The promise and risks of big data

May 30 2014, by Nikki Comeau

To its proponents, big data offers a big promise: insight into complex—and critically important—questions in health care, science, business and more. But its detractors say it poses big risks for individual privacy. Enter Dal's new Institute for Big Data Analytics, poised to explore this challenging new field of study.

Big data is the buzz term used to describe data sets that are huge, flow fast and often contain different forms of data. Computer scientists and data analysts have come to use the three key "Vs"—volume, velocity and

variety—to identify situations that require [big data](#) strategies. Though key, those three Vs don't cover everything. Big data also has to consider the veracity, volatility and validity of data sets. Needless to say, big data is complex: it's a challenge to collect, manage, store and analyze. But the last V sums it up quite nicely: big data can lead to big, valuable solutions.

A premature baby sleeps in a hospital incubator, monitoring devices set up to track heart rate, blood pressure, body temperature and more. In the past, those vital signs would have been checked at regular intervals—perhaps once an hour—with deviations signaling the need for medication or some other intervention. But what if, instead of just checking a half dozen vital signs once an hour, a computer monitored thousands of readings continuously? And what if the data from dozens of babies were analyzed to find correlations between vital sign shifts and the later development of infections or other health problems?

In the past, analyzing millions—even billions—of bits of data and mining it for these kinds of insights was impossible. It was literally too much information, the interrelationships too complex to unravel. But today, with increased computing strength and complexity, researchers are able to examine what's come to be called "big data," with the possibility of finding valuable insights in that stream of information more and more likely.

In the case of the preemies, for instance, researchers in the Artemis Project at Toronto's Hospital for Sick Children used big data strategies to track babies' vital signs and discovered that changes in a baby's heart rate can indicate infection prior to any other signs or symptoms—an early warning that can have life-saving implications.

Those possible benefits—in [health care](#), science, business and more—are what excites Dalhousie's Stan Matwin, Computer Science professor and Canada Research Chair in Visual Text Analytics. Dr.

Matwin is the director of the Institute for Big Data Analytics at Dal, the first academic research institute of its kind in Canada. Since its official launch last summer, the institute has sealed several research deals with partners locally, nationally and internationally, to study topics ranging from traffic patterns in big cities to targeting search-engine users with ads for a specific online retailer.

As well, the institute has conducted big data workshops for small businesses in Nova Scotia, teaching entrepreneurs the value that may be embedded in the data they can or do collect—everything from cell phone location data to GPS data from moving vehicles.

"We actually think about this data as an asset," explains Dr. Matwin. "What can we do to massage this data, how can we use algorithms on it, how can we extract [knowledge] from it? And, knowledge, as we know, is power."

Big data and health care's big picture

The benefit of tracking and analyzing vital signs in preemies is clear. But are there possibilities for improving the overall delivery of health care by collecting and analyzing even more massive amounts of data? Adrian Levy, department head and district chief of Community Health and Epidemiology in Dalhousie's Faculty of Medicine believes there is. A keen observer of technological advances in medicine and elsewhere, Dr. Levy sees an opportunity to explore big data strategies that could improve overall health care efficiency and delivery.

"Almost half of provincial and territorial budgets in Canada are being consumed by health-care budgets," he explains. "So really, it's among one of the biggest social concerns of any developed country in the world, including here in Nova Scotia and in the Maritimes." It's an area of particular concern for Dr. Levy, as principal investigator of the Canadian Institutes of Health Research-funded Maritime Strategy for Patient-Oriented Research. The strategy is focused on the implementation of innovative medical approaches; delivering high-quality, cost-effective health care; and ensuring patients receive intervention at the right time, leading to better health outcomes.

"As opposed to every other sector in society where we've seen huge productivity gains from improvements in computing speed, health care, up until now, has remained remarkably impervious to the benefits [of the whole IT revolution]," says Dr. Levy. Advances using big data in medicine have been happening, but they tend to be specific to an area of

care or practice—like the preemies example—versus an approach that looks at overall systems and delivery.

Dr. Levy cites challenges like confidentiality issues that make IT integration across the many units in health-care environments difficult, but he still believes there's a role for big data to play. That's why he has been consulting with Dr. Matwin.

