Office of the
Privacy Commissioner
of Canada

The Age of Predictive Analytics:

From Patterns to Predictions

Report prepared by the Research Group of the Office of the Privacy Commissioner of Canada

August 2012

Table of Contents

Introduction

In both the public and private sectors there is an overall fascination with predicting how people will behave: What will people purchase? How do they use technology? When will someone behave badly, break the law, or commit fraud? danah boyd has noted this shift, saying "it's no longer about what you 'do' but about what you 'might do', and that includes what other's do where it implicates you or might influence you."[1] This type of analysis isn't an entirely new phenomenon, but the form of that analysis is moving from predictions based on experience, intuition and critical thinking, to one that is based on a technological analysis of raw data – predictive analytics.

In a recent article *How Companies Learn Your Secrets,*[2] and book *The Power of Habit: Why we do what we do in life and business*,[3] New York Times reporter Charles Duhigg described how certain corporations use "predictive analytics" to understand consumers' shopping and personal habits in order to market to them more effectively. Specifically, Duhigg shed light on the practices of American department store *Target*, and reported that the company used predictive analytics as a means to uncover which women were likely to be in their early stages of pregnancy so that it could target advertising to them before any other company. The algorithm that was behind the analytics was dubbed the "pregnancy-prediction algorithm."

The development of the pregnancy-prediction algorithm exemplifies the emergence of a corporate trend that places huge importance on internal "Data Analytics Departments" or "Data Science Teams." Duhigg reported that Target has about 50 employees whose sole job is to find trends and patterns hidden in the data that Target collects on its customers.[4] In the development of the pregnancy-prediction algorithm, the team of data scientists tested theories and analyzed patterns in customer data, historical data from baby registries, and demographic data purchased from data brokers, and found that certain patterns and data linkages could be made that revealed predictable shopping patterns with women who were pregnant. According to Duhigg, the driving force behind the development of the algorithm was a theory that people's buying habits are more likely to change when they go through a major life event, such as a pregnancy, and therefore targeting advertising campaigns to such customers could reinforce a habit of shopping at Target stores.

The purpose of this research report is to develop a better understanding of the concept of predictive analytics, which is the underlying process described in Duhigg's article. Predictive analytics is a general purpose analytical process that can be applied in sectors as diverse as retail to boost sales, law enforcement to predict crime, and health programs to monitor for disease outbreaks. In that regard, it is not a straightforward concept to define or describe, and the privacy implications could vary from privacy neutral to privacy invasive, depending on its application. Moreover, it is important to acknowledge that predictive analytics is a process that is closely intertwined with previously known notions of data mining; however the inferences extend beyond retrospective pattern analysis to a result that is more prospective and anticipatory.

[1] danah boyd. (2011). "Networked Privacy." *Personal Democracy Forum.* New York, NY, June 6. [html]

[2] The New York Times. *How Companies Learn your Secrets* by Charles Duhigg, Published: February 2012.

[3] Duhigg, C. 2012. *The Power of Habit: Why we do what we do in life and business.* Doubleday Canada, Random House Canada Ltd.

[4] Duhigg, C. 2012. *The Power of Habit, p. 190*

Much of the research available on predictive analytics and referred to in this paper relates to the U.S. context and the practices of American companies. In the absence of hard data we still do not know precisely to what extent companies in Canada use predictive analytics, or for what purposes. Nevertheless, by looking at the U.S. context and research we can make some inferences about the current or future practices in Canada. By monitoring the trends in predictive analytics, we can move towards a better understanding of how it may be used by companies, organizations or government, and how it might impact individuals in the Canadian context.

In that regard, this paper will be a starting point to explore some of the underlying themes and conceptual frameworks for understanding predictive analytics, and to discuss a few of the various applications of this technology in the private and public sectors.

This paper will:

- examine the concept of predictive analytics;
- provide an overview of the context for predictive analytics;
- identify some of the applications in the private and public sectors;
- outline some of the broader privacy implications it can raise for individuals and for society as a whole; and
- examine predictive analytics in relation to privacy principles in *PIPEDA* and the *Privacy Act*.

The Concept: "Predictive Analytics"

The first question many people ask is how does *predictive analytics* differ from the *data-mining* that governments and companies have been doing for some time? Data mining is defined as "the *process* of discovering interesting patterns and knowledge from *large* amounts of data."[5] Both data mining and predictive analytics are processes that apply sophisticated mathematics and statistical analyses to data in an attempt to discover knowledge and patterns. Although they are related concepts, perhaps even synonymous in terms of process, predictive analytics gives us new clues as to how data mining practices are advancing and becoming increasingly intelligent.

Predictive analytics marks a progression from simply identifying patterns to making predictions based on patterns. Computerworld (2006) defines *predictive analytics* as "the branch of data mining concerned with forecasting probabilities."[6] From this definition we see that predictive analytics is a concept that is more uniquely forward-looking**,** and when personal information is the raw data, predictive analytics is the process attempting to forecast our future behaviours or intentions. SAS, one of the world's largest business analytics companies, says predictive analytics is about "revealing previously unseen patterns, *sentiments* and relationships [emphasis added]."[7] So where data mining describes the exploratory process of finding patterns and knowledge within data, predictive analytics then attempts to leverage that knowledge derived from data to anticipate meaning and make predictions about the future.

