University of Rochester

BIG DATA IS THE FUTURE

"From the dawn of civilization until 2003, humankind generated five exabytes of data. Now we produce five exabytes every two days . . . and the pace is accelerating."

—Eric Schmidt, Executive Chairman, Google

BIG DATA ANALYTICS

"Big Data" is big news—we often hear how it is transforming science, business, and our everyday lives. Yet for all of the excitement around this idea, stories about Big Data often overlook what is most important about the field—not just that we are swamped with data, but, rather, that scientists and engineers are creating powerful new techniques for collecting, managing, and analyzing huge data sets. The generality of these methods breaks down traditional barriers between disciplines. For example, the data structuring and management methods originally developed for displaying and linking text on the World Wide Web turn out to be a key technology for integrating large scientific databases.

This publication presents a series of case studies of research at the University of Rochester in the area of Big Data analytics. Analytics can be thought of as "sense-making"—finding meaningful patterns and high-level concepts in data. For example, at the low level, a CAT scan is simply an image of a slice of the body. Analysis algorithms can segment the image into different organs. At the highest level, analytics can summarize the data in an extremely concise and meaningful way, such as "normal" or "diseased."

Big Data analytics has emerged from the combined efforts of researchers in computer science, statistics, and economics and in the physical, life, and social sciences. The area goes by several different names—for example, it is called "machine learning" or "data mining" in computer science and "predictive modeling" in finance. However, the researchers share an ever-growing toolkit of mathematics and computational methods. It is not uncommon to find a team including scientists trained in physics working on problems in finance, or for techniques developed for processing natural language text to find a use in genetic analysis.

In addition to researchers who represent a broad spectrum of traditional disciplines, Brendan Mort of the Center for Integrated Research Computing (CIRC) describes the center's state-of-the-art, high-performance computing facilities and consulting services, which enable these projects and many others.

These projects are just a sample of Big Data research at the University of Rochester. Work in Big Data is continuing to grow as a high-priority area for expansion in all of the University's colleges and schools.

In fall 2012, the University hosted a national Forum on Big Data, which brought renowned researchers from across the nation, leaders of federal research funding agencies, representatives of high-technology companies, and faculty and students from the University of Rochester together for three days of presentations, discussions, and networking. Video presentations from the forum can be viewed online by visiting **www.rochester.edu/rocdata/forum2012.**

I hope you find this overview informative and exciting. Big Data is the future, and the University of Rochester is helping make it happen.

Henry Kautz
Chair, Department of Computer Science
Head, Rochester Big Data Initiative

The University of Rochester researchers whose work is highlighted here represent a broad spectrum of traditional disciplines.

[Matthew Blackwell, political science] uses Big Data analytics to understand political campaigns and other issues in American politics.

[Daniel Gildea, computer science] develops systems that translate between human languages, such as English and Chinese, by training on huge corpora of parallel texts.

[Henry Kautz, computer science] is data mining social media such as Twitter in order to track and predict disease outbreaks.

[Jiebo Luo, computer science] works on image understanding, developing systems that can automatically label images, videos, and other kinds of multimedia.

[Rajeev Raizada, brain and cognitive sciences] uses pattern-based fMRI analysis in order to understand the way the brain encodes and processes information.

[Huaxia Rui, Simon School of Business] focuses on how business can make use of data from social media sites such as Twitter and Facebook to improve decision making.

[Vincent Silenzio, psychiatry] works on suicide prevention among at-risk youth and explores the use of online social networks to gather data from people who are otherwise difficult to identify or survey and to provide peer support.

[Robert Strawderman, chair, biostatistics and computational biology] provides an overview of the University's growing portfolio of research in large-scale medical statistics and how it is changing our approach to personalized medicine and health care.

[Axel Wismueller, biomedical engineering] develops novel, intuitively intelligible computational visualization methods for the exploratory analysis of high-dimensional data from biomedical imaging.

Super Computing Takes the Pulse of Politics

Matthew Blackwell

With the advent of the Internet and super computing, political scientists are discovering novel ways to monitor the mood of the public through blog texts, Facebook postings, Twitter messages, and even the expressions on candidate photos posted to the web.

Add to that list campaign contributions. Matthew Blackwell is developing methods to use contribution data, now available in online databases, to follow the pulse of campaigns. In particular, he has developed a novel way to identify the critical moments when a political campaign either takes off or falls flat.

"Campaigns rarely have smooth trajectories," says the assistant professor of political science. "Instead, they tend to go through discrete phases punctuated by decisive turns, up or down." For example, a convincing debate performance by a candidate can give the campaign a surge of support, while a gaffe may totally break it.

Traditionally, political observers have tracked such change points by relying on polls. The problem is that in smaller campaigns, like state legislative contests or less contentious congressional races, polls can be scarce. With campaign contributions, "we can take all of the data that the federal government collects on campaign contributions to get a sense of a race on a day-by-day level," says Blackwell.

Using statistical analysis, including a Bayesian change point model and a Markov Chain Monte Carlo estimator, Blackwell is able to tease out change points amidst the normal ebb and flow of campaign contributions. Recently, he applied this approach to the campaign of Herman Cain during the 2012 Republican presidential primaries. He found that the change points predicted by his model corresponded almost to the day with major events in Cain's campaign.

"You find that change points tend to happen when there's a lot of news about a campaign," Blackwell explains. "More people are paying attention to the candidate, and so the campaign gets more money. By looking through this data, we can get a sense for what causes these things."

With the advent of the Internet and super computing, political scientists are discovering novel ways to monitor the mood of the public through blog texts, Facebook postings, Twitter messages, and even the expressions on candidate photos posted to the web.

Blackwell is optimistic that rigorous analysis of data can clarify certain ideas in political science. Already he's discovered something surprising: "There's a feeling in political science that campaigns don't matter a lot. People are supposed to base their decisions on things that no one can control, like the state of the economy. But the data reveals that voters very much do pay attention to the horserace."

He predicts that political commentators will increasingly embrace data-driven methods. The 2012 presidential election, for example, proved the predictive power of data aggregators over traditional polls alone or pundit forecasts, he says. "There's a feeling among a lot of people that politics can't be predicted," says Blackwell. "But I think that this is just a change in technology, and you'll see a slow acceptance of these kinds of models."

