—so neat that they've formed a venture capital-funded start-up business, Corpora, that makes use of the technology for applications in areas such as health care, insurance, pharmaceuticals, government agencies, and public opinion tracking.

Such ingenuity, combined with vast quantities of information and high-performance computing, is changing the parameters of knowledge.

We call ours the “information age”—an era marked by an endless digital trail revealing what we do and where we do it, and much that's happening within us and without us.

Anyone who has used the Internet is already familiar with the ways businesses have seized on that trove of information to predict and guide the choices we make as we purchase books and shoes, vacations and music.

But the possibilities of “big data”—the fast-emerging shorthand term for the efficient analysis and problem-solving application of vast quantities of data—are profound for science, medicine, and other areas of research. Through high-performance computing, creative computer science, and new bonds of collaboration, researchers find themselves at the brink of what many predict to be a new age of investigation and



advances in knowledge—comparable, the *New York Times* has suggested, to the introduction of the microscope and the telescope.

Rochester is at the

forefront, pairing teams of researchers and computational scientists with supercomputing technology to transform data into knowledge.

Applications range widely. Why do countries go to war? Curt Signorino, associate professor of political science, is using data mining tools drawn from genetics and finance to compare data on every combination of countries from the years 1900 to 2000, creating an explanatory model that fits the data more than three times better than standard techniques.

How can energy flow through the power grid to make sure that electricity is reliably delivered to people where they need it, when they need it? Mark Bocko, professor and chair of the Department of Electrical and Computer Engineering and director of the Center for Emerging and Innovative Sciences, is studying the dynamic behavior of the power grid and how to control it with the tactical use of data—what has become known as the “smart grid.” He and his team are developing imaging, sound, and vibration sensors that will sort through information at the source, curbing the amount of transmitted data so only the most useful is passed along.

How can we better fight the flu, which claims the lives of 30,000 to 40,000 people each year in the United States? David Topham, vice provost and professor of microbiology and immunology, and colleagues are working to build a computer model of the immune system that will allow for simulations of infections and possible vaccines before the flu strikes—thereby speeding production of effective vaccines, and saving lives.

Pedro Domingos, associate professor of computer science at the University of Washington, who will be a featured speaker at a conference on big data to

be held at Rochester in October, says there are few if any fields that will be untouched. “Science, in just about every area, without big data will grind to a halt. It will be a field of diminishing returns.”

Kautz, who is director of an initiative for big data in Arts, Sciences & Engineering, says complex problems in science, mathematics, engineering, and the social sciences have traditionally been approached by breaking them into smaller pieces, understanding how each works, and then deducing solutions to the larger problems.

But systems science—a broad and interdisciplinary field underpinning big data that studies the behavior of complex physical, biological, artificial, and social systems—has upturned that approach, focusing on the whole instead of the parts and ushering in a new scale for problem solving. It provides a fresh capacity to see how things interrelate and influence each other, from the molecular level to entire populations.

“I’m an immunologist,” says Topham. “I was trained in cell biology, so I like to study individual cells.”

Formerly, he would collect a blood or tissue sample, isolate the cells, and from experiments on them, garner a few elements of data. Now, when he and his colleagues

CREATIVE SOLUTIONS: Henry Kautz, professor and chair of computer science, says “big data” often requires a new approach to problem-solving in science. “What I think we need are more people thinking of extremely creative ways to use these machines,” he says.



carry out clinical studies, they pursue many more dimensions of cellular investigation.

Computational approaches are going to “allow us to identify biological relationships between cells and proteins, microorganisms and the host, that we wouldn’t otherwise have been able to detect, and then understand how these affect our ability to respond to vaccines or disease,” he says. A computational take on science has inverted the relationship between experimentation and analysis. Carrying out experiments used to consume about 75 percent of researchers’ time, and analysis the remaining 25 percent, but “I would say that’s reversed now,” he says. “You can do one experiment, and it will take weeks to analyze the data.”

While computers, computational methods, and data collection are advancing rapidly, these are still early days for big data. When researchers talk about the data now available, they seem to reach almost instinctively for metaphors of water: a deluge, a flood, a relentless torrent of information to be channeled and controlled.

“It’s not just more streams of data, but entirely new ones,” says the *Times* about what it terms a “data flood.” An influential report on big data issued last year by

SCIENTIFIC VIEW: As part of work that explores the planet as a complex system, researchers are using large data sets to bring an interdisciplinary approach to earth science, says Carmala Garziona, associate professor and chair of earth and environmental sciences.

McKinsey Global Institute, the research arm of the global management consulting firm McKinsey & Company, invoked the idea of “large pools of data that can be captured, communicated, aggregated, stored, and analyzed” today.

“We don’t know how to manage this information. It’s like drinking from a fire hose—how do you control it so that you don’t become overwhelmed?” says David Williams, dean for research for Arts, Sciences & Engineering and the William G. Allyn Professor of Medical Optics.

As critical as the

availability of data is the capacity to select from and organize it—to sort out the most useful elements, the most meaningful patterns, the formerly unrecognized connections—and transform the flood of data into something of practical value.