[5] Han, J., Kamber, M., Pei, J. 2011. *Data Mining Concepts and Techniques* (Third ed). Elsevier Inc.: p.6 and 8

[6] Computerworld. 2006. QuickStudy: Predictive Analytics. By Jan Matlis, published October 9, 2006. http://www.computerworld.com/s/article/267042/Predictive_Analytics (accessed online April 15, 2012)

[7] http://www.sas.com/technologies/analytics/datamining/

Predictive analytics is a general purpose analytical process that enables organizations to identify patterns in data that can be used to make predictions of various outcomes, not all of which have an impact on individuals. Software products are now becoming more readily available to companies to implement analytics into their business models.[8] They can be used by organizations to avoid risk, make unprofitable customers more profitable, retain profitable customers, reduce business expenses, identify fraud, avoid process failures, or even to analyze the effects of health treatments.[9] Emerging capabilities include real-time analytics and the analysis of unstructured information such as text.[10] The key point to note is that the types of decisions that can be made based on the outcomes of analytics are always advancing into new realms. The trend is a shift away from data mining that presents findings that are mere aggregations or characterizations of patterns in data;[11] predictive analytics is characterized by its ability or attempt to forecast, anticipate, or infer.

The Context for Predictive Analytics

Technological innovation and the shifting nature of consumption on the Internet have played an important role in the emergence of predictive analytics as a tool for business and governments. The devices we use and the convergence of different technologies have prompted new channels and sources of data, leading to the proliferation of data at exponential rates. In recent years we have seen the term “Big Data” emerge to describe this trove of information that is garnered from people’s everyday activities, things we consume, and our interactions with other people and objects. This Big Data has in turn become increasingly valuable to organizations and making predictions has become part of the formula for success. The following section will capture some of the essential elements of the overall context and describe how technology plays a role as a catalyst for predictive analytics, how Big data is the essential ingredient, and how the desire to achieve success through data-driven decision-making is setting this trend in motion.

The catalyst*: the platforms and incentives make all of us the product*

The online environment is at the root of the proliferation of data and ultimately, the emergence of tools to analyze it. Most of our online activities, such as the use of social media, applications, companies requiring us to register and create shopping profiles, etc., all prompt or persuade us to reveal information about ourselves. Sometimes we do not have control over what is requested, for example when we must fill out required fields in order to activate a service or purchase something online. Other times people share personal information willingly in a social manner or in exchange for a benefit or perceived benefit. The Internet of free platforms, free services and free content is laced with incentives, rewards and benefits to participation, whether it be for the convenience of the platforms, the enjoyment we get from being social, or the lure of deals on products we are interested in. The platforms and incentives are all designed specifically to encourage individuals to offer

[8] For example, IBM “Analytical Decision Management Tool”, and similar predictive analytics products provided by SAS, Oracle, etc.

[9] SAS 2012. *Drive Your Business with Predictive Analytics*.

[10] Schwartz, Paul.M (2010) *Data Protection Law and the Ethical Use of Analytics*. The Centre for Information Policy Leadership Hunton & Williams LLP

[11] Millar, J. (2009). “Core Privacy: A Problem for Predictive Data Mining”, in Lessons from the Identity Trail: Anonymity, Privacy and Identity in a Networked Society, Ian Kerr, Val Steeves, Carole Lucock (Eds.), Oxford University Press. 103 – 119.

up information about themselves,[12] and as aptly put by Professor Zittrain in 2011, "If what you are getting online is for free, you are not the customer, you are the product."[13]

As consumers in particular, a great deal of personal information can be extracted from our consumption patterns and online activities. Social media companies and other corporations are shifting the behavior of consumption on the Internet and linking people and objects together through our social interactions and through various different technologies: "the use and convergence of the web, mobile phones, electronic financial systems, biometric identification systems, RFIDs, GPS, ambient intelligence, and so forth, all participate in the automatic generation of data which become available for still more pervasive and powerful data mining and tracking systems."[14]

These technologies can all give companies insights about what is happening now, or what will happen soon; "faster than real-time" is the new goal for digital products.[15] Organizations are all vying to get access to unique information and that encourages each company to seek new avenues for data collection."[16] The real-time transactional customer data and the integration of multiple online platforms and technologies will only make the descriptions of patterns of behaviour and predictions of future behaviour increasingly accurate and meaningful.[17]

The ingredients*: our data crumbs*

Our contemporary society is a data driven world. The concept of Big Data has been around for some time, but the scale of the concept is rapidly escalating. It is not only about the amount of data, but also how data are increasingly interrelated.

Big Data is a concept closely intertwined with predictive analytics because the data points are the ingredients that feed the application of predictive algorithms. In this contemporary and tech-driven society almost everything we do produces a stream of personal information.[18] Between the user-generated content that is offered up by individuals, and the personal information that companies extract from our consumption activities, it is becoming increasingly possible to capture a fairly detailed picture of how we organize and go about our daily lives: "We have gone from having little bits and pieces about us stored in lots of different places off- and online, to having fully formed pictures of who we are. All digitally captured and stored."[19] It is becoming harder to maintain distinctions between our offline self and our digital activities, particularly as we interact with handheld mobile technologies.

