Where Big Data Meets Law

Patricia Kosseim, Senior General Counsel, Office of the Privacy Commissioner of Canada

CIAJ Conference, St. John's Newfoundland
"Privacy in the Age of Information"
October 15-17, 2014

Big Data is a transformational, disruptive technology that fundamentally challenges the way we generate knowledge, how we see the world around us and what we understand "truth" to be.

Actually, big data is not a new technology per se, but rather a new technological trend. Simply put, big data allow us to process huge volumes of data across varying sources, using much more powerful algorithms, to identify underlying patterns and correlations that can predict future outcomes.

Big data is poised to change just about every aspect of our lives - our laws and legal traditions being no exception.

An early precursor of how "big data meets law" was a study conducted by researchers at Harvard and Washington Universities in 2002 to compare approaches for predicting the outcomes of 68 pending cases on the US Supreme Court docket that year. They designed a computational forecasting model that analyzed hundreds of earlier Supreme Court decisions and compared results with the qualitative forecasts of 87 legal scholars who had clerked with, or practiced before, the Supreme Court and were intimately familiar with all the arguments. Needless to say, the machine won!

While the making of a great John Grisham novel at the time, big data have since elevated the stakes for both lawyers and judges in much more fundamental ways. I will describe just 5 of them today.

1. **Big data challenge our traditional concept of causation.**

   That two or more factors may be shown to be correlated -- however logical or spurious -- does not prove the validity of the findings, and says nothing about their causal relation.

   A heralded example of the promise of big data was Google Flu Trends. In 2009, Google published findings in the prestigious scientific journal *Nature*, showing how it was able to match certain search terms with the number of influenza cases being reported by the Centers for Disease Control. It mattered not whether people who used these search terms were actually sick, and even less what may have caused the flu; rather, Google's claim was its ability to track the use of these search terms to predict possible disease outbreaks in real-time, ahead of public health surveillance reports.

In 2013, scientists reported that Google was in fact significantly overshooting the actual prevalence of flu, attributing this error to two methodological flaws: first, an over-reliance on big data to the exclusion of traditional forms of data collection and analyses (what some commentators have dubbed "big data hubris"); and second, the ever-changing hidden algorithms used by internal Google engineers to improve its commercial services, thereby jeopardizing the validity, reliability and replicability of its findings.

In the legal context, inferences based on correlations emerging from big data analyses will increasingly make their way into court rooms in the form of "expert evidence". It will be the unenviable role of judges to assess their validity and assign their appropriate weight in a given case.

2. Big data challenge traditional thresholds that shield us from the eyes of the state.

*R*ather than look for the proverbial needle in the haystack, big data allow law enforcement to troll entire haystacks to see whether there are any needles to look for. Big data can reveal potential associations between individuals and "persons of interest" within two, three or more degrees of separation. Big data can detect suspicious activity based on travel patterns, buying behavior or meta-data associated with online communications or search activity. Much of this is intended to be analyzed pre-warrant, before there are reasonable grounds to believe or suspect anything, since the whole purpose of these endeavors is to uncover new leads based on probabilistic inferences.

How then to connect these patterns of activity with the identifying information needed for law enforcement to pursue new leads are the questions currently being debated in Bill C-13.

Weaving a common thread through decades of section 8 jurisprudence, Justice Cromwell, for a unanimous court in *R* v. *Spencer*, re-affirmed that the reasonable expectation of privacy analysis must extend beyond the discrete piece of data being sought, to also consider all the other interconnected data that may potentially be revealed about an individual. In an age of big data, this includes not only that which we can hardly see, but also that which we can barely *imagine*.

3. Big data challenge our traditional regulatory models.

A new concept of algorithmic regulation is emerging as a possible alternative to traditional rule-making. Rather than codify rules that try to anticipate all possible scenarios, are inefficient to enforce and difficult to adapt to evolving reality, big data could allow certain activities to regulate themselves based on dynamic algorithms and real-time feedback loops.

For example, we all know speed limits exist to ensure safety on the roads. Even with road-side radars and cameras, enforcement still requires intervention and deployment of many police officers. Imagine a GPS-enabled system that could automatically report speeding from the vehicle itself and send e-tickets to car owners directly. Better yet, imagine a dynamic algorithm able to detect road conditions and traffic levels, and adjust speeding limits according to the associated level of risk. Rather than focus on the rule itself, "thou shalt not pass 100 km per hour", algorithmic regulation would focus on the desired outcome, "ensure safety on the road", and modulate itself accordingly.