“It’s a bit like prospecting in the old days of mining, because you’re looking for nuggets of gold,” says Rob Clark, dean of the Hajim School and interim senior vice president for research.

But it’s not a passive search. “I think ‘big data’ is a term, like ‘cloud,’ that’s getting thrown around so much that it’s getting distorted,” says David Lewis, vice president for information technology and CIO. “To us, ‘big data’ is doing something with the data—you’re doing the analytics.”

Such analysis has emerged as a national priority. In March, the White House’s Office of Science and Technology Policy announced a “Big Data Research and Development Initiative” aimed at bringing together research universities, industry, and nonprofit organizations with the federal government to take advantage of the opportunities big data offers for science and innovation.

“The technology for generating new data is always far ahead of our ability to analyze it. It has become a major, global problem,” says Topham. “The real data comes when you can relate different kinds of data, find the connections—and that’s very difficult to do. It almost requires intuition.”

Intuition is a tough thing to teach, but through courses in data mining, biostatistics, and algorithms, students are acquiring the skills needed to swim proficiently in a sea of data. Clark says the secret lies in teaching students the basics of how to manage information, big or small, “to extract kernels of useful information from data sets.”


Such extraction is changing scientific research across the disciplines. In Arts, Sciences & Engineering, earth and environmental sciences and chemical engineering have to take a big data approach.

“We really view it as refining the tools for supporting the mechanics of research,” says Carmala Garziona, associate professor and chair of the earth and environmental sciences department. “We’re basically moving from a very discipline-oriented science, where you would have a group of researchers who’d look at some very specific aspect of the earth, to a much more interdisciplinary science, where groups of researchers are working across disciplinary boundaries to understand how the earth behaves as a complex system.”

The kind of transition she describes is one taking place across disciplines, says Washington’s Domingos. “I think a mental shift has to happen in how scientists think about doing science.” Graduate students and researchers early in their careers have been professionally formed in an environment of computational approaches, but for more established scientists, he notes, big data requires an adjustment to a new way of pursuing research questions.

At the Medical Center, it’s an approach that is swiftly becoming central. The University, New York State, and IBM have partnered to establish the Health Sciences Center for Computational Innovation. It’s home to the IBM Blue Gene/Q supercomputer, making Rochester one of the five most powerful university-based supercomputing sites in the country.

“It’s one of the most powerful supercomputers dedicated to health research in the world,” says Topham, director of the HSCCI. “The Blue Gene/Q lets you run experiments that otherwise wouldn’t be possible.”



POWERHOUSE: At peak performance, IBM's Blue Gene/Q supercomputer can make 209 trillion calculations a second.

A Super Supercomputer

Rochester is one of the first academic homes in the nation for IBM's next-generation supercomputer, designed to 'make knowledge out of data.'

A pivotal piece of big data research at Rochester is the Health Sciences Center for Computational Innovation (HSCCI), which in August became home to the next-generation supercomputer built by IBM—the Blue Gene/Q. It makes Rochester one of the five most powerful university-based supercomputing sites in the nation.

"It will be one of the most, if not the most, powerful supercomputers dedicated to health science research in the world," says David Topham, director of HSCCI. The University created the center in 2008, in partnership with IBM, and began work with what is now the previous generation of supercomputer, the Blue Gene/P. In collaboration, the University, New York State, and IBM have upgraded the center with the Blue Gene/Q. The Center for Governmental Research estimates that the project could create 900 jobs in the community and generate \$205 million in new research funding over the next 10 years.

At peak performance, the BlueGene/Q can make 209 trillion calculations per second. It's 15 times more powerful than the previous generation supercomputer—and has the computing power of about 20,000 laptops. The supercomputer, which will enable scientists to sift through mountains of data and create complex models, has vast potential for applications in medicine, and Rochester scientists are applying high-performance computing to research programs in vaccine development, brain injury, and cardiac disease.

Topham calls it "truly a new domain" for research, as supercomputing technology allows researchers to ask fundamentally different questions about health, creating "knowledge out of data."

At universities around the country, just a decade ago, when researchers needed to summon more computing power than could be found in a conventional machine, they created what were known as "Beowulf clusters"—100 or 200 desktop computers linked together. It took a lot of space, produced a lot of heat, and consumed a lot of energy—and it was only sustainable for a day or two. Other researchers would cluster their own servers, requiring special air conditioning and running up electricity bills.

Faced with that insupportable situation, faculty at Rochester came together to find a solution.

"They had different research domains, but they had the same core issue: they needed a facility that could help them accelerate their research results," recalls David Lewis, vice president for information technology and CIO.

The solution they hit upon was a shared center—what is today the Center for Integrated Research Computing (CIRC), which advances research through high-performance computing. It supports 150 research groups across the University.

"We started with 16 pilot users. Now we have more than 500 researchers" from 35 University departments and centers, says Brendan Mort, director of CIRC.

"The results have been exponential," says Lewis. "We've built this as a community, and people have contributed hardware and people. They feel they can get bigger outcomes by being part of a shared resource rather than trying to build their own thing."