[12] From *The Atlantic*. It's Not All About You: What Privacy Advocates Don't Get about Data Tracking on the Web. By Alexander Furnas, Published March 15, 2012. http://www.theatlantic.com/technology/archive/2012/03/its-not-all-about-you-what-privacy-advocates-dont-get-about-data-tracking-on-the-web/254533/ (accessed online April 11, 2012)

[13] Quoting Jonathan Zittrain last summer in http://news.harvard.edu/gazette/story/2011/06/hyper-public-spaces/

[14] Gutwirth, S. and Hildebrandt, M. 2010. 'Some Caveats on Profiling'. In *Data Protection in a Profiled World*. S. Gutwirth et al. (eds.), Springer Science & Business Media B.V. 2010.

[15] CNN Tech. *"Should we fear mind-reading future tech?"* By Andrew Keen, posted June 19, 2012.

[16] Schwartz, Paul.M (2010) *Data Protection Law and the Ethical Use of Analytics*. The Centre for Information Policy Leadership Hunton & Williams LLP

[17] Pridmore and Zwick, 2011, p. 271

[18] The Vancouver Sun. *The erosion of anonymity: Today's digital world forces us to share more of our personal information.* By Misty Harris, Published April 2, 2012. http://www.vancouversun.com/technology/erosion+anonymity/6396220/story.html

[19] Terrence Craig and Mary E. Ludloff. 2011. *Privacy and Big Data.* O'Reilly Media Inc. p.5.

With the scale of Big Data ever-increasing, so too is its value. Some have argued that "data is the new oil,"[20] meaning the new commodity to be refined (analyzed) and then exploited. The explosion of consumer data has created an entire industry that purports to draw meaning from that data. Companies are resorting to collecting vast amounts of data on consumers in order to gain a competitive advantage.[21] For many, the only way to survive in the global economy is to "embrace and leverage the power of information."[22] The exploding data deluge, comprised of our personal information as the raw data, and the increasing capacities for storage and information technologies, are contributing to this rise in the use of analytics.[23] The true value comes from measuring our behavior and inclinations in fine detail and as a basis to make predictions about future events.

The impetus*: to know something before it happens*

The desire to derive meaning from data is what motivates a great deal of the collection, storage, and dissemination of information, and why there is a push to find more and more advanced techniques for information analysis.[24] It is now about learning whatever we can about people, their attributes, and past actions in an effort to understand their predispositions and predict future actions.[25] Helen Nissenbaum, in her examination of pivotal technological transformations, describes this tendency as an "unbounded confidence placed in the potential of information processes and analysis to solve deep and urgent social problems" and explains that "the confidence fuels an energetic quest both for information and for increasingly sophisticated tools of analysis."[26] Nissenbaum reminds us that this trend is not limited to the context of commercial interests, but that the appeal of analytical methods and tools such as predictive analytics will be sought after by a variety of decision-makers in different sectors, for example, in the financial sector, insurance and credit-card companies, health/hospitals, national security and law enforcement.[27] Depending on the application, the combination of big data coupled with the desire to predict and the intelligent capacity of predictive algorithms could be what truly magnifies the implications for privacy.

[20] see this argument developed in: http://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/

[21] Pridmore, Jason and Detlev Zwick. 2011. Editorial: Marketing and the Rise of Commercial Consumer Surveillance. *Surveillance & Society* 8(3): 269-277.
http://www.surveillance-and-society.org

[22] SAS 2012. *Drive your Business with Predictive Analytics.*

[23] Schwartz, Paul.M (2010) *Data Protection Law and the Ethical Use of Analytics*. The Centre for Information Policy Leadership Hunton & Williams LLP

[24] Nissenbaum, Helen. 2009. *Privacy In Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

[25] Nissenbaum, Helen. 2009. *Privacy In Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press. p. 44

[26] Ibid, p 42.

[27] Ibid, p.45

The Applications: Who is trying to predict what?

The growing appetite and emphasis on using data to drive decision-making can have several influences; sometimes it's about anticipating outcomes that will generate more profits, and sometimes about managing risk or preventing negative outcomes. In a forthcoming book chapter, Ian Kerr characterizes different purposes for utilizing predictive technologies and analytics.[28] One type of prediction he calls "*preferential predictions*", which are an attempt to anticipate individual preferences or inclinations, often for the purpose of tailoring offerings of products and services. Another type is "*preemptive predictions*", which are an attempt to anticipate and prevent certain types of actions that are likely to generate social risk.[29] These two concepts can provide a basis for understanding some of the outcomes being sought by different applications of predictive analytics. The following section will offer some examples from both the private and public sectors to illustrate the desired outcomes of *preferential* and *preemptive* predictions.

Targeted Advertising

As we saw with the Target news story, the application of predictive analytics can help companies be more effective in targeting advertising with the view to increasing profits. Companies want to be able to infer their customers' preferences, identify prospective or profitable customers, and to target appropriate products and services at precisely the right moment.[30] Terry O'Reilly calls this activity *Hyper-Targeting*, and says that this practice will allow companies to send perfectly-tailored advertisements directly to individuals, based on deep knowledge of that individual's personal life, at the exact moment they are about to buy something.[31] The advertising context is a good example of how the potential for large profits can shape how companies' value and use personal information and what motivates them to use predictive analysis to better understand who we are, what we want, and when we want it, in real-time. Predictive analytics will enable companies to effectively perform this type of targeted advertising, and is a clear illustration of how a preferential prediction could be a lucrative purpose for using predictive analytics in the private sector.