Technology change is redefining the practice of political science as well. For decades, researchers have relied on large, canonical databases that were compiled by numerous researchers and an army of assistants. "It's become much easier for researchers to pull down completely original sets of data in a relatively short amount of time very cheaply and then analyze that on their own very quickly," says Blackwell. "There's been a very real increase of neat projects with cool new data."

"What's exciting about Big Data is that it's leading people in a number of fields to realize they have similar problems," he says. "The more data researchers collect, the more people look to different fields to see how others have solved similar problems. And the more this kind of iteration goes on, the more people start to come together."

Blackwell is excited about the collaboration. "Even though disciplines use different terminology, researchers realize that there are fundamental issues that are similar across fields. Maybe there are creative solutions that we can find for problems in our own discipline from the approach others have developed."

Daniel Gildea

Computers Learn the Fine Art of Translation

Just a few decades ago, anyone seeking a translation of a text written in a foreign language had to find a capable person to translate it.

Now, computer scientists and linguists have created automated translation programs that can roughly translate back and forth in many of the world's major languages. And while those programs are still imperfect, they are steadily improving, thanks to the continuing work of researchers like associate professor of computer science Dan Gildea and his colleagues at Rochester.

Gildea works in the field known as machine translation, an area of natural language processing. Machine translation is a big challenge for computers as it not only requires knowledge of the languages that are being translated but also an understanding of idioms, double entendres, and often even pop culture.

Gildea generates algorithms that can translate from one language to another. These algorithms can be applied to any language, but he and his team have been concentrating on translating from Chinese into English.

China's growing economic power and the increasing number of Chinese Internet users—currently about 450 million, or one and a half times the U.S. population—ensure a growing demand for such translations.

Translating from Chinese into English has some intrinsic challenges—whether it's a machine or a human doing the translation. For example, verbs don't have tenses in Chinese. So to understand whether the English translation of a verb should appear in past, present, future, or conditional, it is not enough to simply look at the verb in Chinese. The computer or the human doing the translation needs to find a word somewhere else in the sentence—such as "today," "later," or "yesterday"—that provides that information.

This is not straightforward for a computer. The real challenge comes in how to apply powerful statistical techniques to create the algorithms a computer will use to translate. The algorithms are effectively the logic in the computer's "brain," and they need to learn how to translate.

Just as approaches to teaching a foreign language have changed over time, Gildea explains, so have the models used for machine translation.

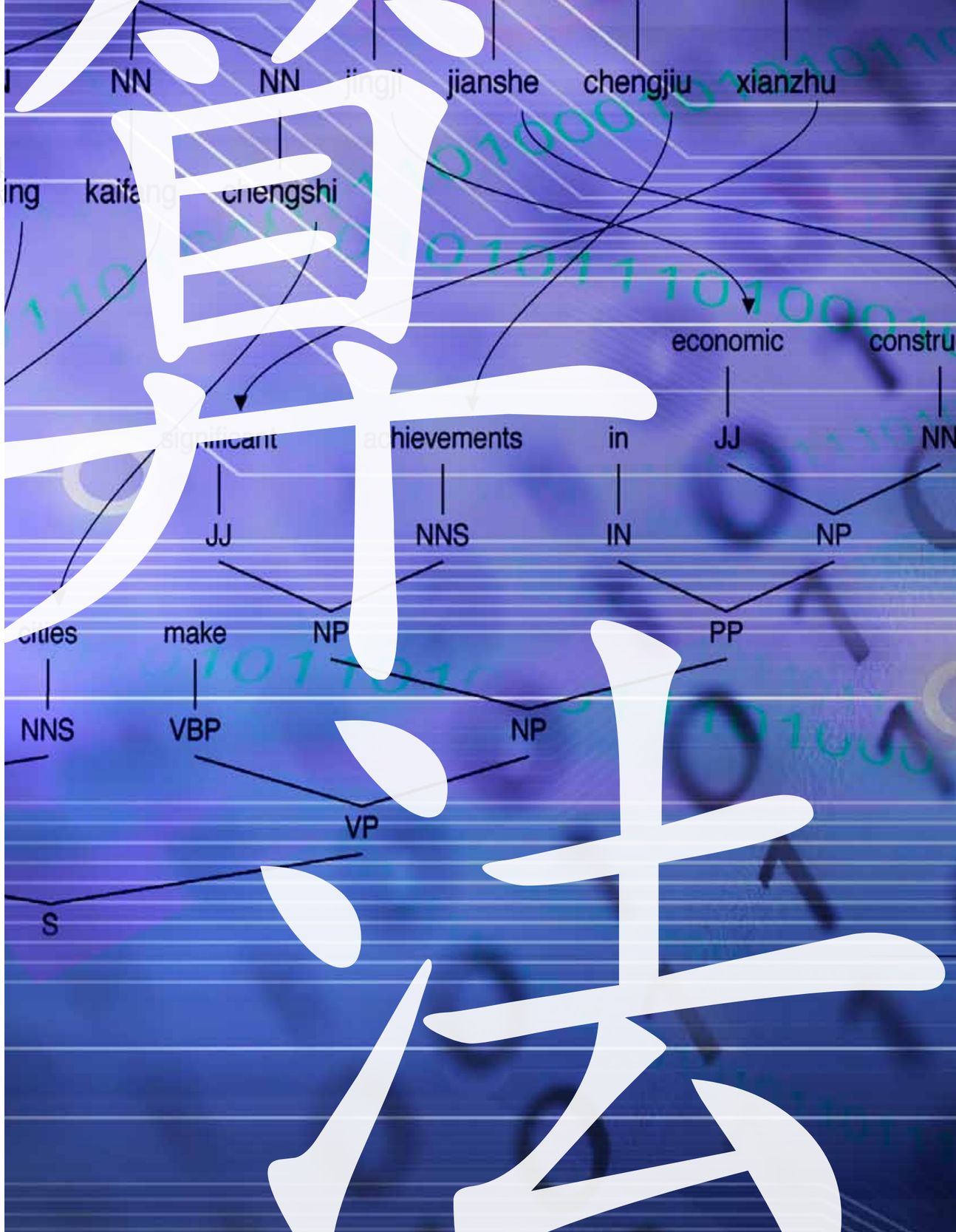
For example, when teaching a foreign language these days, teachers no longer require students to become proficient in the grammar before speaking to them in the foreign language. Similarly, recent approaches to machine translation do not require the computer to have "memorized" all grammar rules of the language in advance.

