



Big Data for Public Good: A Primer

Final Report

March 31, 2014

Prepared by:

Nordicity

For:

ICE Committee



Table of Contents

1.	Introduction	4
1.1	Project Rationale	4
1.2	Project Mandate	5
1.3	Approach & Methodology	5
1.4	Big Data Factoids	5
1.5	Practical Definitions for Big Data and Open Data	6
1.5.1	Big Data	6
1.5.2	Open Data	8
1.6	Creators and Users of Big Data and Open Data	8
2.	Use of Big Data to Address Public Policy Challenges	10
2.1	Case Studies: An Overview	11
2.2	Public Policy Challenge 1: Urban Transportation	12
2.2.1	Case Study: Chicago Shovels	13
2.3	Public Policy Challenge 2: Democratic Participation in Government and its Institutions	15
2.3.1	Case Study: Online Building and Fire Codes Consultation in Alberta	16
2.4	Public Policy Challenge 3: Urban Planning and Operations	17
2.4.1	Case Study: Social Planning - Wellbeing Toronto	18
2.4.2	Case Study: Operations Planning in New York City	19
2.5	Public Policy Challenge 4: Scientific Research Collaboration	21
2.5.1	Case Study: Big Science Infrastructure	22
2.6	Public Policy Challenge 5: Economic Development	23
2.6.1	Case Study: Using Data for Generating Regional Growth in an Urban-Rural Region in Washington State	24
3.	Issues Arising from Big Data and Open Data	27
3.1	Issue 1: Privacy & Trust	27
3.2	Issue 2: Security	28



3.3	Issue 3: Standards & Interoperability	29
3.4	Issue 4: Collaboration	29
3.5	Issue 5: Making Public Data Open	29
4.	Final Observations	31
	Appendix A – Glossary of Terminology	33

1. Introduction

1.1 Project Rationale

'Big data' and 'analytics' are watchwords of the present era of information explosion and connectivity. The increased use of big data should help evidence-based decision-making in public and private arenas, provided that the data is accurate, accessible, and used correctly.

Big data can be marshalled into the service of the public good. Building big data into the workflow of decision-making should bring about efficiencies in the planning and delivery of public services.¹ One estimate of the potential benefits of the application of big data is a 2.5% to 4.5% in savings in the delivery of public services.² Among other benefits are greater citizen participation in governance and the stimulation of economic development.

While this emerging and growing field of research and data usage can assist policy makers, it can simultaneously pose challenges as to what data should be made available – the open data question. How to protect privacy, provide security, and how to open up data previously privately or publicly held are other major policy questions arising from the promise of big data.

The rationale for this 'primer', then, is to inform policy makers as to the promise of big/open data to help meet public policy objectives, and to identify the issues that 'open data' will raise as governments try to leverage big data.³

This primer examines the concepts of both big data and open data, and as well, identifies their potential use in addressing public policy challenges. It also explains and emphasizes the importance of big data analytics, which is illustrated by example in several of the case studies considered.

¹ There are other useful reports that help substantiate this assumption. For example, a [road map for rolling out Big Data projects in government was released in 2012](#), called "*Demystifying Big Data: A Practical Guide to Transforming the Business of Government*." The report was produced by the [TechAmerica Foundation](#), a nonprofit education organization, in consultation with technology experts in the White House, Internal Revenue Service, Centers for Medicare and Medicaid Services and other agencies.

² Andy Potter of Deloitte at the March 26 2014 ICE/IPAC public forum. Another study conducted by McKinsey Global Institute in 2011, indicates at the macro level that big data analytics and the analyses and means of visualization of such data can potentially transform the global economy, make significant improvements to organizational performance and work to improve national and international policies. See *Big Data - The Next Frontier for Innovation, Competition, and Productivity*, the McKinsey Global Institute, November 2011, (see online at: www.mckinsey.com/mgi).

³ For example, 'Privacy by Design', is a theme of the Ontario Government Office of the Information and Privacy Commissioner; for details see: <http://www.privacybydesign.ca/index.php/big-data-calls-big-privacy-big-promises/>. The current Commissioner, Dr. Ann Cavoukian, has recently addressed the emergence of big data and privacy (e.g., on January 24, 2014 in a speech entitled "Big Data Calls for Big Privacy – Not Only Big Promises" (see: <http://www.privacybydesign.ca/index.php/big-data-calls-big-privacy-big-promises/>), and in another speech on February 12, 2014 called "Data, Data Everywhere – The Need for Big Privacy in a World of Big Data and Surveillance" (see: <http://www.ipc.on.ca/English/Resources/Presentations-and-Speeches/Presentations-and-Speeches-Summary/?id=1376>).

1.2 Project Mandate

Nordicity was contracted by the Intergovernmental Committee for Economic and Labour Force Development in Toronto (ICE) to prepare a short report on big data and open data.

The purpose of this project is to create a basic primer of big and open data with particular emphasis on public policy challenges and opportunities. While case studies are used to illustrate how big data can affect decision-making for public policy and operations, they are not intended to be comprehensive. The project scope is limited as the issues raised are not exhaustive nor comprehensively treated. While we make observations at the end of this report, no recommendations are put forward.

1.3 Approach & Methodology

The overall approach and methodology adopted for this project involved three levels. First, we undertook an environmental scan of relevant online resources. Second, we conducted six targeted interviews with relevant subject matter experts concerning big data/open data issues and challenges. A related outreach was attendance at a timely ICE/IPAC⁴ seminar on Big Data for Public Good. Third, we identified and researched several case studies illustrating different aspects of public policy challenges and use issues with respect to the adoption and use of big data/open data, along with the role analytics and visualization play in making use of this data.

What follows is an introduction to the concepts of big data and open data in this section. Consideration is then given to public policy challenges which can be addressed by using big data/open data in Section 2. These policy challenges are further illustrated through examples based on types of applications derived from a literature review (largely through Internet research). Then, in order to better appreciate the significance of these examples, a real-life case study for each policy challenge is presented.

The discussion then turns in Section 3 to five key generic issues associated with both big data and open data, issues which are being addressed to enable big data and open data networks. Finally, in Section 4 the report makes observations that are based on this primer.

1.4 Big Data Factoids

The following ‘factoids’ about big data provide examples of the scope and scale of emergent data sets in the rapidly growing world of digital data.

As this list of factoids illustrates, some would say we are swimming in a sea of data ... and the sea level is rising rapidly. If so, we will either sink or swim depending on how we handle big data.

⁴ Institute for Public Administration of Canada

- *Each engine of a jet on a flight from London to New York generates **10 terabytes of data every 30 minutes**, where one terabyte is one trillion bytes.*
- *In 2013, Internet data, mostly user-contributed, will account for **1,000 exabytes**. (An exabyte is a unit of information equal to 10^{18} bytes = 1000 petabytes = 1 million terabytes = 1 billion gigabytes)*
- *Open weather data collected by the US National Oceanic and Atmospheric Association has an annual estimated **value of \$10 billion***
- *In 2010, the German Climate Computing Centre generated **10,000 terabytes of data per year***
- *The amount of Internet traffic per second in 2008 exceeded all of the Internet traffic in 1993*
- *Every day we create **2.5 exabytes** of data*
- ***90%** of the data in the world today has been created in the past two years*
- *Every minute **100,000 tweets** are sent globally*
- *Google receives **two million** search requests every minute*

1.5 Practical Definitions for Big Data and Open Data

The following presents working definitions of big data and open data. As there are many terms used in the discussion of these subjects, we have included a glossary as Appendix A.

1.5.1 Big Data

Big data can be defined as ***data sets that exceed the boundaries and sizes of normal processing capabilities, thus forcing a non-traditional approach to analyzing these data sets.***

Big data can be classified into two types - what are called '**structured data**' and '**unstructured data**'. The former is the kind of data that comprise such familiar notions as tables of data in rows and columns, as one might find in a financial spreadsheet. Unstructured data, on the other hand, is characterized by all forms of data that do not represent a set pattern, such as a letter, memo or newspaper. Unstructured data is predominantly text and images.⁵

⁵ For more discussion on a breakdown of structured and unstructured data types, see: <http://smartdatacollective.com/michelenemschoff/187751/7-important-types-big-data>.

The growth in big data is three-dimensional. There is an increase in **volume** (amount of data), in the **velocity** of data transfer (speed of data in and out of an application), and in the **variety** of data (range of data types and sources). These are often referred to as the three V's.

A key challenge from this three dimensional growth arises from the fact that most organizations have their own ways of dealing with and processing data. Without standard formats or processing methods, it can be very difficult for users to work with different data sets. Therefore, processes, standards, and formats that drive interoperability of data are key components of a viable big data ecosystem. Interoperability enables the useful meshing of different products and standards, such as mobile phones from different manufacturers all working on the same network. For example, consider the possible interoperability issues involved in the following very small sample of data types and sources:

- web server logs and Internet mouse-click data
- social media activity reports
- mobile-phone call detail records
- transaction data from bill payments to bank account records
- information captured by a wide variety of sensors in machines or in the environment

As the 'Internet of Things' generates massive data from interconnected sensors, it will add enormously to the huge challenge of management and leveraging of value from this data.