Social science by social media

Predictive analytics is becoming a tool that helps draw valuable insights from collective user-generated and unstructured data that reveals things about human behaviour and communication patterns, sentiment analysis, and patterns of social influence. "Google searches, Facebook posts and Twitter messages, for example, make it possible to measure behaviour and sentiment in fine detail and as it happens."[32]

[28] Kerr, Ian (preprint) *Prediction, Preemption, Presumption: The Path of Law After the Computational Turn** forthcoming in Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology, eds. Mireille Hildebrandt & Ekaterina De Vries.

[29] Kerr, Ian (preprint) *Prediction, Preemption, Presumption: The Path of Law After the Computational Turn** forthcoming in Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology, eds. Mireille Hildebrandt & Ekaterina De Vries.

[30] Schwartz, Paul.M (2010) *Data Protection Law and the Ethical Use of Analytics*. The Centre for Information Policy Leadership Hunton & Williams LLP

[31] CBC Radio. *Under the Influence with Terry O'Reilly* on *Hyper-Targeting*. Aired April 28, 2012. http://www.cbc.ca/undertheinfluence/season-1/2012/04/28/hyper-targeting-1/

[32] From the NYT. Big Data's Impact in the World. By Steve Lohr. Published February 11, 2012. http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?_r=2&pagewanted=all

Facebook has the most extensive data set ever assembled on human social behaviour and its team of data scientists are looking for innovative ways to mine the troves of data for insights into human communication and social behaviour.[33] Since Facebook collects data from users interacting in real time, its data science team is in a unique position to be able to experiment and analyze patterns and motivations behind human social behaviour, preferences, and interactions. For example, its data can give insights into why and the extent to which ideas or fashions spread between people or the extent to which a person's future actions are influenced by communications with their friends. In real-time, Facebook could track a social trend or calculate a country's "gross national happiness" by analyzing key words and phrases that signal positive or negative emotions towards various things.[34] It could prove incredibly lucrative for Facebook to sell 'insights' mined from its data to companies who want to know how to induce people to share content or click on ads, or to economists and other researchers studying human social behaviour.

However, generating revenue is not the goal of all organizations interested in this type of application. A new initiative by the United Nations called "Global Pulse" is seeking to conduct sentiment analysis of messages in social networks and text messages to help predict job losses, spending reductions or disease outbreaks in a given region. The goal for the UN is to use early-warning signals and then direct assistance programs in advance of problems, such as preventing a region from slipping back into poverty.[35]

Law enforcement and intelligence

Law enforcement and intelligence agencies have for a long time been using data mining and profiling techniques to predict or identify potential threats or criminal activity. In a society that is increasingly preoccupied with risks and threats,[36] there is a continuous concern that anyone can be the "bad man" and enthusiasm for predictive technologies that preempt or prevent conduct that is perceived to generate social risk.[37] Predictive analytics products are becoming more popular amongst law enforcement agencies, and are already being implemented in the U.S. to help law enforcement forecast "hot spots" based on times and locations of previous crimes, combined with incident records, and historical and sociological information about criminal behavior and patterns.[38] These pre-crime detection technologies continue to be developed and tested, some already claim that they can predict when crimes will be committed and who will commit them, before they actually happen.[39] IBM's analytics tool touts that predictive analytics will help police move from "sense and respond" to "predict and act."[40] Other programs seek to analyze behaviour and attribute

[33] Technology Review: What Facebook Knows. *The Planetary Experiment.* From the Editor. Published August 2012. p. 12

[34] Ibid.

[35]From the NYT. Big Data's Impact in the World. By Steve Lohr. Published February 11, 2012. http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?_r=2&pagewanted=all

[36] A term coined by Ulrich Beck, 1999.

[37] Kerr, Ian (preprint) *Prediction, Preemption, Presumption: The Path of Law After the Computational Turn** forthcoming in Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology, eds. Mireille Hildebrandt & Ekaterina De Vries.

[38] Such as IBM Blue Crush (Memphis), Watson (Washington), http://www.cnn.com/2012/07/09/tech/innovation/police-tech/index.html

[39] Developed by Richard Berk, a professor at the University of Pennsylvania, the software is already used in Baltimore and Philadelphia to predict which individuals on probation or parole are most likely to murder and to be murdered. http://abcnews.go.com/Technology/software-predicts-criminal-behavior/story?id=11448231

patterns that are associated with criminal or terrorist activity.[41] Mandates for public safety and national security are commonly preoccupied with predicting which individuals are most likely to be a terrorist or commit crimes. With this as a goal, the heightened interest in preemptive type predictions only continues to grow.