Instead the computer is taught by analyzing the same text in two languages. This is the model Gildea and his team use, which is also used by hundreds of researchers and by programs like Google Translate.

The researchers repeat this process many, many times so the computer can start to recognize certain words, sentences, and grammatical constructions. Vast amounts of data in the form of translated news and websites that exist on the Internet are the perfect training material for these computer translators. The more texts that are fed to the machine, the more likely it is that similar constructions will be present in different texts for the computer to compare.

Gildea generates algorithms that can translate from one language to another. These algorithms can be applied to any language, but he and his team have been concentrating on translating from Chinese into English.

It is the role of Gildea and other researchers to develop algorithms that extract information from the texts by observing patterns. These are stored in a hierarchical structure called "semantic trees," from more general sentence structures to more specific word endings. For example, a phrase that contains "taller than" or "quicker than" will always require an object, for example, a "him," "her," "Peter," or "the dog." And the next time a translation into English requires this phrasing, the algorithm will know it needs to find the appropriate word to fill in the gap.



Jiebo Luo

Each Image Tells a Thousand Words

After being a researcher at Kodak for more than 15 years, it's not surprising that Jiebo Luo would focus his University research on images.

Now an associate professor in the computer science department, he has, like many of his colleagues, become intrigued by the challenges and possibilities that Big Data pose. But unlike many of his colleagues who are focusing on the information contained in the words and sentences on the Internet, he analyzes images.

"It really is the case that each image contains a thousand words," said Luo to kick off his presentation at the Big Data Forum in October 2012 at Rochester. For example, the iconic painting of the Mona Lisa would "contain" visual words (i.e., visual patterns and features) like "hair," "face," "color," "hands," and maybe even "smile."

Working with images brings its own problems, however. Not only do you have to mine the web to find images, often in diverse formats, from which you can extract information—you also need the computer to be able to understand, somehow, what the image is about.

Training a computer to identify certain features is the first step. But these features need to be unique or identifiable in some way. "When you look through the viewfinder of your digital camera as you're about to take a picture these days, you'll probably see little squares around everyone's faces," Luo says. "This is the kind of technology we use."

He and his colleagues also try to teach a computer to identify other objects, such as a car or a castle. But it's not easy—many of us have seen our camera insist that the tree next to the friend we are photographing is also a person. "It is easy for a human to identify a chair; we all know what they look like," Luo adds. "But it's nearly impossible for a computer to do so; chairs can look just too different." So it is even more impressive that Luo and his team have been using images to predict election results.

Looking at the trends of pictures of what products people post and comparing these with past sales data has allowed the researchers to come up with predictions of how various products are selling.

The use of social media during the 2012 presidential campaign was widely discussed. Every new debate involved the "most ever" simultaneous tweeting—until the next one. And a lot of these media included pictures.

As the outcome of election night became clear, President Obama tweeted a picture of himself with his wife, Michelle, entitled "Four more years." . . . That tweet went viral globally and almost instantly; within just a couple of hours it became the most retweeted message since Twitter began in March 2006. It captured the moment, and in retweeting this picture people expressed a view.

Luo explains that, in a sense, every time an image is uploaded or downloaded from the Internet it is like a vote in favor of or against a candidate or position. He finds that drawing on this information—across millions of Internet users—is like tapping into the wisdom of the crowd.

So Luo and fellow researchers taught a computer to recognize pictures of Barack Obama, Mitt Romney, and the vice presidential candidates, as well as pictures of campaign signage, and to collect these images so the researchers could analyze them. And by gathering these pictures they also get other data. When a picture is uploaded, other information is usually included, such as time and location. Putting all this information from the images together with real-world polling data as the training data and using state-of-the-art data mining techniques, they called the winner correctly in all the swing states.

But their model was more sophisticated than just whether an uploaded picture was of Obama or Romney. Accounting for things like whether the picture was meant as a criticism or endorsement of whom it represented, predictions are made more accurate.

And they have found their model can be used not only for elections but also for predictions of market sales of products.

"There might be a lot of people searching for the 'iPhone 5' on Google or tweeting how they really want it, but talk is cheap," says Luo. "If you're uploading an image of a specific product, it is quite likely you actually own it."

Looking at the trends of pictures of what products people post and comparing these with past sales data has allowed the researchers to come up with predictions of how various products are selling.

Luo and his team are still working on this, but their next research paper might tell how a specific company's stocks might do before the company has announced it.

Rajeev Raizada

Our Neural Fingerprints

Imagine a day when neuroscientists will use a brain scan to diagnose the underlying causes of learning disabilities like dyslexia and to detect such impairments long before children experience difficulty or, potentially, failure in school.

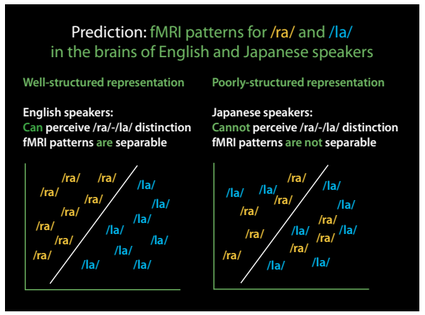
With advances in neuroimaging techniques and the computational ability needed to sort through these data-rich scans, that day may arrive sooner than you expect.

Cognitive scientists like Rajeev Raizada, who will be starting as assistant professor in the Department of Brain and Cognitive Sciences in July, are laying the foundation for such diagnostic abilities by turning to functional magnetic resonance imaging (fMRI). Unlike X-rays, CAT scans, and other types of brain imaging, fMRI involves no surgery, dyes, or exposure to radiation and can be safely deployed over time, providing a risk-free way for scientists to watch our brains in action.