—Kathleen McGarvey

For example, Jean-Philippe Couderc, associate professor of cardiology and assistant director of the Heart Research Follow-up Program Laboratory, and colleague Coeli Lopes, assistant professor at the Aab Cardiovascular Research Institute—along with Jeremy Rice of IBM’s Watson Research Center—plan to use Blue Gene/Q in modeling the heart to test drugs’ effects on the organ.

In 2004, the Food and Drug Administration launched an initiative designed to bring medical breakthroughs to patients more quickly while ensuring safety and reducing the costs of drug development. Key to that effort has been developing better ways to test the cardiac toxicity of drugs—a leading cause of drugs being removed from the market.

Together with the FDA, the University in 2008 established an electronic repository of electrocardiography data—the Telemetric and Holter ECG Warehouse, or THEW—to help foster research in the field. The database is part of the Center for Quantitative Electrocardiology and Cardiac Safety, funded by a \$2.3 million grant from the National Institutes of Health and a part of the University’s Heart Research Follow-up Program. It brings together an international network of academic researchers, pharmaceutical and medical device companies, and government regulators. Data from the center is provided to academic and private research organizations to help them design and validate new tools and methods to detect abnormal cardiac activity.

The repository makes Rochester the hub of a heart research wheel that spans the globe, from academic institutions and industry in Europe to Asia to South America. “We are the only academic group in the world that provides an open resource of ECG data for drug safety evaluation,” says Couderc. Last year alone, 25 publications were produced from repository data. Together with Lopes and Rice, Couderc is using data from THEW to model effects of drugs on cardiac cells, using a computerized model of the heart system produced by the National Library of Medicine in cooperation with IBM.

From an anatomical point of view, the model’s an excellent representation of the heart, he says. He and his team are being trained in using the Blue Gene/Q computer to evaluate the effects of drugs on a wedge of the heart—its inner and outer layers—and the millions of different cells that form them. They check the results of the model against the documented results shown in THEW’s records.

“IBM brings a unique tool; the University brings unique data sets, enabling such tools to have a significant impact on drug-safety evaluation and medicine,” he says.

But one of the most

important ingredients in that equation is the imagination and inventiveness of people engaged in research. “We lead with people, not with computing,” says Lewis.

It’s an emphasis that others echo.

“This is all about the people. In fact, the people are the far more valuable and important component of this partnership,” says Topham. “Yes, you need hardware—but you can’t use the hardware if you don’t have the right people, and IBM has a very strong interest in disseminating the knowledge of how to use these tools to deal with important questions.

“Our health sciences researchers have the questions. We have the patient population. We have the ability to generate the data. We just need the tools to analyze it. And together, the University and IBM can do much more than either one of us can do on our own.”

The collaboration doesn’t extend only to IBM researchers; big data is strengthening ties between researchers all over campus. Big data allows “the creation of teams of investigators,” says Williams. “I think we’re going to see a lot more collaboration between investigators at different institutions” because of improved communications and the ability to transmit data.

The studies being carried out through HSCCI require many people, and a wide variety of expertise, says Topham. “I have pediatricians, infectious disease

specialists, immunologists, neonatologists, researchers in genomics and microbiomics, computational people, data management—we have a huge data management core to deal with all that data.”

Kautz’s vision for big data at Rochester focuses not on supercomputing hardware—“we have that well in hand,” he says—but on the people who use them.

“What I think we need are more people thinking of extremely creative ways to use these machines, as well as other resources.”

With the rise of big data, computer scientists take on a pivotal role in the research of many fields.

“One aspect of providing computational support to a physical scientist is saying, we’re programmers. You tell us what to do and we’ll run that package. But there’s this other role in terms of helping think more deeply about the problem, because that’s the new way of solving the problem. And you need to do both.”

Bringing big data to bear on the field of environmental sciences, for example, will “require a strong collaboration between computer scientists and earth scientists,” say Garziona. “Ultimately, the department plans to hire “earth scientists with a strong computational bent”—Vasilii Petrenko, an atmospheric chemist, joined the faculty last year, and John Kessler, a chemical oceanographer, came on board for this academic year—“or computer scientists who are capable of tackling problems in other disciplines with a healthy dialogue that enables them to find a solution to computational problems.”


And solutions found in one area can, at the computational level, provide keys to other areas, far afield.

Algorithms that Kautz’s students are developing for mining social network data from Twitter, for example, might turn out to be relevant for doing computational biology. “That kind of thing happens all the time,” says Kautz. “You look at a problem with the right level of abstraction and you realize, ‘Gee, we can think of both of these things as a network, and we’re trying to find certain patterns.’”

In his lab, Topham says, researchers are studying vaccines and immune responses, but the methodological advances they make “could apply to cancer, to development, to cognition. They could apply it to environmental questions.”

The possibilities excite him.

“I’m hoping we get to do some really important research, that we solve some long-standing questions,” he says. “Developing new vaccines. Imaging the brain better so that you can tailor treatment more effectively. Understanding individual cells’ behavior, in the brain or in the immune system—these are the key questions that have just been lingering out there in the field.

“And we now have the technologies to study these things in ways we didn’t before.” 

The University will host a conference, The Rochester Big Data Forum 2012, October 4–6. To learn more, visit www.rochester.edu/rocdada.