Location tracking

Mobile smartphones and tablets are making it possible and popular for someone to "geo-tag" their location in real-time. Mobile apps and online services are not only encouraging and facilitating this trend, but there is now evidence from a recent study that suggests predicting someone's future location by tracking mobile phone usage is proving quite accurate and effective.[42] In this study, an algorithm was able to predict a mobile phone user's future GPS coordinates to within approximately 1,000 square metres. When the prediction took into account additional information from a single friend, the future location could be predicted to within 20 metres.[43] Even without geo-tagging, the study also showed that location could be predicted with similar accuracy using the geographic location of cell network towers.[44] The ability to track and predict an individual's movements could be appealing for companies interested in tailoring advertisements based on preferential predictions, or to law enforcement seeking to predict and preempt criminal activity. Companies would be able to predict an individual's movements so it can offer a tailored deal at precisely the right moment. For example, a company using the technology could predict when you go on coffee break and what location you tend to go so that it can offer a special coupon just as you are heading out the door.[45] Alternatively, law enforcement could use the technology to track and predict the location of certain individuals who are suspected criminals. Provided that law enforcement could obtain a warrant to access the GPS location data, the algorithms would enable them to monitor patterns in a suspect's movement and intervene when the algorithm suggests future movement to an unusual area.[46]

Fraud Prevention

There are also areas in which the government or the private sector could utilize predictive analytics for fraud prevention. Government programs concerned with fraudulent transactions or claims for government benefits, insurance and credit reporting, could use pattern and trend analysis aimed at detecting and deterring fraud or

[41] The Department of Homeland Security has a few projects ongoing such as 1) the *Future Attribute Screening Technology* (FAST) mobile module project aims to develop technologies that can screen people for certain behavioral attributes associated with committing violent acts or other crimes; and 2) the *Predictive Screening* project aims to derive observable behaviors that precede a suicide bombing attack and develop extraction algorithms to identify and alert personnel to indicators of suicide bombing behavior.

[42] "Musolesi's research into what he calls mobility patterns, which he recently published as part of his research at the University of Birmingham in the U.K. Recently he won Nokia's Mobile Data challenge by predicting the movements of 25 volunteers working in a town in Switzerland. He used GPS data, telephone numbers and their texting and calling history to do it, and the algorithm was at times able to predict where these volunteers were heading to within 20 square meters. Crucially, the algorithm was only this precise when it also tracked the movements and data of each volunteer's friends." Quoted From: http://www.forbes.com/sites/parmyolson/2012/08/06/algorithm-aims-to-predict-crime-by-tracking-mobile-phones/

[43] http://www.forbes.com/sites/parmyolson/2012/08/06/algorithm-aims-to-predict-crime-by-tracking-mobile-phones/

[44] http://www.forbes.com/sites/parmyolson/2012/08/06/algorithm-aims-to-predict-crime-by-tracking-mobile-phones/

[45] http://www.forbes.com/sites/parmyolson/2012/08/06/algorithm-aims-to-predict-crime-by-tracking-mobile-phones/

[46] Ibid

false claims. For example, Service Canada's Integrity Services Branch is utilising statistical software as part of a Predictive Risk Analysis pilot project designed to detect Employment Insurance (EI) fraud and abuse.[47] The idea is that the predictive risk tool would analyze multiple databases and significantly improve the identification of EI applicants who have been overpaid. Each file that is flagged for review by the system would then be investigated. The program represents a shift to automated fraud detection and general risk management, which is now facilitated through the use of the analytics tools.

The Privacy Implications

It is over simplistic to presume that all or most data analytics is entirely problematic from a privacy point of view.[48] Predictive analytics can take many forms, and the privacy implications will vary according to the context in which it is used, as well as the scope and implementation. Big Data and intelligent predictive analytics could, on the one hand, help advance research, innovation, and new approaches to understanding the world and make important and socially valuable decisions in fields such as public health, economic development and economic forecasting.[49] On the other, advanced analytics prompt increased data collection, sharing and linkages, as well as has the potential to be incredibly invasive and intrusive, discriminatory, and yet another underpinning of a surveillance society. The following section will contemplate some of the broader individual and societal privacy implications that can arise with the implementation of predictive analytics.

Predictive analytics can be "creepy"

While it is not a universal reaction, predictive analytics in certain contexts can prompt a "creepy" or unsettling feeling of being under the gaze of an omniscient observer who knows something about us and our behavior.[50] danah boyd has argued that the collection of data in and of itself is not a violation of privacy, but explains that "piecing it together and using it to "stare" is a serious violation of privacy norms."[51]

Mis-characterizations and inaccuracies are certainly a problematic outcome of analytics; however, accurate predictions could be an even greater invasion of privacy in certain contexts. Millar points out that the outcomes of predictive analytics can reveal things that form part of what he refers to as our *"core private information."* He explains that the use of predictive analytics and data mining can violate core privacy when it reveals an individual's *unexpressed* desires, beliefs or intentions to which only the individual would have first-person access.[52] He further explains that inferences made about us through the use of predictive algorithms, or deep analysis by a team of trained specialists with access to stores of data, could lead us to feel as though

[47] According to a PIA Summary posted by Human Resources and Skills Development Canada (HRSDC) on http://www.hrsdc.gc.ca/eng/access_information/privacy/PhaseII.shtml

[48] Nissenbaum, Helen. 2009. *Privacy In Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.
p.50

[49] From the NYT. Big Data's Impact in the World. By Steve Lohr. Published February 11, 2012. http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?_r=2&pagewanted=all

[50] *The Atlantic*. It's Not All About You: What Privacy Advocates Don't Get about Data Tracking on the Web. By Alexander Furnas, Published March 15, 2012.

[51] danah boyd. "Networked Privacy"

[52] Millar, J. (2009). "Core Privacy: A Problem for Predictive Data Mining", in Lessons from the Identity Trail: Anonymity, Privacy and Identity in a Networked Society, Ian Kerr, Val Steeves, Carole Lucock (Eds.), Oxford University Press. 103 – 119.

our core privacy was violated because they are often inferences that are beyond what could otherwise be easily observed or known by others about us.[53]

In other words, presumptions that are made about people based on activities that are easily observable are not generally alarming to people. On the other hand, presumptions that are derived from a deeper and broader inspection of our activities, or that are done with technical assistance may be unexpected or go beyond our reasonable expectations.

