2015-09-09

DRDC-RDDC-2015-L345

Produced for: DRDC CSS Executive

Scientific Letter

Big Data for Safety and Security

Background

Purpose

The intent of this Scientific Letter is to inform members of the executive of Defence Research and Development Canada (DRDC) Centre for Security Science (CSS) of developments in the domain of 'big data' and the associated implications for the Canadian Safety and Security Program (CSSP).

Introduction

Big data is a term reflecting the massive amounts of electronic data that computer networking as well as other applications produce every day and which may be put towards analysis and decision-making [1,2]. While much big data inevitably remains "raw" or "unstructured" at any given time, algorithmic approaches termed "big data analytics" are increasingly being devised to extract "signal" from the "noise". There is growing interest in exploring how advances in big data analytics can help obtain new insights in a variety of domains, including safety and security. Big data and big data analytics accordingly provide an emerging capability to obtain real-time, and even predictive, understanding of phenomena and problems that might otherwise remain nebulous, "wicked," or comprehensible only long after the fact or with more uncertain sampling procedures [1,3].

The ability to quickly and efficiently make sense out of big data is also an increasingly essential component of the policy intelligence cycle for addressing local, national, and global challenges [4,5]. It is foreseen that in coming years, more public and private sector decisions will be based on big data analytics rather than on experience, intuition, or statistical sampling, making decisions for the most part more rigorous and evidence based [2]. Advanced analytical capabilities and new data sources will also inform program and policy development, evaluation, compliance, efficiencies, and information sharing [6]. As such, it is important for the CSSP to continue to fully comprehend and make use of the opportunities of big data and its analysis to advance its mission.

While big data holds opportunities it also presents some risks. Institutional risks presented by big data and big data analytics include:

- The need to keep tabs on an exponentially growing amount of data as well as its variety and evolving forms and uses, which requires expense as well as technological expertise and relationships;
- Risks of false positives, in the form of patterns or statistical correlations mistakenly or deliberately imposed, as well as false negatives caused by the skewing of available data towards more digitally connected regions or sectors [1]. This risk entails the ongoing need for human judgement, institutional oversight and policies concerning big data analytics, including policies to protect individuals from aggregate-decision making based on big data analytics [1,2,7]; and
- Privacy, accessibility, intellectual property and other ethical issues around the storage, sharing, and use of big data [8]. Framework policies [1] are essential to maintain the security, privacy, and ethical use of big data and big data analytics capabilities as shortfalls may lead to the loss of public trust, and, in turn, of value [7].

Definitions

Big data typically refers to large and complex data sets requiring advanced and unique data storage, management, analysis, and visualization technologies [9]. To some degree the definition of big data is subjective as well as evolving, as there is no fixed threshold where data becomes big data and the definition varies over time and from sector to sector [1]. Big data is sometimes characterized in terms of "the four Vs": exponentially growing *Volume,* diverse *Variety,* including unstructured and previously unstructurable material; high *Velocity* in real time; and the need for *Veracity* of data and analyses to warrant decision-makers' and citizens' confidence and trust [10].

Big data analytics are computer algorithms and artificial intelligence techniques, including natural-language processing, sentiment analysis [7], pattern recognition, machine learning, and computer vision [7], used to seek correlations or patterns in data from one or more datasets that help to explain what is going on and why, thereby improving the evidentiary basis of decision-making [10]. Examples of analytical techniques include *predictive analytics*, which identify correlations and patterns among data sets to hone in on particular sites of risk and benefit for intervention, *causal analytics*, which use information from data sets to model the outcomes of different approaches to a problem, thereby improving decision-making [11], and *network analytics,* which is used to study the dynamic nature of social networks [12].

Sources of Big Data

Sources of big data can include, but are not limited to: electronic communications [9] and the associated metadata [13]; digital images and video [7]; user generated content on social media and other websites; financial transactions; search engines, clickstreams, and log files [14]; solicited crowdsourced data [1]; contents of cloud computing [7]; deep packet inspection data [14]; flight, immigration, and other types of records and logs [1]; digitized overhead video and imagery from satellites and UAVs [15]; digitized libraries and archives [16]; GPS/GIS readings from tracking devices [17]; readings from sensors in the Internet of Things (IoT) ecosystem, including smart meters, industrial control system data as well as data provided by environmental and infrastructural sensors [10,14]; radio-frequency identification tags [9]; somatic data from health and fitness devices [18]; and biometrics obtained from wearable technology, smartphones or other technologies which recognize physical features such as faces, irises,

retinas, DNA, and fingerprints as well as behavioral elements such as voices, signatures, gaits and keystrokes [19,20]. A supplementary source of big data is the linking of diverse data sources, for instance from public sector and private sector sources or from sources across agencies [1].