These new applications of big data raise fundamental questions about who gets to set the rules, how to assess the fairness and accuracy of collected data and how to restrict use of information for its intended regulatory purpose. Collection of speedometer data for road safety purposes is one thing, but also tracking location information "just because" is another.

4. Big data challenge our traditional fair information principles.

Given the new big data paradigm, some major commercial players have begun calling for a watering-down of the traditional OECD fair information principles of consent, purpose specification, and collection/use limitation, willing to trade up other principles like accountability instead -- even suggesting the idea of creating internal ethics boards to review new algorithms before market deployment.

In a world where personal information has become the new oil, and where commercial algorithms, like product ingredients or new formulae are preciously guarded as trade secrets, one can't help but be a little skeptical about the proposed trade-off.

Consumer ethics boards, while very laudable in my personal view, might not provide us the requisite level of third party assurance *unless and until* they achieve what research ethics boards have strived for decades to achieve in the scientific world: true independence from the institutions whose practices they are intended to review, and an external governance regime that “watches the watchers”.

Two days ago, in Mauritius, off the coast of Africa, data protection authorities from around the world passed a unanimous resolution, recognizing the benefits of Big Data, but calling on all parties who make use of Big Data to continue to respect key fair information principles. Further, the resolution urges users of big data to demonstrate that decisions around the use of big data are fair, transparent and accountable and that any profiles and algorithms used be continually assessed from an ethical perspective.

5. **Big data challenge our understanding of harm.**

Many recall how Target combined purchase patterns with basic demographic data to identify women likely to be in their second trimester of pregnancy (apparently a gold mine for retailers) and send them personalized ads. Whether Target got it right 100% of the time was less important than getting it right most of the time; in fact the only reason this enhanced marketing practice came to light through a 2012 New York Times article was because Target “got it” right in one infamous case before a teenager’s own father did.

But what if Target or other advertisers get it wrong? Is the practice any less offensive? Not necessarily, according to women who recently reported consternation at receiving baby-related ads, months after they painfully miscarried.

Or according to a recent Harvard study led by Dr. Latanya Sweeney who found racial bias in ads connected with certain search terms used in Google and Reuters. When searching black-identifying first names (such as DeShawn, Darnell and Jermaine), a higher percentage of ads offering services for criminal record checks appeared, than was the case when searching white-identifying names (such as Brad, Jill and Emma).

What about when big data are used not only to sell our identities, but to *shape* our identities? When big data track our friends and activities on social media sites in order to predict our political leanings and unleash last-ditch efforts to influence our vote? (as a data analytics company based in Ottawa was recently contracted to do by the “Yes” camp of the recent Scottish referendum.)

Or when click stream data are used to profile us into certain interest categories and show us tailored versions of the daily news reinforcing initial biases and depriving us of a more complete understanding of the world's events?

Commodifying who we are, *inferring* who we are, or *shaping* who we are seems intuitively at least, to injure our identities and offend our sense of dignity.

While laws recognize non-pecuniary harms resulting from breach of privacy, courts have been hesitant to award damages beyond a perceived notional cap.

Increasingly, however, courts will be asked to infer harm resulting from use of big data, even in the absence of psychological evidence. As we enter the obscure world of hidden algorithms and as individual control over one's personal information approaches N = 0, I suspect that the right to dignity will become more operative in these cases relative to the concept of autonomy. And over time, courts will become more comfortable inferring harm as the Supreme Court did in *A.B.* v *Bragg* using "common sense and logic".

Conclusion

In a seminal Foreign Affairs article last year called "The Rise of Big Data", Kenneth Cukier and Victor Mayer-Schoenberger summarized these and other challenges in a most insightful way:

"Big data is poised to reshape the way we live, work, and think. A worldview built on the importance of causation is being challenged by a preponderance of correlations. The possession of knowledge, which once meant an understanding of the past, is coming to mean an ability to predict the future. The challenges posed by big data will not be easy to resolve. Rather, they are simply the next step in the timeless debate over how to best understand the world."

Thank you.