Opaque processes and outcomes

Predictive analytics is usually an opaque process. Even where a company acknowledges that it will use personal information in analytics, for example in a privacy policy, we know that most people do not read or understand the complex and arduous legal language in which they are usually written. A recent poll commissioned by the Office of the Privacy Commissioner of Canada found that only 50% of Canadians who responded to the survey indicated they “rarely” or “never” consult privacy policies, and a majority (62%) of Canadians felt that the privacy policies of Internet sites they visit are somewhat or very vague in terms of giving them the information about what the company will do with their personal information.[54] Moreover, even where people read the fine print of a privacy policy or terms and conditions, human beings in general are not very good at weighing the impact of consequences that could be well into the future, and the risks increase as we disclose more, something that the design of social media conditions us to do.[55]

For the most part, the extraction of personal information happens without consumers knowing exactly how much they have really provided. People usually do not have any choice as to what personal information they have to give up, or knowledge of who has access to it or how it is used.[56] While individuals do have some control at the front end, in terms of what they post and share on the Internet and what transactions they complete, their choice to participate is not usually based on a comprehensive understanding of how their data is being manipulated behind the scenes and beyond the moment of their transaction.

Some researchers refer to this as an ‘a*symmetrical information flow.’*[57] Companies, organizations, and governments are all trying to learn more intimate details about consumers and citizens, their behaviours, habits, intentions, etc, but individuals know proportionally very little about the organizations with whom they interact.[58] The complex processes that underlie predictive analytics or data mining techniques are usually quite lost on many individuals, if not most. This leaves people perplexed and completely unaware of the reasons why they may have been “denied a loan, targeted for a particular political campaign message, or saturated with ads at a particular time and place when they have been revealed to be most vulnerable to marketing appeals.”[59] Even if companies were to include detailed references to their activities, for example in

[53] Millar, J. (2009). “Core Privacy: A Problem for Predictive Data Mining”, in Lessons from the Identity Trail: Anonymity, Privacy and Identity in a Networked Society, Ian Kerr, Val Steeves, Carole Lucock (Eds.), Oxford University Press. 103 – 119.

[54] OPC *Survey of Canadians on Privacy-Related Issues*. Conducted by Phoenix Strategic Perspectives Inc., January 2013. http://www.priv.gc.ca/information/por-rop/2013/por_2013_01_e.pdf

[55] Technology Review: What Facebook Knows. *The Curious Case of Internet Privacy*. By Cory Doctorow. June 6, 2012

[56] Manzerolle, Vincent and Sandra Smeltzer. 2011. Consumer Databases and the Commercial Mediation of Identity: A Medium Theory Analysis. p. 326 and 331.

[57] Manzerolle, Vincent and Sandra Smeltzer. 2011. Consumer Databases and the Commercial Mediation of Identity: A Medium Theory Analysis. *Surveillance & Society* 8(3): 323-337. http://www.surveillance-and-society.org

[58] Citing Turow (2006), from Manzerolle and Smeltzer. 2011, p. 327

[59] Andrejevic, Mark. 2011. Surveillance and Alienation in the Online Economy. p.287

a privacy policy, it seems unlikely that individuals or consumers would have the motivation or capacity to learn as much about them as they do about us.

Moreover, when predictive analytical techniques are used by government it will often be the case that the analytical process and its outcomes are tightly concealed from the public for reasons of national security or public safety. Regardless of the specific application many of the prediction algorithms and software applications are opaque because they “are subject to copyright and trade secret laws, so the public does not get to know who wrote them, how they work or whether the assumptions upon which they are based are sound.”[60]

Discrimination and damage to reputation

Where predictive analytics are used to make inferences about future behaviours and intentions, there is a risk that individuals will be profiled and categorized and potentially become the subject of discrimination according to those predictions. If we consider this issue in the context of targeted advertising, the concern would be that profiles of consumers may lead to “exclusion of access to goods and services” or “price discrimination based not on the goods and services, but on the identities or profiles of customers.”[61] In this scenario, dynamic pricing responds to personal characteristics such as wealth (ability to pay), urgency of need, vulnerability to certain enticements, or marketing approaches.”[62]

The issue is not simply about the collection of information, but also what is inferred from the information. Reputation can be a gatekeeper to services, and it is easy for reputation to be built upon inaccurate information or information taken out of context.[63] It is one kind of problem to be excluded from receiving a certain type of advertisement, but it is potentially much more damaging to be excluded from a government program or subject to a disproportionate level of scrutiny based on mis-information.

Preemption could undermine due process

The aim of predictive and pre-emptive analysis is to make assumptions about what will happen before it even happens. From an ethical perspective, the careless and excessive adoption of technologies that anticipate wrongdoing before it even occurs could have a significant impact on our traditional models of justice, due process and individual freedoms.[64] The due process concept requires that individuals have an ability to observe, understand, participate in and respond to important decisions or actions that implicate them.[65] The opaque nature and complexities of analytics could make it more difficult to accomplish this level of transparency and fairness in process.