"The brain has about 100 billion neurons, and they send electrical impulses to each other in a few thousandths of a second," says Raizada. Scientists use fMRI to indirectly capture that electrical activity by picking up the increases in blood oxygen that occur when thousands of neurons become active. "If you pump your muscles, the body sends more blood to the area," Raizada explains. "When neurons are active, the circulatory system increases their blood supply to provide more oxygen and glucose."

"By a lucky quirk of nature, oxygenated blood has a slightly different magnetic signal," he explains. The scans virtually segment the brain into a three-dimensional grid of about 40,000 pixels known as voxels and, using a magnetic field, measure the changes in oxygen levels in each tiny segment.

The result is a huge amount of information about what's going on inside the skull. But it is precisely that wealth of data that is part of the challenge. An fMRI scan creates about one image every two seconds. Multiply that over time—some studies record data for more than 20 minutes at a stretch—and by the dozen or more participants in the typical study, and the data mushrooms.



That's where super computing comes into play. Researchers are hard at work creating statistical algorithms and other computations to sift through the millions and millions of data points the scans create, says Raizada. First they have to separate out the background "noise," things like random fluctuations in blood flow or the magnetic field of the instrumentation that don't relate to the action being studied.

The ultimate challenge is to home in on the signal of interest, Raizada explains. For example, if researchers ask participants to look at a series of objects, they want to isolate just the brain patterns related to that activity. To find those patterns, Raizada compares the scans of different people all performing the same task and looks for similarities.

Here's the rub. Although brains are broadly similar, thinking patterns are individual. Says Raizada: "Each person has his or her own idiosyncratic neural fingerprint."

To tease out the similarities amidst the difference, Raizada looks at the relationship between one person's neural patterns and compares it to the relationships between others' neural patterns. Using such correlations, he says, is one possible solution to decoding the brain's thought processes. It's a puzzle that neuroscientists are approaching from many different angles.

Ultimately, Raizada believes neuroimaging may prove most useful in diagnosing the source of cognitive problems missing from behavioral tests alone. For example, two children may have outwardly similar difficulties with reading, but a brain scan may show that the impairment arises from different sources. One child may be struggling with attention issues, while the other child may have problems with phonological awareness.

Such distinctions are critical, notes Raizada. "Different types of impairment call for completely different types of treatments."

Big Data could provide an early warning system on disease outbreaks

Henry Kautz

Big Data research at the University of Rochester could eventually help authorities identify global disease outbreaks in their earliest stages and track their spread.

It is the next step in a project that made international headlines. Henry Kautz, chair of computer science, and Adam Sadilek, now a postdoctoral fellow, demonstrated that they could predict which Twitter users would get the flu—up to eight days in advance—by "mining" the social media network for tweets of people reporting symptoms in the New York City area.

They used the GPS tags embedded in the tweets sent from cell phones to track those persons' encounters with other Twitter users, whose own risks of becoming ill could then be calculated and tested.

The approach has the potential to dwarf previous methods for health monitoring in scalability and immediacy. The project has expanded to include social network data gathered from cities and airports around the world. "Today, it can take months to figure out where in the world a disease outbreak originated—and meanwhile people from that area will be carrying the disease all around the world," Kautz says. But by applying large-scale machine learning methods to a social network like Twitter, "in a matter of days, we could say there's a disease outbreak in Los Angeles and it looks like the point of origin could be Buenos Aires."