Given these characteristics, big data requires the use of new frameworks, technologies, and processes of management. Technically, institutions using big/open data need to understand the differences between big data and traditional data warehousing and business intelligence practices. Big data has become the new frontier of information management.

When is data 'big'?

Deciding when data crosses the boundary between being 'data' in the conventional sense and 'big data' is a fuzzy domain. Normally, this is the point where data sets exceed the processing capacity of conventional database systems.⁶ This distinction is a rather nebulous concept, experiencing problems similar to those encountered in defining what is considered 'cloud computing.' It is a technology-dependent working definition. From a data perspective, the notion of big data is most frequently characterized by the three V's, noted above (volume, velocity, and variety), as a way to view the bounds between conventional and big data.⁷ Furthermore, individual data sets may not on their own constitute big data but when data sets are merged or 'fused' and explored or 'mined' for patterns, data can become 'big data'.

Analytics is the process of examining data, from a variety of data sources and formats, to deliver insights that can enable decisions in real or near-real time. Various analytical concepts such as data mining, natural language processing, artificial intelligence, and predictive analytics can be employed

⁶ IDC uses the figure of 100 terabytes as the threshold for big data – one terabyte = 1000 gigabytes

⁷ The i-Canada Alliance for smart cities concentrates on the gigabit speed as the next frontier for smart cities, which enables the use of big data applications for many purposes – but does not specify a metric for big data.

to analyze, contextualize, and visualize the data. Analytics creates new business value by transforming previously unusable data and by providing new predictive insights and actionable knowledge.

Big data analytics is the process of examining large amounts of data in a variety of types and formats to uncover hidden patterns, unknown correlations, and other useful information. The primary goal of big data analytics is to help organizations make better decisions by enabling users to analyze massive volumes of data in search of meaningful patterns.

In summary, big data analytics functions are unique because they handle open ended 'how and why' type questions, whereas traditional business intelligence tools are designed to query specific 'what and where' questions. Big data analytics processes unstructured data to find patterns, whereas data processed by conventional data warehouse systems use structured and mostly aggregated data.

1.5.2 Open Data

Open data refers mainly to the concept that some data – particularly public data – should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control.

Using that concept, the City of Toronto defines open data as “data that can be freely used, reused and redistributed by anyone - subject only, at most, to the requirement to attribute and share alike.”⁸ Toronto’s open data policy position incorporates the concept of Privacy by Design, as promoted by the Ontario Government Office of the Information and Privacy Commissioner, and discussed further below.⁹ The federal government has established the policy infrastructure for encouraging and supporting open data practices.¹⁰

Governments are creators and users of big data for their own internal needs. While such data could be deemed to be open data and hence freely available to all, there are various privacy, security, and other policy reasons for not making all data ‘open’ to the general public. While universal open data in some ways is a worthy objective for many public databases, it is far from a universally accepted concept.

1.6 Creators and Users of Big Data and Open Data

Government, industry and science enterprises are all generators and consumers of big data. Technology provides the ways and means by which big data is collected, manipulated, stored, and shared by digital means.

⁸ The City of Toronto’s open data policy can be found here:

<http://www1.toronto.ca/wps/portal/contentonly?vgnextoid=7e27e03bb8d1e310VgnVCM10000071d60f89RCRD&vgnextfmt=default>

⁹ See open data portals for Toronto (<http://toronto.ca/open>) and Ontario (<http://ontario.ca/open>).

¹⁰ “Data.gc.ca provides one-stop access to the Government of Canada’s searchable open data and open information, together with open dialogue, as part of the federal government’s commitment to enhance transparency and accountability.” (from data.gc.ca website)



Governments collect vast quantities of data in a wide range of areas (e.g., weather data, economic data, census data, etc.). Traditionally these data sets have been closed and only used in 'silo-like' situations by those who collect the data for a predefined purpose. Today, by recognizing the benefits of sharing data through open data policies, it has become possible to use data in entirely new and unexpected ways that go far beyond the original purposes envisioned for such data.

Moreover, by adopting an open data policy, organizations such as cities now realize that they can purposely collect new forms of data that are intended to be open and shared. This is resulting in new kinds of relationships between cities and their constituent parts, be they citizens, businesses, or various public institutions from schools to city departments to regional agencies.

The old adage that "information is power" is a truth that prevails when one has access to and can use big data and open data. From the perspective of public policy, making these forms of data more widely available to communities of interest is an empowering experience.

2. Use of Big Data to Address Public Policy Challenges

The use of big data to address public policy challenges has been applied across a number of traditional government functions. They include regulation/oversight, public safety, traffic management, and operational efficiency. Governments of different orders (federal, provincial, regional, and local) can examine and address a variety of public policy challenges. For illustration, this report examines the following high-impact areas where open data (and sometimes big data) are being used to address policy challenges:

- Urban Transportation
- Democratic participation in government and its institutions
- Urban planning, city administration, and operations
- Scientific and technical collaboration
- Economic development

For each of these areas, the report identifies broad policy challenges, and provides a few big data applications to illustrate the range of potential situations in which big data can be or is being deployed. A case study that further explores different facets of big data/open data is presented for each area. It should be noted that the project team did not evaluate case studies to select the best in class. Undoubtedly, there are good examples in Canadian and other country, region, or city contexts which could demonstrate similar characteristics.

As was noted in the definition of big data, there have been various stages of evolution of the concept of big data and in the development of solutions to the challenges it presents. Not all big data applications are at the same stage of maturity, and not all meet the threshold of 100 terabytes, for example. Lacking one or more characteristics of big data does not disqualify a solution from 'big data/open data' status, at least as a case study that illustrates other important aspects of big data/open data. To reprise these characteristics, the following are the main facets of big data/open data:

- **Amount of data** – one general estimate places the threshold which defines “big” at 100 terabytes
- **Open data** – whether it represents a case where previously inaccessible data was made more accessible to the general public
- **Multiple sources** – the combination of different data sources, likely from different public agencies in the case of government
- **High-performance computing** – where data processing requirements and other tools exceed traditional capacity of an organization’s IT systems (including cloud computing services being used by the organization)
- **Data analytics and visualization** – where there are new ways of analyzing data to draw inferences and make predictions, including visualization techniques that aid in analysis and problem solving

- **Collaboration** – where there is new collaboration within an organization, across organizations, or across orders of government

2.1 Case Studies: An Overview

The case studies illustrate various organisational and operational needs that must be addressed in order to effectively utilize big data/open data. Behind every application involving public domain data sets is an abundance of institutional, logistical and human factors issues that go hand-in-hand with data use. All of the case studies cited in this primer include the use of open data and/or existing sources of big data that are illustrative of applications that build big data sets. The case studies are validated as a consequence of the data being open and the need for multi-stakeholder collaboration.

Below is a summary of each case against a set of criteria developed to illustrate the various components of big data and open data. Not all cases meet every criterion, but they all tell important parts of the big data/open data story. Moreover, they all indicate various dynamics of people, organizations, collaboration, and the technical status of the data environment. The various criteria show that most aspects of developing big data applications in meeting public policy challenges relate to non-technical concerns, certainly at early stages of adoption by any one or more partnering organizations.

Table 1 - Summary Table of Case Studies Described and Rated Against Selected Status Criteria (Y= Yes; N= No)

Criteria	Case Study by Policy Theme					
Policy theme	Urban Transportation	Democratic Participation in Government	Urban Planning and Operations	Urban Planning, Administration & Operations	Scientific & Research Collaboration	Economic Development
Case name	Chicago Shovels	Online Building and Fire Codes Consultation in Alberta	Wellbeing Toronto	Risk Based Building Inspection – NYC Fire Department	Big Science Infrastructure	Using Data for Regional Growth in Washington State
Initial objective	Improve flow of traffic associated with snow storms	Seek citizen online input to refine codes	Assist urban planners, citizens, NGOs & companies in accessing city demographics	Reduce potential for building fires and enhance efficiency of the inspection process	Provide computing & technical resources for various public and private research projects	Create interagency collaboration to enhance local and regional economic development opportunities
Data ownership	City sources	Federal & provincial	City, federal, provincial & private sector	City sources	Multiple sources	State & local governments

Criteria	Case Study by Policy Theme					
	Urban Transportation	Democratic Participation in Government	Urban Planning and Operations	Urban Planning, Administration & Operations	Scientific & Research Collaboration	Economic Development
Open data policy	Y	Y	Y	Y	N	Y
Multiple data sources (data variety)	Y	Y	Y	Y	Y	Y
Data volumes greater than 100 terabytes	N	N	N	N	Y	N
Requirement for big data tools and computing	N	N	N	N	Y	N
Need for new data analytics/ visualization	Y	Y	Y	Y	Y	Y
Evidence of high degree of collaboration	Y	Y	Y	Y	Y	Y

What follows is a general consideration of each policy theme with a relevant case study (noted above) as examples of efforts to embrace and use multiple data sources in new ways to develop and/or deliver on public policy positions. They are cases of open data and in most cases nascent big data.