"Health care is an excellent source of big data," says Dr. Matwin. "More and more, we see computers infiltrating the health-care world in both the research and the delivery. And not just computers, but different devices that use data in massive amounts, like imaging devices. You have patient data, test data, genetic data. They're coming in totally different forms and just putting them together is a challenge."

How can it all be put together for the benefit of the health-care system? That's the question Dr. Levy and Dr. Matwin are exploring together. Dr. Levy explains, for example, that in some cases, often with patients suffering multiple chronic illnesses, tests can be duplicated. "Our computer systems [that capture data] aren't talking to each other," he says.

Before any type of integration strategy, however, Dr. Levy and Dr. Matwin need to first assess the landscape. They're currently looking at what data sets already exist and how they can best be analyzed and optimized to ultimately reach the goal of better health care in this region.

One project they're poised to launch involves geographic data. Dr. Levy wants to better understand Capital Health District Authority's patients and where they're coming from, since the health authority is the province's main referral centre. The plan is to display the data visually on an interactive map that can be used to better inform policy analysts and decision makers.

Keeping private data private

But while big data collection and analysis may have benefits, confidentiality is a real concern. Will gathering data about preemie babies and infection rates, for instance, put individual children at risk of having their health information tracked and, say, shared with an insurer years in the future so they're denied insurance—or charged more for it?

Dr. Matwin is optimistic that such risks needn't come to pass: he believes that it's possible to collect plenty of data to analyze while at the same time creating security procedures that protect the privacy of those who've provided it. "In every project we do [at the Institute], we think about the privacy issues from the beginning," he says.

It's a concept called "privacy by design," a Canadian idea first proposed by Ontario Privacy Commissioner Ann Cavoukian. It means building systems that accommodate and analyze data with privacy methods already embedded in the original design versus as an afterthought. "If you have a system used to share and publish data information about individuals, and you only start thinking about making this data private by removing identifiable information once you've already built the system, it's too late," says Dr. Matwin.

Existing privacy methods aren't perfect and Dr. Matwin is among several researchers investigating ways to improve information privacy. Adding "noise" to the data—random, irrelevant values—acts as camouflage, and individual data points begin to lose any sense on their own, making it difficult to pull out an individual's data and use it for other purposes. Another method is called anonymization, where an individual data point is made to look like 50 others, 100 others, etc. Dr. Matwin compares it to the scenes in movies where someone escapes into a crowd. "You know, they're looking for you in a busy marketplace and you try to look like everybody else so it's harder to find you."

These two methods, however, require tweaking the data, and some critics argue this degrades its quality. "The dream here is to develop methods that, on the one hand, protect the data and, on the other hand, don't change it at all," says Dr. Matwin.

This magic method, he thinks, is a cryptographic one. "It's like a digital envelope," explains Dr. Matwin. The data's owner would seal an envelope containing raw data and send it through a system that could analyze it without having to actually open it and look inside. The envelope, now containing results, would be sent back to the owner. The method could even combine different sets of data from different owners, which is even harder to accomplish due to the usual legal framework around sharing data sets. This would be particularly beneficial with health data. However, the cryptographic method is still theoretical. Dr. Matwin says we're likely to see significant progress bringing it to the practical level within three to five years.

In the meantime, many citizens are willing to take part in such health-care studies with existing privacy standards in place. "Several focus groups have asked patients about the use of routinely collected administrative health data for patient care, even though they don't stand to benefit," explains Dr. Levy. "Patients want the data to be used. As long as you can assure them that anonymity and confidentiality are protected, people are pleased to see their data being used to improve the system."

Still, that willingness to share data may vary under other circumstances—the collection of data by, say, a retailer or social media company like Facebook or Twitter so they can target consumers with more effective advertising or, more controversially, the collection of national security data with the goal of spotting potential terrorist activity. Are there circumstances in which we should trade some privacy for some other benefit? These are questions Dr. Matwin believes need to be

addressed as big data analytics and technology continue to advance.

"There's a need [for society] to talk about the new deal for data. And it's not something that a bunch of university professors will make happen alone."

Provided by Dalhousie University