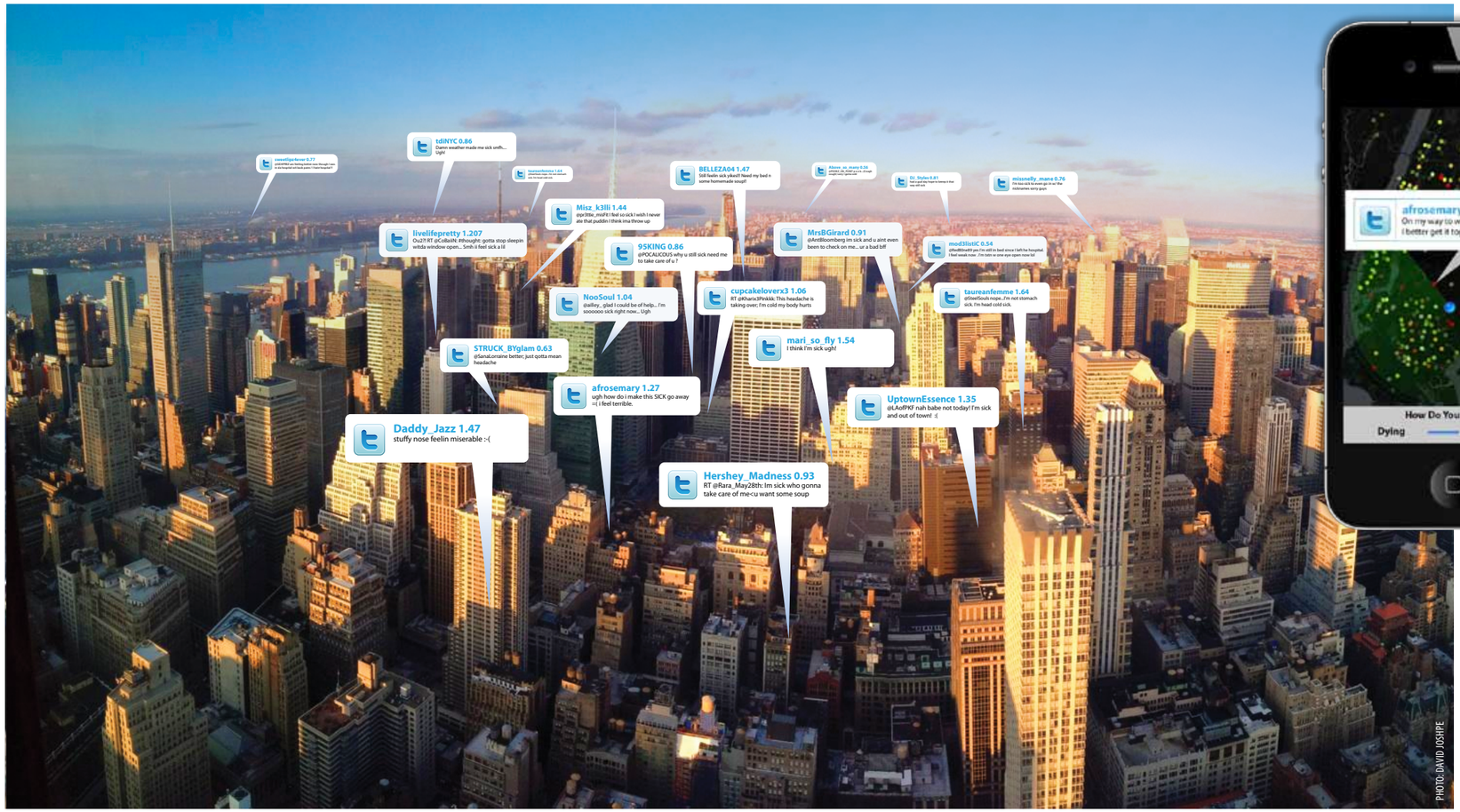


PHOTO: DAVID JOSHPE

Kautz is confident this approach "could give researchers, medical professionals, and organizations like the Centers for Disease Control a sort of early warning system that could be applicable to all kinds of disease outbreaks." In addition to improving immediate response to disease outbreaks, the data can also be mined to help answer fundamental questions, such as how large-scale epidemics emerge from low-level interactions between people in the course of their everyday lives.

Most previous work in computational epidemiology focuses on "simulated populations and hypothetical scenarios," Kautz and Sadilek note. Instead, for their flu study, they used a Twitter search application to collect 16 million "real time" tweets from 630,000 different users in the New York City area during a single month. They zeroed in on the tweets of 6,237 individuals who posted more than 100 GPS-tagged tweets during the study period.

The researchers developed statistical natural-language processing algorithms that identified 2,047 tweets reporting flu-like symptoms. Locations were mapped, other Twitter users who visited the same locations were identified, and probabilistic models were then constructed to predict if and when an individual would fall ill.

Kautz is looking at other areas where the application of Big Data methods to Twitter could bear fruit. Would it be possible, for example, to estimate people's emotional states from their tweets? When depressed people are tweeting with other people who either are or are not depressed, what is the affect—is depression contagious?

Big Data mining of social networks is sometimes equated with Big Brother–like invasions of privacy. But there's an important distinction to be drawn, Kautz says.

"Dozens of companies are data mining your social media in order to try to sell you things," Kautz says. "This may or may not be a good thing. By contrast, our goal of improving national and global health is clearly a benefit to society."

Kautz also notes that his project only makes use of data that Twitter users have explicitly made public—no private messages or other information is involved. "Much of the more commercial work—including what is going on at Facebook and Google—make use of your private data, such as your web search topics, your history of purchases, and the message and pages you only share with friends. Such use or misuse of non-public information is the serious threat to privacy."

The approach has the potential to dwarf previous methods for health monitoring in scalability and immediacy. The project has expanded to include social network data gathered from cities and airports around the world. "Today, it can take months to figure out where in the world a disease outbreak originated—and meanwhile people from that area will be carrying the disease all around the world," Kautz says.

Huaxia Rui

Big Data at the Box Office

"Back at work and recovering from #avatar—fantastic movie!"

"Wow! I wanna see 'the lovely bones'!"

Two types of tweets about two different movies. But from a business standpoint, which type of tweet carries more weight in affecting a product's sales revenues?

Probably the second one, according to Huaxia Rui, assistant professor of computers and information systems at the Simon School of Business.

Rui and two fellow researchers analyzed the impact of four million tweets on box office sales for 63 movies. So-called "intention tweets" from people who hadn't even seen the movies appeared to have a greater effect than "positive tweets" from people who had actually seen them.

The results of their study suggest that online chatter really does matter in affecting sales. Business managers could glean important clues about the popularity of their products and even forecast future sales through careful analysis of Twitter traffic.

Word of mouth has always been regarded as a major influence on whether a product will be a big seller or not, Rui says. With the advent of Twitter and other social media networks, huge numbers of word of mouth messages are easily accessible to researchers and business analysts alike, providing "real time" indications of consumer preferences and reactions. Twitter is especially fruitful because it allows researchers to extract the number of followers each author has.

From June 2009 to February 2010, Rui and his colleagues used a computer program to query Twitter every hour for messages mentioning 63 different movies. They filtered out institutional messages; they used machine learning algorithms to classify tweets into one of four categories: those that showed an author's intention to see a movie and ones that gave a positive, negative, or neutral opinion about a movie already seen.

They then used a dynamic panel data model to measure the effect of this word of mouth traffic on weekly box office sales.

They found that the more chatter there is about a movie, the higher its sales revenues, especially when there is a relatively high ratio of tweets from authors with a high number of followers (for example, 400 or more).

Not surprisingly, positive tweets boosted revenues, and negative tweets decreased them—casting doubt on the old saying that "any publicity is good publicity"!

Rui and two fellow researchers analyzed the impact of four million tweets on box office sales for 63 movies. So-called "intention tweets" from people who hadn't even seen the movies appeared to have a greater effect than "positive tweets" from people who had actually seen them.

The most surprising finding, at least on the surface, is that "intention" tweets from people who had not yet seen a movie appeared to have an even stronger impact on revenues than positive tweets from people who had actually seen one.

Rui suggests this is because of the dual effect of intention tweets: These tweets are a clear indication that their authors intend to see a particular movie, and the tweets not only make their followers aware of the movie but possibly influence them to see it as well. That makes "intention tweets" far more valuable in attempting to forecast future sales.

How might a savvy businessman use this kind of information?

Imagine you're the manager of a retail store, and it is two weeks before Black Friday. If you are scanning Twitter and detect a surge in "intention" tweets showing an interest in one of your products, "That could be useful for determining your staffing and inventory," Rui notes.

With more people using smart phones, tweets even reveal geographic location, which could narrow such staffing and inventory decisions to single regions—even individual stores.

"It sounds futuristic because nobody has done this. But I think it could be useful in the future," Rui says.

Consumers could benefit as well.

Businesses that monitor social media, for example, will likely be more responsive in addressing complaints reflected in negative tweeting, precisely because the tweets will be visible to so many other potential customers.