2.2 Public Policy Challenge 1: Urban Transportation

The general challenge

Faced with increasing capital infrastructure and maintenance costs – let alone public frustration with traffic congestion and delays - governments are seeking innovative ways to manage transportation infrastructure and operations. Such challenges include reducing congestion, municipal vehicle fleet management, and preventive maintenance. Success can be measured by better planning of transit investment, greater involvement of citizens, improved service and traffic flow, reduced operating costs, and through other benefits.

How big / open data can help

Much information is already collected through automated data capture systems, as well as through organization call-takers in areas such as city 911/311 systems. Examples of how collected data is being used to address some transport challenges include:

- **First responder route prioritizing** for emergency services (e.g., real-time traffic light management, traffic congestion sensors, deployment of emergency responders – fire, police, and ambulance/hospital availabilities);
- **Dynamic road maintenance information system with citizen input** (e.g., assimilating data from citizen reports about potholes, downed trees on a road or sidewalk or snow plowing needs; road sensors; and weather data); and
- **Parking management and space availability** (e.g., expiry warning notifications on occupied parking matched with GPS data - made available to mobile phone subscribers).

2.2.1 Case Study: Chicago Shovels

The following case study applies to a seasonally cyclical problem, one also faced by Canadian cities – snow removal from streets and walkways. It also illustrates the organisational and social dynamic that ensues when these kinds of data are made available across organizational boundaries.

Case Study: Municipal Snow Removal in Chicago



The applications described below incorporate multiple data sets, analytics and visualization, civic engagement, and multiple city departments. While not meeting the “big data” volume threshold, the data sets will be growing over time.

Chicago Shovels is an example of the use of the emergence of open data to better optimize municipal services at minimal cost. Chicago’s mayor was apparently a major supporter of this initiative.¹¹ At its heart, **Chicago Shovels** is a dynamic community of city employees and volunteers who serve as application developers creating urban applications based on open public data being used with a view to empowering other volunteer citizens to help their communities cope with winter storms¹². In this case, Internet-enabled software tools help connect the public with city resources and empowers neighbours to come together to help Chicago deal with winter.

¹¹ http://www.huffingtonpost.com/2012/01/03/chicago-shovels-city-lauc_n_1181068.html

¹² A summary video of this initiative can be found here:



There are four application components of the **Chicago Shovels** initiative. They are: *Plow Tracker*, *Snow Corp*, *Winter Apps* and *Adopt a Sidewalk*. We describe each in turn. They all depend on open data and some of these apps are capable of generating very large data sets (i.e., big data).

Plow Tracker map is a map that anyone can access, and in Chicago, data is loaded onto the map via GPS sensors on snow plows. During snow storms this Internet-enabled plow tracker map shows the real-time locations of City plows. One can watch as snow clearing efforts start with the major streets and then move to side streets in order to keep Chicago roadways clear.¹³ By accumulating GPS data and associated times and street names, etc. it becomes possible to create dynamic real-time or near real-time thematic maps that portray different features of the state of the road network. Thus, *Plow Tracker* is in the initial stages of evolving into a big data application.

Chicago Snow Corps is a program that connects volunteers with residents in need of snow removal – such as seniors and residents with disabilities. While winter can be hazardous for everyone in a city like Chicago, it can be especially difficult for elderly and physically disabled residents, who may not have the ability or resources to remove snow from their sidewalks and walkways. Consequently, *Chicago Snow Corps* aims to help minimize potential heavy-snow emergencies by pairing volunteers with blocks where elderly and disabled citizens have requested help. To request a volunteer to shovel one's walkway/sidewalk in case of extreme snowfall, a citizen in need calls a three digit phone number (part of the 311 service). That need is matched by pairing volunteers with blocks where elderly and disabled citizens have requested help. The data sets within this app allow for open access, but have privacy restrictions based on personal information and need to know access. This app is the result of public-private collaboration.







Winter Apps is a web application created by citizen developers who have leveraged selected City of Chicago open data to build applications that help people better navigate Chicago in winter. *Twoinch.es* informs and alerts drivers of winter parking bans. *WasMyCarTowed.com* uses the City's towed and relocated vehicle data to reconnect owners with their vehicles.

Launched in June 2011, this app was the result of a first-of-its-kind competition based on using data from multiple orders of government to make applications that improve the lives of large numbers of residents. The competition led to numerous volunteer-developed applications using open data. A sampling of some of these applications is shown below in Figure 1.

<http://www.cityofchicago.org/city/en/depts/mayor/snowportal/chicagoshovels.html>

¹³ A closer look at the technology that assists in this kind of decision-making:
www.cityofchicago.org/city/en/depts/mayor/iframe/plow_tracker.html.

Figure 1 - Examples of City and Citizen Apps created in Chicago using open data

		
<u>SpotHero</u>	<u>FasPark</u>	<u>Taxi Share Chicago</u>
Reserve a parking spot in Chicago or lease out your own.	Find street parking in real-time.	Share a cab anywhere in Chicago. (Android-only)
		
<u>Clear Streets</u>	<u>iFinditChicago</u>	<u>Techno Finder</u>
See where Chicago plows have been	Info on food, shelter, and medical care in your area.	Find public Wi-Fi networks anywhere in Chicago

Adopt a Sidewalk aims to help others by arranging for neighbourhood volunteers to shovel specific public streets and thoroughfares by registering online using a Google map of Chicago. The mobile application allows a volunteer to claim the sidewalks that they will shovel in a given winter. It also allows the volunteers to share supplies like snow blowers, shovels and salt with their neighbours. It also creates 'bragging rights' by permitting volunteers to post that they have cleared their sidewalk.

In conclusion, Chicago Shovels is a multi-part set of applications that are built upon a common framework using open data that is emerging as a big data application suite. By offering dynamic mapping of plow locations, and releasing the associated data publicly, additional applications have been developed through public/private partnerships. **Chicago Shovels** is an ideal example of what can be done by creating a combination of city and citizen resources to develop useful applications that can be used during winter storms and associated clean-up efforts. Such apps also contribute to engaging citizens in community service.

2.3 Public Policy Challenge 2: Democratic Participation in Government and its Institutions

The General Challenge

Governments are often faced with the real challenge of encouraging the participation of citizens they serve in specific decisions, and in delivering services efficiently. Many complex problems require effective consultation with citizens and affected interest groups and stakeholders. Engaging more than just financially well-off interested parties is a difficult objective to attain when data is not easily accessible. Open data solutions with good data visualization capability, such as those that can show different outcomes from different decisions, can enhance democratic consultation.

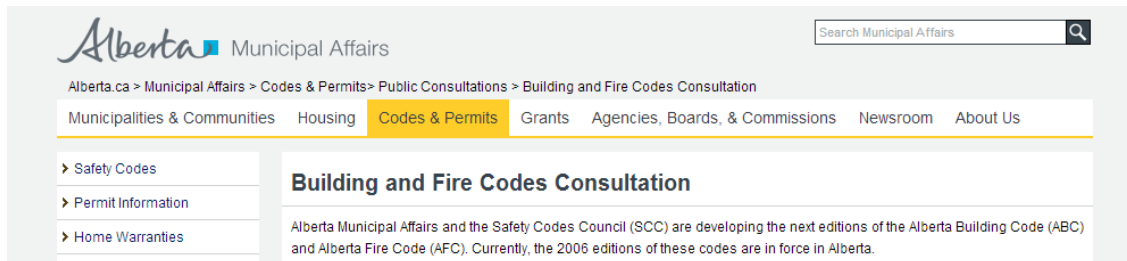
How big / open data can help

Making public data 'open' provides opportunities for governments to engage with the electorate in more meaningful ways. Community and economic data can be made accessible in formats that simplify the process for citizens such that they grasp the outcomes of various public and private initiatives. Other examples that are evolving but are not likely to have yet passed the big data threshold include:

- **On-line consultation through the release of integrated public data sets** in a way that highlights the predicted impact of a specific real property development helps developer consultation with citizens; and
- **Political data bases** that combine voting patterns and socio-economic data can help political parties and stakeholder groups fashion their messages to voter segments.

2.3.1 Case Study: Online Building and Fire Codes Consultation in Alberta

The following case study is an example of government reaching out to citizens with a view to collecting input and views on the development and implementation of new government policy. This case study is about open data being merged from two levels of government for public comment. It is an advance toward big data in that it is merging major data sets and is being displayed visually.



The screenshot shows the Alberta Municipal Affairs website. The breadcrumb trail is: Alberta.ca > Municipal Affairs > Codes & Permits > Public Consultations > Building and Fire Codes Consultation. The main navigation menu includes: Municipalities & Communities, Housing, Codes & Permits (highlighted), Grants, Agencies, Boards, & Commissions, Newsroom, and About Us. A sidebar menu on the left lists: Safety Codes, Permit Information, and Home Warranties. The main content area is titled 'Building and Fire Codes Consultation' and contains the text: 'Alberta Municipal Affairs and the Safety Codes Council (SCC) are developing the next editions of the Alberta Building Code (ABC) and Alberta Fire Code (AFC). Currently, the 2006 editions of these codes are in force in Alberta.'

