AMCTO
THE MUNICIPAL EXPERTS

Policy and Management Briefs

Issue 06 / April 26, 2018

Big Data Policy

CHRISTOPHER RAYNOR
christopher.raynor@york.ca

Terms like "big data", "data lake" and "predictive analytics" are not new[1] but they are being used more regularly in the municipal sector. Many believe these concepts can drive insight, efficiencies and improvements to the way municipalities deliver services by taking huge volumes of data, analyzing them and then deriving new insights. In this way, we equip machines to do the powerful thinking the human brain could not manage at this scale.

This paper will explore the emergence of big data, the theory behind it and some of its potential benefits for local governments. It will also consider some of the inherent risks in using big data and provide some recommended approaches to mitigating or avoiding them.

So, what is "big data?" Some consider it to be the fourth industrial revolution... is characterized by mobile internet, more powerful sensors, artificial intelligence and machine learning.

Devices can collect more information and data than ever before

So, what is "big data?" Some consider it to be the fourth industrial revolution. The first industrial revolution happened around the invention of the steam engine and paved the way for mechanical production. The second occurred in the late nineteenth and early twentieth centuries and allowed mass production with the use of electricity and assembly lines. The third industrial revolution began in the 1960s with mainframe computers and brought personal computers and the Internet. The fourth industrial revolution is characterized by mobile Internet, more powerful sensors, artificial intelligence and machine learning.[2] Other commentators consider that big data is defined by the power to predict.[3]

This fourth industrial revolution is spreading faster than any previous revolution. This is illustrated by the increasing load on Internet Protocol (IP) addresses. Every website on the Internet has a unique numerical IP address (something like 74.25.134.56) which personal computers use to connect to the

ABOUT THE AUTHOR

Christopher Raynor is the Regional Clerk for The Regional Muncipality of York where he oversees the Council secretariat, Access and Privacy and Information Asset Management programs. He also sits on the Region's Information and Technology Governance Board.

Chris has over 15 years of municipal experience and is very interested in the questions raised by the intersection of people and technology.

computer(s) running a particular website. For usability purposes, a different system is used to translate the IP address into something more understandable. Consequently, users only need to know the address google.com rather than the IP address sitting behind it.

Over time the IP address system has evolved. Until recently IP version 4 was used to accommodate all the websites on the Internet and it did this by providing 4.3 billion unique addresses. However, websites and computers are now not the only things connecting to the World Wide Web. Over the last few years many other physical items have become "smart" and connected. These include phones, tablets, thermostats, cameras, fitness trackers, fridges, watches, electrical outlets, lights, televisions and speakers. All these devices need their own IP addresses. In fact, there are so many devices that IP version 6 was launched allowing for trillions of unique addresses.[4]

These devices are connected to the Internet so they can collect data and store it. For example, a fitness tracker can calculate the number of steps you take in a day, the quality of your sleep, your average heart rate and various other data points depending on how sophisticated the device is. Connected thermostats can track your energy use and analyze how it varies depending on the outside temperature and compare your usage patterns with other users in your area to help you understand how energy efficient you are.

These insights come out of big data analytics. Big data is the sum total of all the data points each of these devices collects but also can include data points from other sources such as apps or social media. However, there is not a single generally accepted definition of big data. It typically means data sets that are so large that they cannot easily be managed or analyzed with traditional software systems. Gartner provided a little more clarity by suggesting that big data is characterized by its volume, variety and velocity.[5] In short, we are able to collect and analyze data and information in ever-increasing volumes and speeds and municipalities are investigating ways in which they can reap the benefits of this capability.

Local governments can use big data analytics to increase efficiency and solve complex problems

The ability to analyze large sets of data enables municipalities to detect patterns and deploy scarce resources towards areas of greatest concern. These patterns would be too difficult for the average person to recognize because they are hidden in such huge volumes of information. However, as computer processing power has grown, and continues to grow, it becomes easier and easier for computers to churn through huge datasets and report what they find.

This kind of analysis can help inform objective decision-making and reduces the likelihood of "gut" decisions that may feel right but do not deliver the intended results. A simple example could be an analysis of weather patterns using data from the last several decades. Patterns emerging from this analysis could inform the timing and scale of a winter roads maintenance program in any given year. Similarly, analysis of traffic patterns in a municipality could help inform appropriate signal timings or could even help power dynamic timings that shift throughout the day based on real-time traffic data. Chicago's public health inspection program offers a specific example of how data analytics can achieve positive results.