Rui is actually working on a system called twittersensor to give consumers even more power to check how well companies are treating their customers based on people's discussions on Twitter. And customers may be less likely to find long lines or empty shelves on Black Friday if their local stores have done their Twitter "homework" in advance.

Vincent Silenzio

Social Media Offers Help for the Hopeless

"Just can't go on. Think I'm going crazy. Don't see how I'll still be here in a week or a month or a year."

"Call the Lifeline. Don't be shy. They can really help!"

These two lines from TrevorSpace, an online network that targets youth at high risk for suicide, capture the problem—and a potential remedy—at the heart of Vincent Silenzio's research into using social media as a means of suicide prevention.

The first line is from a young person self-identified as LGB (lesbian, gay, or bisexual). Because of the social stigma they encounter, LGB young people 16 to 24 years old are three to four times more likely than other young people their age to attempt suicide or give serious thought to it, says Silenzio, associate professor of psychiatry, family medicine, and public health sciences.

Their isolation in society means they rely heavily on social media to establish networks of friends.

Silenzio sees in this a great potential to counteract the "toxic influences" that lead LGB young people to the depths of despair. He notes, "Their high rate of Internet use suggests that online social networks offer a novel opportunity to reach them"—to answer their despair with hope, for example, even if it is with a simple message of encouragement like the second line above.

Studies by Silenzio and other researchers show that large segments of "hidden" populations—including drug users and prostitutes—could be reached online by "peer-driven" messages. A public health message could be sent to a relatively small number of members recruited from that population, and the online network's own connectivity—through "respondent-driven diffusion"—could take over to spread the word far and near with laser-like precision to the people it most needs to reach.

The key is understanding exactly how the "topology and features" of social networks can be used to expand the reach of positive interventions—and block the transmission of "toxic influences."

That is where Big Data comes into play.

In one study in 2008, Silenzio and four other researchers used an automated data collection program—a "web crawl"—to examine people's publicly accessible Myspace data to find those who openly identified themselves as LGB. The crawl was then extended to those individuals' online friends—and the friends of those friends—until a network of 100,000 LGB individuals had been mapped.

A series of Monte Carlo simulations was run using computational methods to replicate what would happen if a real message were actually sent to members of the network. A variety of starting points was used. For example, what if various combinations of five, 10, or 15 randomly selected individuals were chosen to start the chain? What if they were given five or 10 coupons or alternative incentives to spread the message to other members of the network? The simulations showed that as many as 18,409 individuals could be reached.

One of the key findings: What matters most is not how many individuals the process starts with but the number of peers they recruit along the way. Increasing the number of coupons from five to 10, for example, caused a far more "dramatic increase" in the final sample size reached, compared to doubling or even tripling the number of initial participants.

The key is understanding exactly how the "topology and features" of social networks can be used to expand the reach of positive interventions—and block the transmission of "toxic influences."

Another of Silenzio's goals is to develop social-enabled computer applications that could be put in the hands of teachers, clergy, and others who have close contact with LGB young people.

The applications would instruct them in the kinds of messages and support that can help point a distressed youth away from thoughts of suicide and would provide tools to disseminate these types of messages through social networks.

Ultimately, Silenzio hopes, his work in this area will become "superfluous" because society will have become more accepting of LGB young people and because stronger networks of support will be available to them.

At that point researchers will not have to devise ways to reach what is now a "hidden" population, with peer-driven messages launched online.

Radiology of the Future

Axel Wismueller

When people talk about Twitter as a source of Big Data, Axel Wismueller simply smiles.

Those 400 million tweets a day (as of June 2012) represent about 56 gigabytes of information.

By comparison, the daily production of biomedical images by a single radiology practice is approximately a terabyte, Wismueller says. (One terabyte equals 1,000 gigabytes.) And there are hundreds of radiology practices.

"So this is real Big Data," observes Wismueller, associate professor of imaging sciences, biomedical engineering, and electrical and computer engineering.

This abundance of medical images is already creating bottlenecks in obtaining timely diagnoses.

Wismueller understands this all too well. Not only does he head a biomedical imaging research group at the University of Rochester, he is a practicing diagnostic radiologist at its Medical Center. When he began practicing in the 1990s, Wismueller says, he read 100 to 150 images a day. Now, wielding a mouse at his computer, he scrolls through 30,000 to 50,000.

Looking at such numbers "exceeds information-processing capabilities of the human brain and could eventually make it difficult for us to maintain the highest professional standards that we are committed to provide to our patients," Wismueller says.

And the problem is only going to get worse, he believes. A surge in aging baby boomers is creating additional demand for biomedical images to detect and treat such diseases as Alzheimer's, breast and prostate cancer, osteoarthritis, and osteoporosis. But the number of radiologists is stable or declining.

Professor Wismueller's lab uses advanced pattern recognition and machine learning technology to explore and visualize huge files of medical data in novel ways. "Computational radiology" helps bridge the gap between fundamental engineering research and clinical practice.

"The only remedy is computer-aided analysis of biomedical images," Wismueller says. Indeed, he is convinced that computer-aided analysis will transform the field within 10 to 20 years.

Imagine, for example, that all those images being taken every day, along with all the related health reports, lab tests, and diagnoses, could be stored in an accessible database.

A radiologist confronted with a hard-to-diagnose condition—one of the interstitial lung diseases, for example—could use computational methods to "mine" the database for similar-looking images with similar associated test results for which specific diagnoses had already been determined.

It would help the radiologist narrow the range of possibilities and arrive at a speedier diagnosis for his own patient.