This case is about multi-stakeholder use of open data with a view to developing new government policy on building and fire codes with direct input from residents. In this case, the province of Alberta conducted online consultations to seek public comment using open data from the federal and provincial governments. Specifically, Alberta Municipal Affairs, a government department, and the Safety Codes Council (SCC), a provincial public/private stakeholder group, set out in 2011 to develop the next editions of the Alberta Building Code and Alberta Fire Code to replace the existing codes in force in the province since 2006.¹⁴

As part of the code development process, Municipal Affairs and the Council sought online public input to the proposed changes to Alberta-specific code requirements for the next editions of the codes. Proposed changes were submitted to Municipal Affairs by the public and stakeholders or developed by the department. All the proposals were then reviewed and approved for inclusion in

¹⁴ This material is principally sourced from: www.municipalaffairs.alberta.ca/cp_building_codes_standards.cfm

the consultation by Building and Fire sub-councils of the SCC. The consultation was open from June 5 to August 31, 2012.

Respondents were linked to a web page that provided mouse click access to review all the proposed code changes. Once inside the survey, clicking on a numeric code item reference opened a copy of the proposed code change in a separate browser window.

For each code proposal, the commenter could select one of four options: a) support the proposal, b) support the proposal with revisions, c) not support the proposal, or d) provide no opinion. Comment boxes to suggest revisions or explain why one did not support a proposed change were also available.¹⁵

In conclusion, the above online consultation was created through a public/private initiative and open to all for comment, but was primarily dependent on informed views from Albertans. This online consultation represented a federal/provincial sharing of open and unstructured data in the form of building and fire safety codes. Some of the data was from previously-published sources (i.e. the national codes). Additional data was created through proposals made with respect to regional/local needs unique to Alberta. The collection of various stakeholder views on each specific part of the code and its proposed revision had the possibility of generating large volumes of data for analysis by the government, thereby enabling them to create a consensus.

2.4 Public Policy Challenge 3: Urban Planning and Operations

The General Challenge

Beyond the traditional role of land use planning, a city or regional jurisdiction faces investment and deployment decisions for the delivery of social services and more operational services (like first responders).¹⁶ The social services planning processes and enhancement of overall operational effectiveness are constant challenges in times of budgetary restraint and the general stress of modern urban life – and present ongoing challenges to such jurisdictions as well as cross-jurisdictions

How Big/Open Data Can Help

Examples of big data applications in urban settings, and potentially involving three levels of government include:

¹⁵ It was anticipated that the revised codes would be approved and implemented into law in 2013, however that has not yet happened. Consequently details on the trial have yet to be published.

¹⁶ The future of cities does not fit easily within disciplinary boundaries. Traditionally, urban research has been the domain of social scientists, while architects, urban planners, and policymakers implement academic findings into real practice. However, the rising availability of city data and the computational resources to model and simulate the complexity of cities brings new groups of specialists and partners into the traditional resource mix of city apps, opening up new possibilities for understanding, managing and building cities.

Urban and social planning

- **Interactive master planning** with thematic maps and associated information resources suitable for use by both municipal officials and citizens alike – through the use of zoning data, historic development patterns, economic indicators, demographic factors, and real estate -transactions – all with good visualization formats to highlight appropriate synthesized data.
- **Cultural and tourism planning** – using geo-spatial inventories of cultural activities, attractions, and artifacts; attendance records, spending profiles, and ticketing data on frequency and origin of participant – for predicting the economic and social payback of new cultural initiatives.

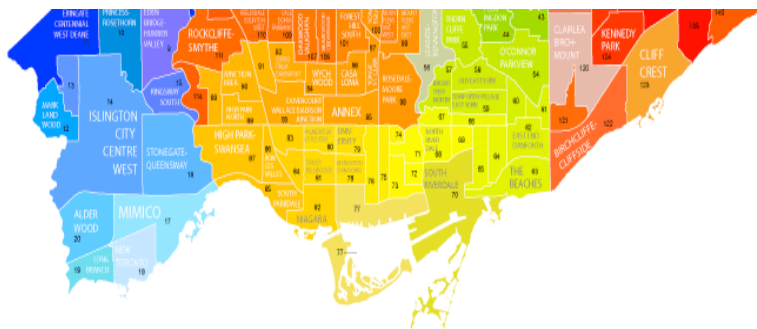
Operations

- Garbage collection routing, police deployment, city owned facilities availability, road maintenance priorities, and dynamic traffic signaling – bringing together the data driven by each of these formerly isolated data sets to improve crew deployment, service support integration, and even social services delivery.
- Cross-agency analytics can more effectively address crime, public safety, and quality of life issues while improving government services, e.g. matching crime data to school bus hours and poverty indices.

The types of datasets in these that are being collected and analyzed are quite varied. They range from survey information and tax roll information for economic indicators, to detailed geographic information systems data which maps each and every tree, house, or utility pole into a large database. The key to big data analysis is the fact that data that was once only available in isolated forms can now be combined with completely different data sets to understand and investigate correlations. This application will be explored in the cases studies below – one for planning and one for operations.

2.4.1 Case Study: Social Planning - Wellbeing Toronto

The following case study from the City of Toronto is illustrative of how social development processes can be improved by adopting an open data approach.



The use of large and multiple data sets to aid in social development in Toronto is exemplified by a platform called **Wellbeing Toronto**. Developed by the City through an extensive consultation process

with many interest groups and data suppliers, it represents a combination of open data and emerging big data related to city neighbourhoods. This map-based visualization tool helps evaluate community wellbeing across Toronto's many different neighbourhoods.

Using a standard Internet web browser, *Wellbeing Toronto* allows anyone to select and combine a number of datasets that reflect neighbourhood wellbeing. The results can be presented as easy-to-read maps, tables or graphs. There are more than 160 indicators that can be searched and viewed, such as: tree cover, crime, environment, health and transportation. In fact, searches can be multi-variable ranging from one to 20 indicators at a time and can be grouped accordingly. These indicators are classified into eleven categories such as demographics, civics, economics, housing, safety and culture.¹⁷ This app is moving toward the status of becoming big data as more and more years of data is made available. (Data is currently available for two reference years 2008 and 2011.)

Wellbeing Toronto was developed to meet the needs of a variety of users: decision-makers who need data to support neighbourhood-level planning; residents who want information to better understand the communities in which they live, work, and play; and businesses that need indicators to learn more about their customers, or to plan their business.

This initiative was put in motion with approval in 2001 by Council of the City's Social Development Strategy as part of Toronto's plan for the future. It proposes a set of specific strategic directions to guide the City's course in providing social programs and services and strengthening communities. All of this data also can be used by residents to better understand the characteristics of different neighbourhoods, and help them decide where they may like to reside. By the same token, this information can be used by policy makers and administrators to make more informed and efficient decisions on where to devote resources and make improvements.

In conclusion, the processes associated with *Wellbeing Toronto* have included consultation with many stakeholders and interested parties, as well as delivery of large sets of data that can be viewed from a single variable up to 20 variables and can be visualized in multiple formats such as graphs or tables. This platform is useful to a wide range of users including city planners, citizens, businesses, other levels of government, and researchers, to name only a few.

2.4.2 Case Study: Operations Planning in New York City

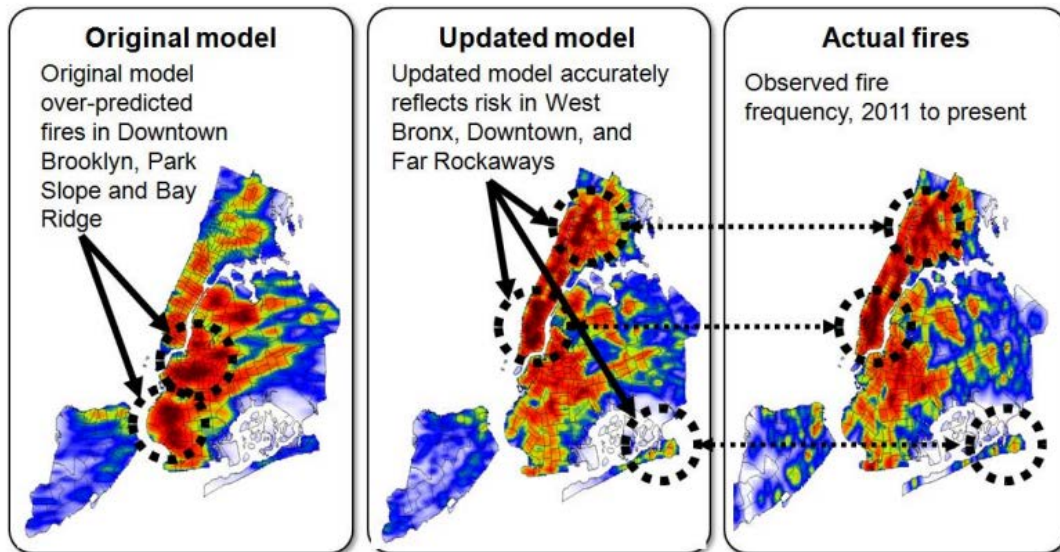
An example of large urban data sets being used for operations planning is drawn from New York City's experience. The New York Fire Department (FDNY) has inspection jurisdiction over roughly 300,000 buildings in the city and inspects some 25,000 building a year. The Department seeks to use

¹⁷ For details see:

<http://www1.toronto.ca/wps/portal/contentonly?vgnextoid=4209f40f9aae0410VgnVCM10000071d60f89RCRD&vgnextfmt=default>.

data mining and analytics to reduce fires and their severity by looking at various trends and indicators.¹⁸

Figure 2 - Graphic illustrating NYFD's Risk Based Inspection System (RBIS)



Analysts at FDNY say that some buildings are linked to characteristics that make them more likely to have a fire than others, poverty being one example of a key indicator. Other factors that correlate with deadly fires include the age of the building, occupancy, whether it has electrical issues, the number and location of sprinklers, and the presence of elevators.