Chicago more than 7,000 "high-risk" food establishments – those where chefs handle raw food. With only 35 inspectors, it was very difficult for the city to comply with legislated inspection requirements so they used data analytics as a means to prioritize inspections. The system uses expected variables such as date since last inspection and number of previous violations but also other, less expected, variables including three day temperature trends, burglaries, sanitation complaints and whether or not an establishment has a tobacco licence. By sending inspectors to establishments recommended by the algorithm, Chicago increased its rate of finding violations by 25 per cent. [3]

This kind of analysis can help inform objective decision-making and reduces the likelihood of "gut" decisions that may feel right but do not deliver the intended results.

Data analytics can also help make sense of very complex systems such as the London transport network. Using data collected from the fare payment card, "Oyster," Transport for London staff were able to examine the patterns resulting from ore than four million daily passenger movements around London. As a result they found the transport network comprised many large passenger hubs and they are starting to use that information to develop contingency plans in the event of a major failure on any tube line or elsewhere in the system. [7]

These examples show how predictive analytics can improve systems and help focus resources. Assuming the underlying data sets are valid and accurate, then there is no doubt that big data can yield positive results with relatively low risk when applied against "things."

Using analytics to make people-based decisions requires additional caution

Where big data and predictive analytics become more problematic is when they are applied against people. Machine-based predictive decisions towards people can have wide-ranging impacts both on individuals and society as a whole.

One reason for this is the algorithms on which the analysis is based and how

those algorithms are designed. Put simply, an algorithm is a step-by-step process for solving a problem. Although they are generally thought of as a computer function, humans use algorithms all the time. For example, my algorithm for buying a new car might be:

1. Research new cars
2. Test drive a variety of cars
3. Buy the fastest car

Somebody else's algorithm for the same outcome could be:

1. Research new cars
2. Test drive a variety of cars
3. Buy the cheapest acceptable car

In data analytics, machines derive their findings based on algorithms designed by humans. While it seems like the human touch might help soften the cold calculations of a computer, it is possible that the human element is contributing to some of the problems.

The example above showed two algorithms with the same outcome – purchasing a new car – but the definition of success was different. One algorithm found the fastest car that was tested and the other found the cheapest. This definition of success is where algorithms can be problematic. Rather than being objective, unimpeachable processes for rendering solid decisions, algorithms can be influenced by a number of factors such as policy direction, profit margins, lobbyists and inherent biases.

For example, machine-based "risk assessments" are used within a number of US jurisdictions to calculate the likelihood of a repeat offence. This score is then used to determine sentencing and probation programs. However, examples of such systems' findings can be alarming.

In one case in Florida, a teenaged girl was caught riding a six-year old boy's scooter down the street. She didn't go too far before discarding it but she was seen and the police were called. She was arrested and charged with burglary and petty theft in the amount of $80. The previous year in the same area, a man was caught shoplifting approximately $80 of tools from Home Depot. This man had already served five years in jail for armed robbery. The teenaged girl, on the other hand, had a couple of misdemeanours while a minor. Disturbingly, the risk assessments processed against both people suggested the man was a

low risk offender while the girl was high risk. Was it significant that the man was white and the girl was black? Maybe or maybe not, but due to the opacity of the algorithm it is impossible to know. In any case, within two years, the man was jailed for eight years and the girl had no further offences. This casts doubt on the accuracy of the algorithmic analysis and raises questions about its objectivity.[8]

A further example of a faulty or ineffectively-designed algorithm involves the sensitive issue of child abuse. The Illinois Department of Children and Family Services (DCFS) used an algorithm to predict cases of child abuse in the state. It worked by mining DCFS and assigning a risk score to children.

More than 4,000 children were assigned a risk score of 90+, with 369 under ge nine receiving a 100 per cent chance of death or serious injury within the next two years. Unfortunately, at least two children were found dead despite multiple DCFS mistreatment investigations and neither was flagged by the algorithm. Consequently, the system is being discontinued because the predictive analytics "didn't seem to be predicting much."[9]

Even if the algorithm is well designed, it is not always clear how it is rendering its decisions.