[60] Kerr, Ian (preprint) *Prediction, Preemption, Presumption: The Path of Law After the Computational Turn** forthcoming in Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology, eds. Mireille Hildebrandt & Ekaterina De Vries.

[61] Nissenbaum, Helen. 2009. *Privacy In Context: Technology, Policy, and the Integrity of Social Life*.

[62] Nissenbaum, Helen. 2009. *Privacy In Context: Technology, Policy, and the Integrity of Social Life*.

[63] From Jennifer Barrigar’s privacy conversations presentation to OPC on Apps.

[64] Kerr, Ian (preprint) *Prediction, Preemption, Presumption: The Path of Law After the Computational Turn** forthcoming in Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology, eds. Mireille Hildebrandt & Ekaterina De Vries.

[65] Ibid

Predictive Analytics and PIPEDA

Private sector companies in Canada that carry out predictive analytics using personal information will have to ensure their practices are in compliance with the privacy principles contained in *PIPEDA*. Consent, purpose and use limitation, openness and transparency, and accountability will be some of the key areas to reflect on when we examine how predictive analytics is utilized in Canada in different contexts. The following section will identify some of the trigger points where predictive analytics could raise concerns under PIPEDA.

Knowledge and consent

Consent is a key governing principle under PIPEDA, requiring that individuals have a basic understanding of how their information will be used in order to provide meaningful consent for its collection and use. The issue is that a traditional model for consent is somewhat difficult to apply to the complex and dynamic scenario of data analytics. While people value privacy, they seem to be largely willing to exchange their personal information for online goods and free services. Companies tend to obtain consent for these practices by including an ambiguous "notice" to individuals, and that notice is typically buried deep within the confines of a privacy policy that contains complex legal language. Some argue that this form of consent has no meaning unless individuals have a genuine awareness of the profiles that are being compiled and fully understand how their data is being manipulated: "To know which of your data you want to hide you need to know what profile they match; to know if you want programs and profiles automatically adapted to your behaviour, you need to know when and how this happens… [and] these initiatives should not merely be left to contingent market incentives."[66] It is extremely onerous to ask individuals to derive this understanding and provide meaningful consent by reading the convoluted language in privacy policies. Moreover, the reality is that mergers and subcontracting, data sharing agreements between companies and organizations, the scope of data collection points and the linkages being made, and the capacities of the predictive algorithms that only data scientist can really understand, all make this technical and business environment complex and variable over time.

Openness and transparency

The complexity and variability of the online and business environment also poses problems for transparency. Achieving transparency should mean that information handling practices are conveyed to users in a way that is relevant and meaningful to the choices they must make.[67] When we consider the power dynamics and the information asymmetries, the goals of organizations versus the goals of individuals to be social or to utilize innovative technology, and the complexity and largely hidden analytical tools such as predictive analytics, it is no wonder that transparency is a difficult privacy principle to observe. Nissenbaum does not think it is possible to explain the current online advertising ecosystem in a useful way without resorting to a lot of detail. Typically, an organization will explain its personal information handling practices using complex legal and contractual language contained in lengthy terms and conditions or privacy policies, which are generally not read or understood by a majority of people. Nissenbaum's term the "transparency paradox,"[68] captures the

[66] Gutwirth, S. and Hildebrandt, M. 2010. 'Some Caveats on Profiling'. In *Data Protection in a Profiled World*. S. Gutwirth et al. (eds.), Springer Science & Business Media B.V. 2010. p.38

[67] Nissenbaum, H. (2011) 'A Contextual Approach to Privacy Online'. http://www.amacad.org/publications/daedalus/11_fall_nissenbaum.pdf

[68] From The Atlantic. 'The Philosopher Whose Fingerprints Are All Over the FTC's New Approach to Privacy' by Alexis Madrigal. Published March 29, 2012.

essence of this problem. She explains that if a privacy policy finely details every flow, condition, qualification, and exception, it is unlikely to be understood, let alone read;[69] however, summarizing information handling practices in a more simplistic style is no more helpful because it omits the important details that are likely going to make a difference for privacy.[70] Describing the practices and intended outcomes of predictive analytics is already challenging due to its intricacies and complexities, and there should be clear concerns raised where organizations obscure the details of its activities within ambiguous privacy policy statements.[71] One should be mindful of "simulated transparency,"[72] whereby the approach to transparency actually ends up being incomprehensible to the users and overly permissive to the company.

Accountability

Accountability is a key governing principle for organizations that to implement predictive analytics. Being an accountable organization is about more than simply having privacy policies or designating a chief privacy officer. It is about having a business model that gives effect to all the privacy principles, and their underlying meaning.

Ethics is at the foundation of the fair information and privacy principles, and other notions such as appropriate flows of information, reasonable expectations of privacy and contextual integrity. It is fundamentally about acting in consideration of the effects on others and in that way should always play a key role in assessing the privacy implications of the different applications for predictive analytics. Paul Schwartz underscores the importance of ethical analytics and a contextual approach to understanding its implications. He emphasizes that essential components to responsible and ethical use of analytics are the privacy principles, such as notions of accountability and proportionality.[73] He provides some useful overarching ethical considerations that companies and organizations should respect before using predictive analytics:[74]

- Ethical use of analytics should be driven by a company's obligation to be a socially responsible actor.
- An organization's processing, analysis, and decision-making through analytics should respect cultural and social norms about acceptable behaviour, and the use of "sensitive" information.
- A company needs to be accountable; acknowledging that analytics could have a negative as well as beneficial impact on individuals.
- A company should assess the impact of its use of analytics on the trust in the company held by a wide range of stakeholders.