It is, however, difficult to absorb all the relevant factors at once. For this reason, New York officials have built an algorithm that assigns each one of the city's 330,000 buildings liable for inspection with a risk score. When fire officers go on inspections, the application provides a list of buildings, ranked by their risk score that they should visit first. An example of how an analysis of pre- and post-deployment of this app produced dramatic results follows. By comparing pre-deployment with respect to inspections conducted and severity of violations versus a similar comparison post-deployment, it was found that the first 25% of inspections in a given year identified 21% of the severe violations whereas the first 25% of inspections using the app yielded 71% of the severe violations. This outcome is but one of the benefits of mining this kind of building data.

In conclusion, the FDNY example shows how big data solutions can help identify priorities for fire inspection. Similar kinds of apps with appropriate data sets could be constructed and shared in respect to many city department mandates. Again as we see in this example, data sets will multiply

¹⁸ This is elaborated further in NYC's Annual Report on Analytics, found here: http://www.nyc.gov/html/analytics/downloads/pdf/annual_report_2013.pdf

and expand, likely requiring more big data infrastructure to accommodate the exponential growth of big data.

2.5 Public Policy Challenge 4: Scientific Research Collaboration

The General Challenge

Scientific research builds, uses, and shares big data among national/global research teams. The collaborative research process that such work relies on requires effective big data frameworks, capacity, and procedures. This means not only having access to large amounts of data, but having the tools and processes available to solve the complex problems that present themselves.

The general challenge is to ensure that big data infrastructure and tools are good enough for strong participation by Canadian academic and industrial research in global cooperative efforts. If that infrastructure and those tools suffer, the recruitment of top notch researchers becomes more difficult and Canada gets more isolated in scientific research – whose data sets are expanding exponentially.

How big / open data can help

By its very nature, the generation and use of big data across many disciplines improves the performance of scientific research, including genomics, proteomics, nuclear research, astronomy, medical sciences (like brain research), and modeling large-scale engineering structures. Organizing the data in a meaningful way can advance research, and mankind's general understanding of the world around us in many ways.

Some current big data scientific research projects include the following examples:¹⁹

- When the **Sloan Digital Sky Survey** (SDSS) began collecting astronomical data in 2000, it amassed more in its first few weeks than all data collected in the history of astronomy. Continuing at a rate of about 200 gigabytes per night, SDSS thus has amassed more than 140 terabytes of information. When the **Large Synoptic Survey Telescope**, successor to SDSS, comes online in 2016, it is anticipated that it will acquire that amount of data every five days.
- The **NASA Center for Climate Simulation** stores 32 petabytes of climate observations and simulations on one of its large high performance computing clusters.
- The world's largest set of data on **human genetic variation** is freely available on the Amazon Web Services cloud. It measures 200 terabytes – the equivalent of 16 million file

¹⁹ Big data also has emerged in computational social science. For example, researchers have used Google Trends data, a software tool and associated data sets, specifically to search for trends in the data in order to demonstrate that Internet users from countries with a higher per capita gross domestic product are more likely to search for information about the future than information about the past. The findings suggest there may be a link between online behaviour and real-world economic indicators. The results hint that there may potentially be a relationship between the economic success of a country and the information-seeking behaviour of its citizens captured in big data. This kind of analysis would have been impossible to quantify without the ability to mine and visualize big data sets.

cabinets filled with text, or more than 30,000 standard DVDs. Decoding the human genome originally took 10 years to process; now it can be achieved in less than a week and, from a technology perspective, the **DNA sequencers** have divided the sequencing cost by a factor of 10,000 in the last ten years.

One of the key means by which that big data can be leveraged is through the ability to undertake complex visualizations of large datasets. These visualizations can help interpret what the results actually mean. High performance computing is thus often used as a tool for visualizing massive data sets. The following is an illustration of the power of visualization.



A visualization created by IBM of Wikipedia edits. At multiple terabytes in size, the text and images of Wikipedia are a classic example of big data (from: http://en.wikipedia.org/wiki/Big_data).

Note that 1 TB = 1000000000000 bytes = 10^{12} bytes = 1000 gigabytes.

2.5.1 Case Study: Big Science Infrastructure

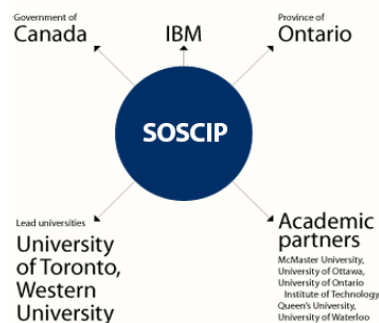


This case study is a description of how the **Southern Ontario Smart Computing Innovation Platform (SOSCIP)** plays a key role in big science through provision of big data infrastructure. SOSCIP was launched in 2012 and its digital processing research infrastructure is shared among universities, governments, and increasingly companies – particularly small to medium size enterprises (SMEs).²⁰ The diagram (overleaf) illustrates the principal partners who created this \$210 million initiative. Big data project themes at SOSCIP include cities, health care, energy, water, and computing. The consortium is supported by ultra-broadband networks: ORION, the Ontario advanced optical network

²⁰ For more information see: www.research.utoronto.ca/edge/edgenet/summer2012/big-data-big-impact/#sthash.mi46Hif7.dpuf

that provides provincial broadband networking for interdisciplinary digital research; and CANARIE, Canada’s national broadband research network which allows for Canada-wide and international connectivity to similar research and education networks around the world.

A major milestone in the establishment of SOSCIP was the installation at the University of Toronto of a high performance computer known as Blue Gene/Q, which is designed for big data analytics and was ranked at its launch as the 65th most powerful supercomputer in the world — and is the most powerful computer in Canada.



So far, SOSCIP appears to have achieved many of the milestones that were set out at the onset of the initiative; for example, there are more than 40 projects underway among seven university partners and over half of those projects involve partnerships and collaborations with the private sector. In addition, twenty one postdoctoral fellows provide key technical and research expertise in supporting the consortium’s partners.²¹ The plan in the coming years is first to grow the community of highly skilled technical and research personnel and then to expand strategic private sector partnerships. Both initiatives aim to leverage the powerful computational resources available through SOSCIP in analysis of big data sets and visualizations of these analyses.

In conclusion, the SOSCIP consortium has created a southern Ontario-based distributed computational platform for the benefit of researchers seeking answers to a vast range of big data and computationally intensive problems in big science and many other sectors of the economy. At the same time it is building a base of highly skilled knowledge workers in the Greater Toronto Area. Because of the growth of big data applications and the amount of big data, the state of Canada’s digital infrastructure in the future is dependent on continuing public-private partnerships and investments.

2.6 Public Policy Challenge 5: Economic Development

The General Challenge

Forecasting economic development outcomes as a result of public and private investment have traditionally been very complex tasks.²² This complexity is due in part to limited data sets and the associated analytics needed to understand impacts. It is also the lack of a robust predictive capacity to estimate outcomes of economic development decisions.

²¹ For a progress report of this project, see <http://sosscip.org/impact-report/cities.html>

²² There is in fact growing international collaboration, e.g. EconomicDevelopment.org, a Canadian-based website, which is devoted to sharing economic development news and resources from around the world. As a community of experts, professionals, and members of the public engage in the dialogue, big data/open data approaches should develop.

Another general challenge is to develop open data policies that stimulate private sector investment in the commercialization of software and services designed to process, display, and market previously inaccessible public data. Users or customers for such new services are government, the private sector, and academic institutions. While this report does not present a case study involving this economic development challenge, there is ample evidence of the robustness of the sub-sector of information and communications technologies (ICT),²³ and of new companies raising substantial capital based on big data solutions.²⁴

How big / open data can help

Through the use of big data/open data, economic development can be stimulated. Developing collaboration within and across communities of interest is a critical step in utilizing and sharing big data/open data when applied to these areas.