The problem is exacerbated because many of the algorithms in play today are proprietary and therefore closed systems. Northpointe designed the algorithm used in the Florida case and Northpointe is a for-profit company that does not reveal its methodology.[10]

Since it is unclear what factors are considered in the risk assessment, their weighting and the accuracy of the underlying data it is difficult to objectively assess how well constructed the algorithm actually is. In fact it might contain a fundamental flaw but few people would ever know. The algorithm may continue erroneously impacting lives in significant ways for several years before such a flaw is detected. This was exactly the case in Illinois where the Chicago Tribune unearthed a significant number of data entry errors as well as design failures, meaning there were no linkages between cases involving siblings or investigations into different adults in the same house.[11]

Even if the algorithm is well designed, it is not always clear how it is rendering its decisions. Maybe its use of historical data is perpetuating a faulty status quo or perhaps it is discarding employment candidates based on predictive behaviours that are not especially important to the hiring organization but that are important to the algorithm. These dangers are why Cathy O'Neil refers to algorithms as "weapons of math destruction".[12]

Increased understanding of data and analytics supports better

decision-making

Municipalities need to be careful that they apply critical thinking to big data analysis. Simply assuming that the analysis is always right can have significant impacts on individuals' lives, such as the young girl from Florida or the families in Illinois.

To conduct this critical thinking, municipalities need the right people in place. Staff tasked with using the decisions and recommendations surfaced through big data analytics have to be able to understand them. Unfortunately, the Corporate Executive Board (now Gartner) cautioned "there's an odds-on chance that someone in your organization is making a poor decision on the basis of information that was enormously expensive to collect."[13]

The Board's research suggested organizations tend to quickly hire a few data analytics experts and expect that expertise will trickle down through the organization. They found this trickle down rarely happens and as a result, the rest of the organization is unable to render good decisions based on data analysis. The Board suggested that organizations are made up of "visceral decision makers" who ignore data and go with their gut, "unquestioning empiricists" who only trust analysis and data and "informed skeptics" who apply judgement to the analysis. After noting that organizations should develop more informed skeptics, the Board indicated that only 38 per cent of workers and 50 per cent of senior managers fit into this category.[14]

...Organizations tend to quickly hire a few data analytics experts and expect that expertise will trickle down through the organization... this trickle down rarely happens...

I have already highlighted some inherent risks of using predictive analytics on people. The problem becomes a lot worse when we consider that decision-makers may not even be able to properly interpret results. The error of assuming correlation equals causation is possible in even the simplest of analyses and becomes more likely in complex situations where analysts have not been properly trained. When these analysts are charged with making sometimes life-changing decisions about people, municipalities should be taking steps to ensure the right people are making these decisions.

Again, the Illinois DCFS case illustrates this. The system took DCFS's daily data dumps and converted them into "real time" risk scores. The company that designed the algorithm states that front line staff should never see these scores but they should instead be analyzed by supervisors who have received training and can prioritize appropriately. However, this was not happening and front line workers were understandably disturbed to receive alerts that "the two youngest children, ages one year and our years have been assigned a 99 per cent probability by the Eckerd Rapid Safety Feedback metrics of serious harm or death in the next two years."[15]

Municipalities can tap into data science expertise through partnerships with universities. The Regional Muncipality of York has developed ties with the Schulich School of Business by hosting events and providing co-op opportunities to students. This raises the profile of York Region as a potential employer following graduation and also benefits the Region by providing short-term data analytics expertise, relatively cheaply.

However, raw data science expertise needs to be tempered with a good knowledge of municipal business and client needs. This is where the informed skeptics come in. They can help data scientists understand operational realities and explain any contextual anomalies that may otherwise be discarded by the data scientist. This kind of close partnership helps to ensure algorithms are designed to meet the needs of both the municipality and the residents it serves.

Municipalities also need to make sure that the datasets they use are fit for purpose. Big data analysis is only as good as its inputs, so if inaccurate data is fed in, inaccurate findings will come out. This is problematic for municipalities because they tend to collect information in silos.

Municipalities also need to make sure that the datasets they use are fit for purpose. Big data analysis is only as good as its inputs so if inaccurate data is fed in, inaccurate findings will come out.

This is partly because privacy legislation makes it somewhat difficult to share information internally but also because municipalities tend to collect specific information for specific services. As municipalities seek to combine more and more of these data sets, they will need to consider the contexts in which each was collected so that the right elements are used to derive insights.