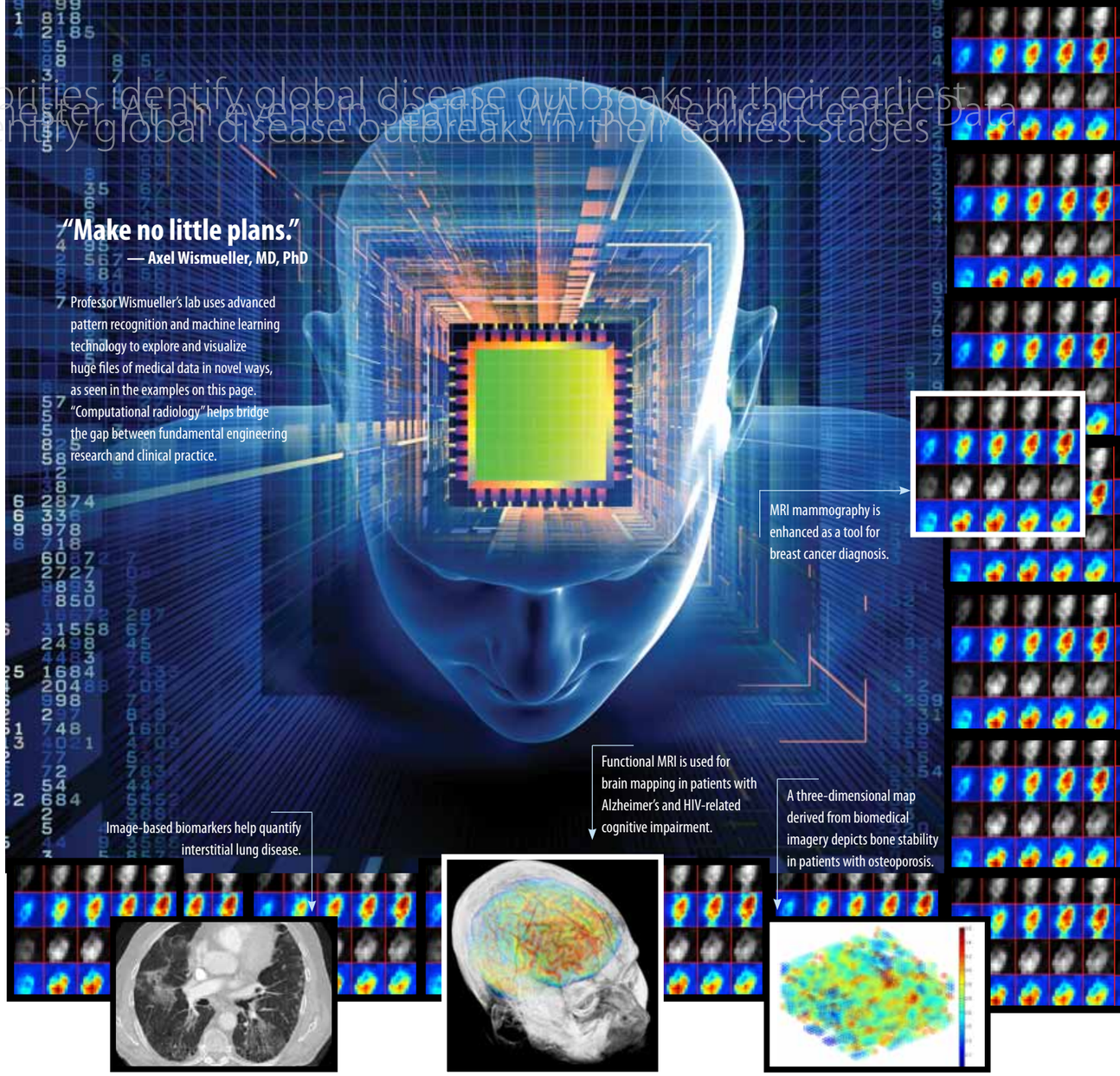
However, several hurdles must be overcome before computer-aided analysis of biomedical images becomes reality. Computer scientists face unresolved questions about how to extract, characterize, and classify all that data from biomedical images. Any subsequent changes in radiology procedures or patient care would have to prove their feasibility in carefully orchestrated, highly regulated clinical settings. And even then, there are the hurdles of government approvals and acceptance by health insurance companies.

Wismueller is not daunted by this. He quotes Chicago architect Daniel Burnham: "Make big plans; aim high in hope and work"—and keeps pushing the frontiers of engineering and medicine, seeking pioneering advances in "computational radiology." His team is applying computational methods in ways that could make it easier to visualize huge files of data and in ways that could significantly advance the diagnosis and treatment of multiple sclerosis, interstitial lung disease, breast cancer, Alzheimer's disease, HIV, and osteoporosis.

For example, Wismueller, with a grant jointly funded by NIH and the German government, is developing a computational framework that uses resting-state functional MR images to examine brain connectivity—how different parts of the brain "talk" to each other. The framework will be used to investigate how brain connectivity changes when antiretroviral therapy is begun for patients who suffer cognitive impairment as a result of HIV.

Another example: Wismueller and his team are applying methods that astrophysicists use to describe the distribution of galaxies to a much more confined space—the mesh-like trebacular structures inside bones. He and the team are converting biomedical images of bone tissue affected by osteoporosis into three-dimensional maps that can identify likely fracture sites and the amount of load these sites can bear. This could help monitor the progression of the disease in a patient and whether the fracture risk is sufficient to require more costly treatments.

Wismueller is convinced that, even as the tools of Big Data transform radiology, "the radiologist will remain at the center of what we are doing just as the patient will remain at the center of why we are doing it."



Genome diagram (opposite page)

Biostatisticians Tame the Human Side of Big Data

The Department of Biostatistics and Computational Biology plays an essential role in Medical Center research:

- from clinical trials of new therapies for neurological and heart diseases to epidemiological studies tracing the effects of mercury exposure on child development
- from predicting suicide risk among veterans to studying the dynamics of complex cellular systems
- in examining the influence of genomic, molecular, imaging, and environmental information on these and other aspects of human health through statistical modeling.

In all of these areas and more, the department's faculty members provide expertise in study design and statistical analysis.

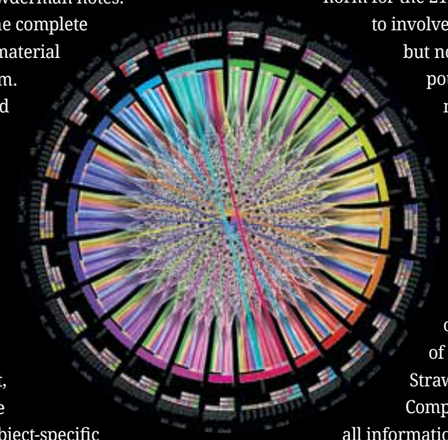
Beyond the support this department provides to other researchers, its members also initiate their own research related to the development of novel statistical methods. Increasingly, in all of these efforts, the department is confronted by the challenges—and opportunities—of Big Data, says Robert Strawderman, who became chair in July.