Examples of economic development and workforce alignment improvements include:

- **Improved investment decisions**, targeting of inbound investment opportunities, and other planning measures could be enhanced in the development of industry and geographic clusters; e.g., through industry specific capabilities, labour force capacity, technology trends specific to an industry sector, public and private investment sources, and linkages to R&D teams;
- **More open public data to enable commercial and non-profit applications**, e.g., specific data and predictive analytics related to various forms of insurance as well as weather and climate data with predictive capability for merchants as it relates to buying habits;
- **Corporate location criteria** and conditions with site comparison capabilities including tax rates, community amenities, and skills concentrations by area; and
- **Business assistance programs** and contacts supported by interactive thematic maps of services, zonings, and land use rights, job skills availability.

2.6.1 Case Study: Using Data for Generating Regional Growth in an Urban-Rural Region in Washington State

The following is an example of a collaborative effort involving regional economic development coupled with workforce considerations. It is predominantly a case of several involved entities organizing themselves and changing working relationships in order to lay the groundwork for future collaboration – with open data policies to access economic development and workforce related data

²³ IDC, a research firm, predicts that [the market for Big Data technology and services will reach \\$16.9 billion by 2015](#), up from \$3.2 billion in 2010. That is, according to IDC, a 40 percent-a-year growth rate — about seven times the estimated growth rate for the overall information technology and communications business. (Source 2012 IDC report: *Big Data Market By Types: Worldwide Forecasts & Analysis (2013 – 2018)*)

²⁴ The [Canadian Digital Media Network](#), the University of Waterloo, Communitech, Open Text and Desire2Learn were awarded matching funds by the federal government in the 2014 budget to create the *Open Data Institute*, which will support the commercialization of Canadian entrepreneurs with big data solutions.

sets. This case is not specific to big data sets, but it describes goals and processes that are essential in order to share big data within and across organizations.

In 2009, the Washington State Legislature passed a bill that called for coordination between workforce and economic development entities. The legislation gave each workforce development area in the state a directive to perform an industry cluster-based analysis to identify crucial sectors in their region.

This case study, called *Targeted Cluster Identification Strategic Alignment*, highlights processes for collaboration (e.g., governance), a critical factor in applying big data and open data to any public good. It also presents and describes the data-driven strategies and data assessments for each of the industry clusters described in the report.

Although a study was previously completed that included the Pacific Mountain Workforce Development Area, home of the state capital, it used out-of-date labour market data and settled on clusters (e.g., coal mining) that reflected old numbers and dwindling industries.

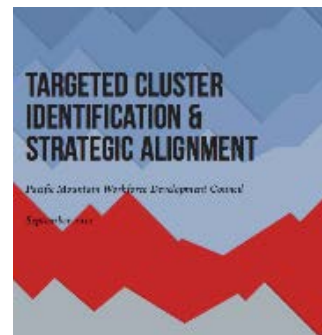
The new state-mandated cluster studies were intended to be part of joint strategic planning between regional workforce and economic development groups. Yet as was evidenced in the Pacific Mountain region, there were natural barriers in bringing both groups together. For example, workforce and economic developers in the region often worked in isolation in moving toward meeting their own goals. Economic developers from county to county typically also worked individually, and they often competed against each other.²⁵

The five clusters that became the final targets for economic promotion and development were:

- Food production;
- Wood product and paper manufacturing;
- Life sciences;
- Chemical products and plastics manufacturing; and
- IT/telecommunications.

Arriving at these clusters and promoting them is documented in the report. From an issues perspective, as discussed above, this initiative spanned all five topics with collaboration and data sharing being the two dominant challenges.

In conclusion, one of the important results of the study was that it has opened avenues for counties, economic development and workforce specialists, and local colleges to have focused, meaningful conversations with business owners about how to grow their businesses and the region in general. This case study shows how economic development and work force readiness teams in a number of



²⁵ A sample of such a study can be found at: www.economicmodeling.com/2012/12/17/using-data-for-regional-growth-how-an-urban-rural-region-in-washington-state-reinvented-its-approach-to-workforce-economic-development/



bordering jurisdictions learned to collaborate firstly within each county and then across jurisdictions. Moreover, it allowed each of these groups to interact in their respective communities in a coordinated way by utilizing a fact-based approach. This case is primarily one that emphasizes the need for cooperation within and among competing jurisdictions and the need to share data in an open fashion.

3. Issues Arising from Big Data and Open Data

In the past few years big data has emerged as a new reality in aiding organizations to develop and deliver on their respective policies and plans for serving their clientele. In the case of a jurisdiction, clientele are citizens, businesses, visitors, and other levels of government, among others.

The collection and associated release of any data can pose challenges to anyone engaged in this field, whether in the private or public sector. An understanding of these issues is critical to what is known as responsible data usage. The following issues are examined briefly below:

1. Privacy & Trust
2. Security
3. Standards & Interoperability
4. Collaboration
5. Making Public Data Open

3.1 Issue 1: Privacy & Trust

Modern cities, with their dense populations of mobile citizens, are incredible generators of big data: mobility data, city resource use data, state of infrastructure data, etc. Each of these kinds of data, if analyzed with modern tools, can supply rich, actionable information useful to other citizens, to city managers, and to policy makers. For instance, it has been shown that mobility patterns of people who are unemployed may be a more reliable source of statistics on unemployment than the data supplied in the traditional government registration process.²⁶ Alternatively, traffic data collected with specialized, inexpensive sensors can, for example, be used to identify sections of the city transportation networks in which vehicles move too closely to cyclists and others, creating potentially dangerous situations. However, as this data is collected, it is important to consider how it will be used in manners that maintain privacy and build trust between all parties.

As there is a general agreement on the value of mobility and similar data, the question of availability of this data arises. For instance, mobile phone operators have mobility data generated by subscribers' accessing various cell towers, but making this data available for research is rarely done (largely due to privacy concerns). This corporate policy posture calls for a rethink regarding the strictly proprietary character of such data. Perhaps aggregated, properly privacy-ensuring forms of mobility data could be made for policy development and delivery purposes.

The above discussion places issues of data privacy and privacy-enhancing technologies at the centre of any big data effort applied to cities. The use of big data may prove to be beneficial; however, privacy-enhancing techniques seem to be a pre-condition for its effective use. Thus, it follows that novel privacy solutions are needed, especially concerning a person's mobility data, be it cell phone

²⁶ Personal communication from Professor Stan Matwin, Canada Research Chair in Big Data Analytics, and Director Dalhousie University Institute for Big Data Analytics (March 2014). See: <https://bigdata.cs.dal.ca/>

calls or mobile use of the Internet. Consequently, privacy and trust are essential considerations following from the effects of these drivers.²⁷

Big data and open data both pose a range of public policy questions with respect to data is collected, used and stored. Central to the treatment of this data is the need to treat personally identifiable information with strict privacy practices while maintaining an appropriate level of security on all the systems used to store, manipulate and transmit data. Anonymizing of personal data is an emerging practice for seeking patterns in and among large populations.

3.2 Issue 2: Security

Security refers to the physical integrity of data and the applications that manipulate and transmit the data. Security represents the encoded protection provided by the information and communications technology infrastructure in accordance with appropriate policy-driven standards. Big data holdings are anticipated to be targets for attack by malware aimed at stealing data for benefits such as economic reward. Consequently, security is a priority technical requirement for big data and associated applications, especially for any 'real-time' control systems such as water works and traffic lights.

Currently there are two primary types of security threats: 'routine vulnerabilities' that are relatively easy to correct; and 'systemic threats' that are insidious and hard to detect and counter.

Routine vulnerabilities comprise a term that is associated with malware, which is specific code that infects machines when a user accidentally or with intent opens a file or visits a website that is purposely infected in order to steal data such as the user's identity.²⁸

Systemic threats are inherently difficult to detect and prevent and are normally used to target specific organisations, rather than average computer users. Examples of big data that are targeted in this way include very large customer lists with personal identification information that can be used to fraudulently gain access to financial assets and other personal records. There have been a number of high-profile thefts of this nature in the past year (e.g., Target stores in the United States²⁹).

The next stage in the evolution of big data security includes a major new component to the Internet, the "Internet-of-Things." As mentioned above, it comprises a wide range of sensor networks that provide a range of functions from controlling critical infrastructure-based systems like the power grid

²⁷ Privacy issues can emerge in other legal forms as well, largely as unintended consequences. For example, police officers not issued smart phones will tend to use their own personal smart phones to record evidence (even video). Courts may allow the seizure of the officer's phone that would contain all the data, including that connected with work as well as the personal information on the phone.

²⁸ Malware is most commonly associated with email attachments that contain the malware unbeknownst to the user, consequently leading to infections that spread via automatically attaching to all email addresses in one's email address book. The consequences typically range from degradation in performance to failure of targeted applications or operating system components. Under such attacks computers also can be hijacked and used as 'bots', which create networks of infected computers in order to perform large scale theft or fraud.