In The Regional Muncipality of York, a simple example is population data. The Long Range Planning team collects census information to provide the most accurate population counts. However, census information is only available at certain points in time. Consequently, other areas in the Region might have to use forecasted provincial growth plan numbers to calculate development charge bylaws. If these two types of population data get combined with many other datasets, it can cause confusion on which one should be used for other analytics purposes.

The problem gets more complicated when discrete business units share data. For example, Community and Health Services data could be combined with Transportation Services data to see what correlations there are between different transportation options and community well-being. However, because these datasets have been collected separately there are likely few, if any, staff who properly understand both datasets and so it would be easy to mistakenly apply a data element to an algorithm thinking it meant one thing when in fact it meant another.

These issues can be addressed through data and information governance. A governance framework sets standards and expectations around who is responsible for maintaining a particular dataset to ensure it is accurate and up-to-date. It also creates standards about what metadata – basically the information used to describe each data element – should be collected so that data elements are understandable to othersm and it is clear which data elements are considered to be the recognized source of truth for, say, population counts. Data governance might also provide that completed algorithms are subjected to the equivalent of a peer review to ensure they are fit for purpose.

The Regional Muncipality of York is starting to establish this kind of governance framework but it is challenging because it represents a significant culture shift.

An ethics-based approach can mitigate privacy concerns

Big data analytics also raises some larger questions around privacy. Privacy legislation, at least in Ontario, is ill-suited to manage issues around big data. The Municipal Freedom of Information and Protection Act (MFIPPA) is ore than 20 years old and did not contemplate the current information ecosystem. MFIPPA's reliance on consent or consistent use as a means to facilitate municipal use of personal information is actually a weakness.

Privacy legislation, at least in Ontario, is ill-suited to manage issues around big data. MFIPPA ismore than 20 years old and did not contemplate the current information ecosystem.

It's common knowledge that almost every interaction with a company or website involves the collection of small or large amounts of personal information. We willingly provide that information in return for free use of services we value such as Facebook, Twitter and the Google suite of products. Rarely do we read the small print of the privacy policies and so we have little understanding of what we have consented to. The same applies to municipal small print.

For example, many municipalities have adopted call centres using customer relationship management software to track calls and outcomes. While this generally provides a higher level of service to resident,s imagine a resident's surprise when, upon calling for bus route information, the customer service person asks them if they need any further follow up regarding their appointment at a sexual health clinic two weeks earlier! That resident may have "consented" to their information being shared but they might not have understood the implications about who would see what information.

That is a straightforward example but consent issues become more complex once large numbers of datasets are aggregated. It becomes very difficult to either track consent or withdraw consent fully or in part. It is also harder to draw a connection between the original purpose of collection and the evolving uses.

Furthermore, in many cases, individuals have no choice on where or how they access government services and so they really have no choice but to consent.

The Municipal Freedom of Information and Protection Act completely fails to address the situation where predictive analytics uses anonymized datasets. Since personal information has often been stripped out of the actual datasets, MFIPPA would not apply and there is no privacy issue from a legislative standpoint. However, the outcome of the analytics on those datasets can affect people as has already been discussed. This is a privacy issue that MFIPPA does not recognize.

For this reason, some people have called for a code of ethics around big data. One proposed framework is based on the following principles:

1. Clarity on practices – be open about what you are doing and when
2. Simplicity of settings – make it easy for people to adjust their preferences
3. Privacy by design – build in privacy protections early and make them the default
4. Exchange of value – ensure people see the benefits of sharing personal information[16]

MFIPPA completely fails to address the situation where predictive analytics uses anonymized data sets.

These principles are a good start in terms of managing relationships with users and clients. They allow individuals to retain control of their personal information and should help maintain trustful relationships with municipalities. They can also help municipalities consider the right thing to do. Sometimes, just because you can do something doesn't mean that you should.

This type of approach is supported by Ontario's Information and Privacy Commissioner. The Commissioner's Big Data Guidelines[17] from 2017 highlighted that big data analytics may be negatively impacted by poor data quality, biased datasets, discriminatory factors and spurious correlations. The guidelines emphasize the need for openness from the outset of a project. Other suggestions, such as consultation with the public and civil society organizations, are useful as a way for municipalities to "peer review" their proposals to ensure they meet the proposed goals in an ethical way.