Methodology

Background research conducted to inform the contents of this Scientific Letter report consisted of related internet searches and review of recent non-academic literature. An academic literature review is outside the scope of the present work.

In order to understand the current landscape and the level of exploitation of big data and big data analytics in the program of work supported by the CSSP, DRDC CSS portfolio managers were interviewed by the authors.[1] The interviews followed an informal, semi-structured format meaning that while they were guided by a set of prepared questions allowance was made for deviations from the script.

The following series of questions guided the interviews:

1. Are big data and/or big data analytics relevant to your portfolio?
2. Are there any past or current projects within your portfolio that exploit big data or related analytical approaches?
3. How do you see big data and big data analytics influencing your portfolio in the future? (Related query: If not already, do you see big data becoming a priority area within your portfolio?).

Statement of Results

The results of interviews with CSS portfolio managers are summarized in Annex A. The majority of the portfolio managers interviewed indicated some relevance of big data to their respective portfolios. In some cases, big data featured high on the portfolios' science and technology (S&T) priority list, while in other cases it was described as an emerging priority. While the interviews also intended to capture past and current CSSP investments in big data; the resultant list of related projects shown in the table includes not only projects that are truly representative of the big data domain, but also projects which the portfolio managers flagged as potentially evolving into incorporating big data approaches. Finally, the 'Future Direction/Project' column in the table captures trends in science and technology that the portfolio managers envision influencing the direction of their respective portfolios towards incorporating big data.

Discussion of Results

Big data is fast emerging as a means of addressing a variety of safety and security problems, a development reflected across the CSSP. On the safety side, as societal resilience and community well-being gain traction as key areas of focus in the psycho-social domain, researchers are increasingly turning to big data to inform their investigations into human behaviour and psychological health. Not surprisingly, big data has become a focal point in emergency management as well, where the challenge surrounds how to aggregate, integrate, and filter the vast amounts of information now being generated from such diverse sources as social media, weather forecasts, risk assessments, and sensors.

[1] It is important to note that, outside of portfolio managers, other CSSP stakeholders, including for example Federal safety and security departments, were not consulted to inform the present work.

Projects are also underway to upgrade first responder services for the wireless era, including interactive IP-based 911 dispatching ('Next-Generation 911'), a dedicated broadband network for first responders; and interoperable communication hubs that will integrate the activity of police, paramedic, and fire services and mobilize preventative health services for high risk individuals. These projects are being designed to accommodate streaming big data sources such as wireless telecommunication feeds, video from body-worn cameras and mobile phones, biometric inputs for identification, access management and health applications, and environmental, optical, and magnetic sensors for updating onsite conditions. Additionally, a new project is examining how to draw on open source intelligence (OSINT) from social media to anticipate and identify violent extremism and other forms of criminality.

Many of the data sources from current CSSP community safety projects will prospectively be suitable for filtering, aggregating, and integrating data to produce real time, retrospective, and predictive analysis for optimizing operational decisions, situational awareness, resource use, and resilience across the community safety sector. In certain projects underway or recently completed, such as the national aggregation and broader linking of fire service and paramedic service reports, ordinary data rather than big data is the object for collection and analysis, but developments in using big data may hold insights for how this material can be analyzed to create operational insights. In view of increasingly complex data management requirements, a number of CSSP projects are also currently developing procedures and governance principles for archiving, accessing, and sharing data and big data across agencies and with coalition partners.

On the security side, numerous current CSSP projects are studying how to use big data for modelling, tracking, and identifying different types of security risks and threats, with further avenues to leverage big data analytics for security applications currently opening up. Risks of natural hazard such as flood, fire, and earthquakes can increasingly be predicted using weather and geological data. Risk modelling of radiological events relies on huge data inputs such as plume dispersion models and meteorological surveys. Border security is beginning to use biometric algorithms to scan for faces of known individuals. E-security makes use of meta-data analysis and deep packet inspection to identify breaches. Faster identification of strains of biological weapons or threats is enabled by data analytics on large databases of relevant information and will be further permitted by genomic analysis of biological strains. Social media is also an increasingly significant source of big data for security risk reduction, providing a basis for investigating the activities and intentions of individuals or groups with access to, or threatening, the use of explosive devices. Agencies responsible for monitoring nuclear non-proliferation are likewise looking to use big data from social media as well as traditional media to monitor for signs of non-compliance by nations and groups.