²⁹ See for instance: <http://www.usatoday.com/story/money/business/2014/01/10/target-customers-data-breach/4404467/>

or municipal water supplies, (e.g., both potable and gray water). To date these sensor systems have very limited built-in security as each sensor is a tiny computer with limited ability to self-protect. The spread of the Internet-of-Things is a looming problem that computer scientists and engineers are actively working on to find reliable security solutions.

3.3 Issue 3: Standards & Interoperability

Standards are a central issue in considering how best to maximize the benefits of using big data/open data. Without standards about formats and encoding procedures, data sets would be incompatible and consequently difficult to merge (or fuse) and equally difficult to mine for specific facts and patterns.

Standards also lead to optimizing interoperability by following set rules such as the format of the data, much like the format of a business letter with layout and content rules such a font type, size and placement on the page. Interoperability, which is primarily a technical matter, also is essential in order to allow smooth collaboration among both data sets and applications that manipulate the data.

3.4 Issue 4: Collaboration

Examining the use of big data in municipal settings has led to identifying numerous projects and domain-specific applications such as real-time traffic and utility management, and various forms of urban planning (e.g., thematic maps). Many of these projects are a result of formal collaborations among various partners both within an organization and across organizations, including public and private groups.

Collaboration within and among organizations is an essential part of any application of big data. Collaboration is, in part, affected by the form of governance put in place to oversee a given big data project. Rules of engagement need to be negotiated by the participating parties and should include the ways and means by which any collaboration is to be managed.

Collaboration is a multi-faceted concept. There is the base level of agreeing to collaborate in terms of people and organizational processes at one end, and collaboration in shared data and possibly even applications for manipulating data at the other end. The nature of the collaboration will be drawn out on a timeline of engagement as sharing moves from agreements to actual data and application exchanges. In the case of High Performance Computing, collaboration often involves joint analysis of a common data set or set of data sets. In order to expedite such collaborations there remains the need for some form of standards for encoding and/or transforming the formats of data such that they become interchangeable to a degree that allows interoperability.

3.5 Issue 5: Making Public Data Open

With the advent of big data archives, aspects of public good are being transformed from provision by public institutions (e.g., GPS data and various kinds of maps) to private services operated at no cost to the user (e.g., the Weather Network's apps, Google maps, and navigation). In these cases, the data is

generated by government at their cost and made available to private concerns for both public good and private gain. Consequently, public-private partnerships are emerging as a common framework around which government and industry collaborate in regard to utilizing big data.

Furthermore, if the Canadian (or other) government was to supply open data, it would likely provide such data as a public good. In so doing, several competing interests must be considered. For instance, it may be necessary to balance the benefits of giving 'land developers' access to and use of big data for business purposes versus the interests of individual citizens. Should certain uses of open data be price-sensitive? For example, the accrual of net benefits at public expense for developers is likely to significantly benefit such business users more than citizens. Consequently, open data falls into an ongoing debate about who pays for government-supplied data, regardless of whether it counts as 'big data.'

Collaboration and the bottom-up approach characterize the aim in developing and deploying most urban big data/open data initiatives. In effect, the emergence of these applications has been opportunity driven, often with a top-down policy declaration to find ways and means to exploit big data holdings to improve and enhance city services.³⁰ It is clear from experiences of many cities that the use of big data and open data is creating both enhanced services and new services in managing and administering the diverse responsibilities of city governments.

³⁰ New York City has faced the common challenge in creating collaboration among various business units of the city administration, in grappling with standards and interoperability issues, as well as addressing privacy and trust, and defining metrics to measure improvement in service. Projects are coordinated by the New York City Mayor's Office of Data Analytics. (See www.youtube.com/watch?v=S6EvneIRiTo).

4. Final Observations

The policy challenges outlined in this paper could be considered as the start of a conversation. More in-depth examination of any one of them, or others not mentioned in any detail, would be a logical next step to be taken (e.g., urban transportation, urban planning and operations, and economic development).

From a policy perspective, 'open government' is a laudable democratic value. Open government means more than "open data," but it is difficult to imagine effective open government without concomitant open data policies.

While being 'good for democracy,' open data raises immediate personal data confidentiality issues. Accordingly, policies need to reflect that concern and protect citizens and organizations. Data security is also a fundamental technical concern. Data security is relevant to data owned or controlled by the state as well as by private sector institutions that are not practicing a policy of open data.

Open data costs money – by organizing data so that it is accessible, by citizen engagement to broadcast its availability, and by rendering it in digital form in the first place. So, open data means that governments have to prioritize what is to be made open.

Big data can help government analyze policy and improve the effectiveness of its service delivery and other operations. Effective analysis of big data along with intelligent use of the patterns examined gives government better information to make decisions, allocate resources, and improve the return on investment in enhanced and new services. Because big data merges data from different sources, it leads to greater collaboration across governments. Big data is collaborative-inducing, especially in the public sector.

Exponential growth of data is particularly apparent in major research projects – particle physics, genomics, biomedical research and other basic sciences. For Canada to be plugged into global research teams, the country needs a digital infrastructure of high-performance computer processing, high speed connectivity, effective 'middleware' solutions, and data storage. Consequently, public policy whose objective is to encourage R&D ultimately involves the generation and use of big data.

Governments at all levels have primary objectives relating to economic development and employment generation. Big data can help in two ways. First, it can be used by industry to make decisions; e.g., corporate location and pointing out features and conditions to enhance inbound investment. Second, big data infrastructure is one of the fastest growing ICT sectors. As well, the business of leveraging access to, and the analysis and visualization of, big data is a ripe area for new ICT entrepreneurs. Data analytics and data visualization are crucial components for making big data useful to government, industry and citizens – and again fodder for innovation and entrepreneurship.

Effective use of big data requires standards, protocols and software that aid interoperability. Effective use of big data engenders a collaborative approach and horizontal analysis, solutions, and service delivery. Government departments gain more from big data when they collaborate.

Big data also goes hand in hand with data analytics, thus big data is an ecosystem – high speed links, computer processing, innovative data leveraging firms, analytics and visualization – so big data opportunities need investment from public and private sources.



Many governments at the local, regional, and national level have grasped the importance of big data. They know that big data and open data require intelligent action. In recent years, Canadian cities have shown well in the international smart cities' awards (e.g. in 2013 Canadian cities took two of the top seven intelligent cities awards as selected by the Intelligent Community Forum).

Beyond our big data pipes and recognition, however, is the 'what next' question, how do we take effective advantage of big data solutions. Big data can be overhyped and turn up all sorts of invalid correlations (i.e., no cause and effect). Deploying big data smartly will be the big test for governments. Just how much effort, what priorities, and how to ensure that benefits are reaped – all are current big data issues.

A further step in the country's development in open/big data would be a **strategic performance measurement** of where we stand (locally, provincially, and nationally). What are the best practices for measuring whether and how we obtaining benefits as a society, and at what price? Are we ahead or behind our competitors? These issues should be addressed in future analysis of big data/open data to ensure that public policy is appropriately incorporated into investments and operating decisions related to big data developments.

Appendix A – Glossary of Terminology

Agile Computing	Agile computing is a group of software development methods based on iterative and incremental development, where requirements and solutions evolve through collaboration between self-organizing, cross-functional teams. Compared with traditional software engineering, agile development is mainly targeted at large complex systems and projects, often involving big data.
Amazon Web Services	A Cloud Computing service offering on-demand delivery of information technology resources via the Internet with pay-as-you-go pricing.
Analytics	Analytics is the discovery and communication of meaningful patterns in data. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance. Analytics often favors data visualization to communicate insight.
Anonymized Data	Data that has any personally identifiable details removed.
Artificial Intelligence	Artificial intelligence is the ‘intelligence’ exhibited by machines or software. It is a branch of computer science that develops machines and software with human-like intelligence.
Bar Code of Life	‘DNA barcoding’ is as a way to identify species. Barcoding uses a short genetic sequence from a standard part of the genome the way a supermarket scanner distinguishes products using the black and white stripes of the Universal Product Code, the strips typical of a product bar code. This project has become a major international big data effort in life sciences that started at the University of Guelph about 10 years ago. (See http://www.barcodeoflife.org/)
Big data	Big data can be defined as data sets that exceed the boundaries and sizes of normal processing capabilities and which force a non-traditional approach to analyze these data sets.
Blue Gene/Q	Blue Gene/Q is an IBM high performance computer that can reach operating speeds in which there are some 10^{15} calculations per second, that is 10 followed by 15 zeroes, a very large number. It also offers low power consumption for such a class of machine when

compared to others on offer. Blue Gene/Q was designed to manipulate big data. The most powerful high performance computer in Canada is a Blue Gene/Q within the Southern Ontario Smart Computing Innovation Platform. (See below for SOSICIP)

Bot Is a software application that runs automated tasks over the Internet. Typically, bots perform tasks that are both simple and structurally repetitive, at a much higher rate than would be possible for a human alone (e.g., gaming bots and auction-site robots). Another, more malicious use of bots is the coordination and operation of an automated attack on networked computers.