Another way that municipalities can mitigate the privacy risks of big data is through Privacy Impact Assessments (PIA)[18]. This tool can help to evaluate risks associated with a big data project or system. A good PIA will examine the project conceptually to understand the desired outcomes and measure the reward of those outcomes against the potential risks of going ahead.

Conclusion

Big data and predictive analytics can provide tremendous insight and sometimes at very low risk. For this reason, they are tools that municipalities should investigate and apply where appropriate. However, municipalities should be cautious when they consider these tools as part of any human services programs. Datasets should be carefully studied to ensure they are accurate and that the analysts using them are properly trained. This training should extend to the ethical considerations of the data analysis and not just the analysis itself.

Most importantly, municipalities should give weight to the ethical implications of a project. Don't ask whether the law allows it. Instead ask: what is the *right* thing to do? ■

NOTES

1. John Mashey, presentation to Usenix '99, dated April 25, 1998 [http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf] Retrieved September 14, 2017
2. The Fourth Industrial Revolution, Klaus Schwab, World Economic Forum, 2016
3. Big Data Ethics, Neil Richards and Jonathan King, Wake Forest Law Review, Vol 49, page 393
4. Why is IP version 6 important for internet users? https://www.lifewire.com/why-is-ipv6-important-to-internet-users-2483451. Retrieved September 23, 2017
5. http://www.gartner.com/newsroom/id/1731916 dated June 27, 2011. Retrieved September 14, 2017
6. Up to Code? An algorithm is helping Chicago health officials predict restaurant safety violations http://www.pbs.org/newshour/bb/chicago-revamps-restaurant-inspections-by-tapping-into-social-media/ Retrieved October 2, 2017
7. UCL Engineering – Oyster gives up pearls https://www.youtube.com/watch?time_continue=103&v=9sAugcb2Qj4 Retrieved October 2, 2017
8. Machine bias https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing Retrieved October 20, 2017
9. Data Mining Program designed to predict child abuse proves to be unreliable, DCFS says, Chicago Tribune, December 6, 2017 [http://www.chicagotribune.com/news/watchdog/ct-dcfs-eckerd-met-20171206-story.html] Retrieved, December 7, 2017
10. See Note 8
11. See Note 9
12. https://www.ted.com/talks/cathy_o_neil_the_era_of_blind_faith_in_big_data_must_end?utm_source=newsletter_daily&utm_campaign=daily&utm_medium=email&utm_content=image__2017-08-22
13. Good data won't guarantee good decisions https://hbr.org/2012/04/good-data-wont-guarantee-good-decisions Retrieved October 20, 2017
14. Ibid
15. See Note 9

FOR MORE INFORMATION, PLEASE CONTACT:

Rick Johal
Director, Member and Sector Relations
rjohal@amcto.com
905.602.4294 x232

Eric Muller
Policy Advisor
emuller@amcto.com
905.602.4294 x234

16. What big data needs: A code of ethical practices https://www.technologyreview.com/s/424104/what-big-data-needs-a-code-of-ethical-practices/ Retrieved October 20, 2017

17. Big Data Guidelines, published May 2017 https://www.ipc.on.ca/wp-content/uploads/2017/05/bigdata-guidelines.pdf Retrieved May 8, 2018

18. See, for example, the IPC Guide to PIAs https://www.ipc.on.ca/wp-content/uploads/2015/05/Planning-for-Success-PIA-Guide.pdf

AMCTO
THE MUNICIPAL EXPERTS

AMCTO
Association of Municipal Managers, Clerks and Treasurers of Ontario

2680 Skymark Avenue, Suite 610
Mississauga, Ontario L4W 5L6

Tel: 905.602.4294
Fax: 905.602.4295
Web: www.amcto.com
Twitter: @amcto_policy

About AMCTO Policy and Management Briefs

AMCTO's Policy and Management Briefs are designed to fill a gap in the discussion of local government in Ontario, by fostering dialoguing and promoting rigorous analysis of important topics facing municipalities across the province.

About AMCTO

With approximately 2,200 members working in municipalities across Ontario, AMCTO is Canada's largest voluntary association of local government professionals, and the leading professional development organization for municipal administrative staff. Our mission is to provide management and leadership service to municipal professionals through continuous learning opportunities, member support, and legislative advocacy.