---

http://www.theatlantic.com/technology/archive/2012/03/the-philosopher-whose-fingerprints-are-all-over-the-ftcs-new-approach-to-privacy/254365/ (accessed online May 23, 2012)

[69] for more on this See the 2008 study by Aleecia M. McDonals & Lorrie Faith Cranor 'The Cost of Reading Privacy Policies', 2008 Privacy Year in Review Vol 4:3.
http://moritzlaw.osu.edu/students/groups/is/files/2012/02/Cranor_Formatted_Final.pdf

[70] Nissenbaum, H. (2011) 'A Contextual Approach to Privacy Online'.
http://www.amacad.org/publications/daedalus/11_fall_nissenbaum.pdf

[71] this was raised generally in Pridmore and Zwick, 2011

[72] The concept "simulated transparency" is expanded on in Balkin, J.M.(1998) 'How Mass Media Simulate Political Transparency'. Yale University. http://www.yale.edu/lawweb/jbalkin/articles/media01.htm

[73] Schwartz, Paul.M (2010) *Data Protection Law and the Ethical Use of Analytics*. The Centre for Information Policy Leadership Hunton & Williams LLP

[74] Schwartz, Paul.M (2010) *Data Protection Law and the Ethical Use of Analytics*. The Centre for Information Policy Leadership Hunton & Williams LLP

An assessment of potential impact is the starting point for organizations who are contemplating the use of predictive analytics. This notion ties in well with the OPC's Accountability Document,[75] which states that "an accountable organization can demonstrate to customers, employees, shareholders, regulators, and competitors that it values privacy, not only for compliance reasons, but also because privacy makes good business sense."

Predictive Analytics and the *Privacy Act*

Government departments and agencies in Canada that carry out predictive analytics using personal information will have to ensure that they are doing so in accordance with the provisions of the *Privacy Act* and in accordance with Treasury Board (TBS) policies, such as the Directive on Privacy Practices and the Directive on Privacy Impact Assessments. The proliferation of data has been a key catalyst and ingredient in the emergence of predictive analytics in the private sector, and the public sector trend is similar in that it is marked by increased information sharing across different programs, increased outsourcing or contracting with private sector companies, and increased collection of information particularly where "intelligence" is being sought.

The *Privacy Act* does place some limits with regards to the collection of personal information, namely that personal information is only collected where it directly relates to an operating program or activity of the institution, and government programs shall, wherever possible collect personal information directly from the individual to whom it relates. Government departments are required to inform individuals of the purpose and authority for the collection of personal information, as well as identifying new consistent uses for personal information. It may be in those new "consistent uses" that we find the emergence of predictive analytics as a tool to replace manual scrutiny and analysis of large quantities of data. However, and particularly with public safety and fraud prevention programs, the decision to implement predictive analytics should be preceded by a careful consideration of the privacy impacts, and thorough assessment of the necessity for using predictive analytics, whether its use is reasonable and proportionate to the outcome being sought, and whether the program will be effective while also being minimally intrusive.[76]

[75] Getting Accountability Right with a Privacy Management Program. http://www.priv.gc.ca/information/guide/2012/gl_acc_201204_e.asp

[76] OPC Guidance Document (2010) *A Matter of Trust: Integrating Privacy and Public Safety in the 21st Century*.

Conclusion

The Target story that initiated this research report was a revealing illustration of the prominence being placed on data analytics and the position of "data scientists" within companies. The pregnancy-prediction algorithm was concerning because it demonstrated how analytics can yield very personal inferences about people, is generally a very opaque process, and can generate feelings of being manipulated or profiled.

This is a fast evolving field and the scale of data aggregation and analysis is escalating at a magnitude that outpaces much of the concerned dialogue around the practice. Predictive analytics is a tool that can be applied in a variety of different ways, and ethical considerations and fair information and privacy principles can help frame a contextual approach to assessing the privacy risks associated with a given application of predictive analytics, and identify the uses of most concern. Applying ethical considerations should always begin with the realization that the decisions made based on the outcomes of analytics can have a negative effect on people, that certain information should not be collected for the purposes of analytics, and that there should be boundaries and reasonable limits on the types of assumptions that can be made about peoples' future intentions and behaviours.[77]

Duhigg claims that predictive analytics experts are saying "someday soon, it will be possible for companies to know our tastes and predict our habits better than we know ourselves."[78] The vast potential and expanding scope for predictive analytics makes it an issue that is very much on the radar of those concerned with privacy. However, its complexity and obscurity make it difficult to grasp the extent to which it is already used in Canada, what outcomes are being sought, to what extent it is effective or not, or to delineate in advance which specific forms of analytics will raise privacy concerns, and those that may not. In that way, this research report is only the first step in monitoring the trends in predictive analytics and acquiring a better understanding of the challenges that lie ahead.

[77] Schwartz, Paul M (2010); Kerr, Ian (preprint) *Prediction, Preemption, Presumption: The Path of Law After the Computational Turn** - draft version of the article, author requests permission prior to citation
[78] Duhigg, C. 2012. The Power of Habit, p. 212