Broadband Broadband refers to a communication bandwidth of at least 256 thousand bits per second. In other words, broadband is a communications channel that can pass at least 256,000 bits/second. Typical internet connections for home and office use range from roughly 1.5 megabits to 10s of megabits/second. One megabit equals 10^6 bits or one million bits.

Business Intelligence Comprises a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making.

Byte A byte is a unit of digital information in both computing and telecommunications. It most commonly consists of eight bits, where a bit can have only one of two values, zero or one.

Canada Research Chairs The Canada Research Chairs is a Federal Program that invests some \$265 million per year to attract and retain some of the world's most accomplished and promising researchers. Chairholders are actively conducting research excellence in engineering, the natural sciences, health sciences, humanities, and social sciences.

CANARIE CANARIE is a Canadian government-supported non-profit corporation, founded in 1993, which maintains a set of leased wide area network links for the transfer of very large data files. The core network consists of 19,000 km of fibre optic cable capable of speeds as high as 100 Gbit/s but generally operated at 10 Gbit/s. The network is used primarily by education and research bodies across Canada, with links to similar networks in other countries. In terms of capacity,

CANARIE currently averages 6,600 terabytes of data per quarter, a terabyte being equivalent to a trillion bytes. To put this into perspective, it would be equivalent to transmitting the complete contents of 3,300 academic research libraries every three months, or roughly one entire academic research library every 40 minutes. (See www.canarie.ca)

CFI	Canada Foundation for Innovation, a granting agency of the Government of Canada that supports research infrastructure initiatives at qualified academic institutions in Canada. (See www.innovation.ca)
Cloud Computing	Refers to the on-demand delivery of information technology resources (e.g., computing and storage) via the Internet, often with pay-as-you-go pricing.
Cluster	In terms of economic development, a cluster represents a concentration of similar businesses and associated support infrastructure (e.g., legal, financial, research and consulting services) that add value to a specific geographic region and its environs.
CRTC	The Canadian Radio-television and Telecommunications Commission, Canada's telecommunications and broadcaster regulator. (See www.crtc.gc.ca)
Data	A set of values of qualitative or quantitative variables; restated, data is individual pieces of information.
Data Mining	A field at the intersection of computer science and statistics, it attempts to discover patterns in large datasets.
Data Warehouse	A storage architecture designed to hold data extracted from transaction systems, operational data stores and external sources. The warehouse then combines that data in an aggregate, summary form suitable for enterprise-wide data analysis and reporting for predefined business needs.
Digital Scholarship Initiative	A grassroots online forum for the sharing of news, ideas, and expertise related to all aspects of digital scholarship in the Greater Toronto Area and beyond.

Exabyte	Is a number representing 10^{18} bytes of digital information.
Gbit/s	A measure of the amount of data that can be transferred across a given network per second, in this case in gigabits per second, in other words billions of bits per second.
GPS	The Global Positioning System (GPS) is a space-based satellite navigation system that provides location and time information in all weather conditions, anywhere on or near the Earth where there is an unobstructed line of sight to four or more GPS satellites.
High Performance Computer	A class of computer that is at the frontline of contemporary processing capacity – particularly speed of calculation which can happen at speeds of a tiny fraction of a second while handling massive amounts of data concurrently. HPC machines are also known by an earlier name called ‘supercomputers’.
ICE	Intergovernmental Committee for Economic Development and Labour Force Development
Internet	A global system of interconnected computer networks that use a standard means to communicate and subsequently serve billions of users worldwide. It is a network of networks that consists of millions of private, public, academic, business, and government networks, of local to global scope, that are linked by a broad array of electronic, wireless, and optical networking technologies. The Internet carries an extensive range of information resources and services, such as the documents found on the World Wide Web (WWW) and the infrastructure to support email, among others.
Internet-of-Things	The Internet of Things (IoT) refers to uniquely identifiable objects and their virtual representations in an Internet-like structure. The IoT is seen as the next major extension to the Internet with machines talking to machines, such as sensor networks, that will generate large quantities of big data.
Interoperability	Is what allows things to work together, such as mobile phones from different manufacturers all working on different telecommunication operators’ networks.

Machine Learning	Machine learning is a branch of artificial intelligence and is concerned with the construction and study of systems that can learn from data (e.g., a machine learning system could be trained about the structure and content of email messages in order to learn to distinguish between spam and non-spam messages. After learning, such a capability can then be used to classify new email messages into spam and non-spam folders.)
Middleware	Middleware is computer software that provides services to software applications beyond those available from the operating system. Middleware makes it easier for software developers to perform communication and input/output, allowing for more efficient application development and operations.
Mobility Data	Data associated with and derived from mobile devices such as cell phones.
Natural Language Processing	Natural language processing is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human languages, such as English or French. It is related to the area of human–computer interaction. Many challenges in this field involve natural language understanding that is enabling computers to derive meaning from human or natural language input and in natural language generation.
NGO	Non-governmental organization
Open data	Open data refers mainly to the concept that certain data – particularly public data - should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control.
Optical Network	Fiber-optic communication is a method of transmitting information from one place to another by sending pulses of light through an optical fibre.
ORION	The Ontario Research and Innovation Optical Network (ORION) is a high-speed optical research and education (R&E) network infrastructure in Ontario. Established in 2002, ORION connects virtually all of Ontario's research and education institutions, including every university, most colleges, several teaching hospitals, public research facilities and several school boards to one another and to the

global grid of R&E and using optical fibre using CANARIE as a bridge to the world. (See www.orion.on.ca)

Petabyte

1 PB = 100000000000000 bytes or 10^{15} bytes which equals 1000 terabytes where 1 TB = 1000000000000 bytes or 10^{12} bytes which is equal to 1000 gigabytes, the size of memory in modern desktop and high end laptop computers.

Predictive Analytics

Predictive analytics encompasses a variety of statistical techniques from modeling, machine learning, and data mining that analyze current and historical facts to make predictions about future, or otherwise unknown, events.

Privacy by Design

'Privacy by Design' is a concept that was developed by Ontario's Information and Privacy Commissioner, Dr. Ann Cavoukian, back in the 90's, to address the ever-growing and systemic effects of Information and Communication Technologies, and of large-scale networked data systems. (See: www.privacybydesign.ca)

Public Good

"...[goods] which all enjoy in common in the sense that each individual's consumption of such a good leads to no subtractions from any other individual's consumption of that good..." by economist Paul A. Samuelson In his classic 1954 paper *The Pure Theory of Public Expenditure*.

Relational Database

A relational database is a database that has a collection of tables of data items, all of which is formally described and organized according to a particular mathematical model. For more on relational databases see: http://en.wikipedia.org/wiki/Relational_database.

Sensor Networks

A sensor network comprises a group of tiny, typically battery-powered devices and wireless infrastructure that monitor and record the state of conditions in any number of environments (e.g., factories, pipelines such as potable and gray water systems, power grids to hospitals, among others). The sensor network generally connects to the Internet, various kinds of private networks, or specialized industrial networks so that collected data can be transmitted to back to a centralized system for analysis and for use in applications.

SMEs

Stands for Small and Medium-sized Enterprises. Industry Canada defines a small business as one with fewer than 100 employees (if the

business is a goods-producing one) or fewer than 50 employees (if the business is service-based), and a medium-sized business as one with fewer than 500 employees.

SOSCIP

The Southern Ontario Smart Computing Innovation Platform (SOSCIP) is an Ontario-based research consortium established in April 2012. The consortium pairs academic and industry researchers with high performance computing to analyze big data within agile computing, health, water, energy and cities. The consortium members include the IBM Canada Research and Development Centre in Toronto as well as seven Ontario universities, led by University of Toronto and the University of Western Ontario. Other participants include McMaster University, Queen's University, University of Ontario Institute of Technology, University of Ottawa, and the University of Waterloo. (See: <http://soscip.org/>)

Structured Data

Data that resides in a fixed field within a record or file is called structured data, examples include spreadsheets and databases organized in columns and rows, so-called relational databases.

Terabyte

The terabyte is a multiple of the unit byte for digital information. The prefix 'tera' represents the fourth power of 1000, and means 10^{12} .

Text Mining

Refers to the process of deriving high-quality information from digitally recorded text.

Thematic Map

A thematic map is a type of map or chart especially designed to show a particular theme connected with a specific geographic area. Such maps can portray data associated with any kind of spatial relationships such as physical, social, political, cultural, economic, sociological, agricultural, or any other aspects of a city, state, region, nation, or continent.

Tweet

A text message associated with the social network called Twitter.

Unstructured Data

Refers to information that either does not have a pre-defined structure or is not organized in a pre-defined manner. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts. Written report, letters, memos, presentations, videos and the like are all examples of unstructured data.



Velocity of Data Transfer

The frequency and speed by which data is generated, captured, and shared.