The department, which has 29 faculty members, boasts a long record of methodological and collaborative research and of educating professionals in the use of statistics. When researchers want to assess the benefits of new therapies, for example, "we're the ones who try to design the studies in such a way that allows an unbiased comparison," Strawderman says.

When there are problems with the data—when subjects drop out of a study in midstream, for example—"we have to account for the potential impact of those types of events on the inferences you want to make. And then there's the actual statistical analysis and reporting of the results. So we are directly involved in all phases of data collection and analysis," Strawderman adds. What complicates that work is the sheer volume of data now accessible to researchers.

For example:

"There's a lot going on in the department dealing with genome data in one way or another," Strawderman notes. A genome is the complete set of genetic material for an organism. In personalized medicine, for example, the design of targeted therapies relies on biomarker information derived from the genome of each patient, hence, on huge amounts of subject-specific "high dimensional" data. "That's the human side of Big Data," Strawderman observes.



For another example, consider the recent focus on health care reform—on moving to evaluation-based health care models. "This is another area where there are potentially massive amounts of data to cope with from various sources," Strawderman notes. "There are government and clinical database sources. And genomic, imaging, and cytometric data are increasingly available for individual patients. Integrating these data sources, each containing potentially large amounts of data per subject, is part of the Big Data challenge."

Statistics developed as a science in the 20th century, focused on answering well-defined questions about a specified population using a modest amount of information sampled from many study subjects. Strawderman characterizes such data as having "a lot of rows with relatively few columns." The norm for the 21st century continues to involve many study subjects, but now each subject has the potential to contribute massive amounts of data, i.e., many more columns than rows. "Such large-scale datasets and associated questions of statistical inference lie beyond the scope of standard methods of analysis," comments Strawderman.

Complicating matters, not all information collected might be relevant, the manner by which data are obtained may be inherently biased, and patterns detected using automated methods are easily distorted by hidden factors not known to the analyst. Strawderman says, "Biostatisticians try to figure out how to formulate the relevant scientific questions and process this information in a way that continues to make some sense."

The norm for the 21st century continues to involve many study subjects, but now each subject has the potential to contribute massive amounts of data, i.e., many more columns than rows.

The department includes two related centers and a division.

- **The Center for Integrative Bioinformatics and Experimental Mathematics, with more than 30 members, is an interdisciplinary research group focused on providing bioinformatics and computational biology support for research in immunology and infectious diseases.**
- **The Center for Biodefense Immune Modeling, with 17 members, is developing models of the immune response to influenza A infection, a potential bioterrorism agent and emerging pathogen. It is also modeling immune responses to influenza vaccinations.**
- **The Division of Psychiatric Statistics, with 13 members, supports innovative research collaborations and coordinated data gathering for studies of human behavior.**

Department, center, and division faculty develop state-of-the-art methods and computational tools to query and analyze the increasingly complex data generated by URMC researchers.

Brendan Mort

BLUE GENE Q

Deep within an unassuming-looking building just off campus lies the Research Data Center—the brain that supports Big Data at the University of Rochester.

The center's computing capabilities are mind boggling. The systems housed here have an aggregate computational capacity of 240 teraflops or 240 trillion calculations a second. That's the equivalent of more than 20,000 laptop computers.

The research facilitated by all this computing muscle ranges from the study of young stars in distant galaxies to simulations of the human heart.

But Big Data wouldn't be such a big deal at the University if somebody didn't maintain those big machines and—even more important—provide the training and support to the researchers who need access to them.

That's where the Center for Integrated Research Computing (CIRC) comes in.

CIRC supports 550 users with individual and group training, and it hosts monthly symposiums where researchers can showcase their work, learn about emerging computing technologies, and participate in collaborative discussions.

Those users, by the way, include faculty members, postdoctoral scholars, research staff, graduate students, and undergraduates from more than 35 River Campus and Medical Center departments and centers. "We support the computational needs of researchers across the University community, including high-performance computing and big data applications," says director Brendan Mort, who heads a staff of six.

Most of those staff members, he adds, are located in Taylor Hall. "We are located with researchers on campus—rather than with the machines at the RDC—to emphasize the importance of collaborating with faculty, students, and research staff and providing them the assistance they need in using the systems," Mort says.

The Research Data Center began to coalesce in 2005 out of recognition of a need for shared computing resources. A faculty team of 17 researchers recommended not only the purchase of a large high-computing cluster but also formation of a center to provide support and training. What was then called the Center for Research Computing received joint funding from the College and the Medical Center in 2007 and was officially launched the following year after purchase of the Linux cluster.

Later that year, IBM donated a Blue Gene/P supercomputer as part of its partnership with the University in establishing the Health Sciences Center for Computational Innovation. The success of that center led, in turn, to purchase of the Blue Gene/Q in 2012 with state funding awarded at the recommendation of the Finger Lakes Regional Economic Development Council.

Now that all that computational power is in place, the next priority is people power. "We need more translators—computational scientists—to work with the scientists in the labs to do the high-performance computing that's needed to analyze huge amounts of data," Mort says.

"There's an untapped world out there. We've got all this data and really fast machines, and now we are looking for the people who understand both the technology and the science well enough to sit between the computer and the lab scientist to help make the connections, so that the research gets done."

CIRC maintains three major computing systems

- **the flagship, a Blue Gene/Q supercomputer that accounts for 209 of those teraflops of computing muscle**
- **an NX cluster that allows researchers to use high-performance computing and Big Data applications remotely in their labs with the ease and convenience of a desktop environment**
- **a Beowulf-style Linux cluster called "Bluehive"—in acknowledgement of the school's mascot, the yellow and blue yellowjacket wasp called Rocky**

Blue Gene Q Gives the University Computing Muscle

Deep within an unassuming-looking building on Science Park lies the brain that supports Big Data at the University of Rochester.

PRINTED ON RECYCLED PAPER · UNIVERSITY COMMUNICATIONS

0816-5M-313SV