Conclusion

Successful big data applications are developed out of well-brokered relationships between operational communities and technology communities. They start from awareness of operational questions that need to be answered, scrutiny into which of these questions cannot be answered using available techniques, and a well-informed overview of the opportunities for insight increasingly being made available by big data streams and big data analytics. Successful big data applications further require willingness to invest in inquiry, expertise, and infrastructure necessary to leverage the data streams and to refine the analytical tools necessary to answer those questions [9].

As an internationally-networked program that brokers science and technology innovation for public safety and security communities, the CSSP is well positioned to ensure big data applications are being developed and operationalized in the Canadian context and to provide advice, best practices, and governance principles for their use. Canada in general is well-placed to develop cutting-edge big data research due to its advanced digital sector and well-functioning digital infrastructure. However, increasingly focused and integrated policy and procedures for big data research and investment in all stages of the technology innovation process are essential for Canada's emergence as a leader in this domain [5].

Prepared by: Suzanne Waldman, Kate Kaminska, Brian Greene . (DRDC — Centre for Security Science).

References

[1] Deputy Minister Committee on Policy Innovation, "DMCPI Discussion Paper on Big Data", 2014. http://www.gcpedia.gc.ca/wiki/DMCPI_Discussion_Paper_on_Big_Data.

[2] S. Lohr, "The Age of Big Data," New York Times, February 11, 2012. http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html.

[3] T. Harford, "Big Data: Are We Making a Big Mistake?" Financial Times Magazine, March 28, 2014. http://www.ft.com/intl/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html.

[4] National Research Council Canada, National Research Council Multimedia Analytic Tools for Security program website. http://www.nrc-cnrc.gc.ca/eng/solutions/collaborative/mats.html

[5] Social Sciences and Humanities Research Council, "Capitalizing on Big Data: Toward a Policy Framework for Advancing Digital Scholarship in Canada," Tri-Council consultation document, 2013. http://www.sshrc-crsh.gc.ca/about-au_sujet/publications/digital_scholarship_consultation_e.pdf.

[6] Shared Services Canada website. http://ssc-spc.gc.ca/pages/itir-triti/itir-triti-afac-030615-pres1-eng.html.

[7] [7] S. Siewert, "Big Data in the Cloud," IBM article, 2013. http://www.ibm.com/developerworks/library/bd-bigdatacloud/index.html.

[8] P. Dussault, MP. "Privacy and Social Media in the Age of Big Data," Report of the Standing Committee on Access to Information, Privacy and Ethics," April 2015. http://www.parl.gc.ca/content/hoc/Committee/411/ETHI/Reports/RP6094136/ethirp05/ethirp05-e.pdf.

[9] IBM, "Optimizing the Business of Government: Big Data & Analytics" Discussion paper prepared for Senator Oliver, 2013. http://www.gcpedia.gc.ca/gcwiki/images/c/c3/Big_Data_Analytics_Discussion_Paper_Prepared_for_Sen_Oliver.pdf.

[10] A. McAfee and E. Brynjolfsson, “Big Data: The Management Revolution,” Harvard Business review, October 2012. http://ai.arizona.edu/mis510/other/Big%20Data%20-%20The%20Management%20Revolution.pdf.

[11] A. Goldfarb, “Data, Big Data, Innovation, and Productivity,” paper requested for Senator Donald H. Oliver, 2013. http://www.gcpedia.gc.ca/gcwiki/images/5/55/BigData_Goldfarb_for_Oliver.pdf.

[12] H. Chen et al., “Business Intelligence and Analytics: From Big Data to Big Impact,” MIS Quarterly, Vol. 36, No. 4, 2012. http://hmchen.shidler.hawaii.edu/Chen_big_data_MISQ_2012.pdf.

[13] R. King. “U.S. Government Spending on Big Data to Grow Exponentially,” BiometricUpdate.Com, August 9, 2013. http://www.biometricupdate.com/201308/u-s-government-spending-on-big-data-to-grow-exponentially.

[14] SQLstream Website, “What is Machine Data?” http://www.sqlstream.com/products/what-is-machine-data/.

[15] M. Underwood, “Big Data Meets the Little Drone,” Syncsort Blog, February 20, 2014. http://blog.syncsort.com/2014/02/big-data-meets-the-little-drone/.

[16] L. Johnston, “Defining the “Big” in Big Data,” May 17, 2012, Library of Congress website. http://blogs.loc.gov/digitalpreservation/2012/05/defining-the-big-in-big-data/.

[17] E. Van Rees, “Big Data Analysis and Location: Leveraging Big Data at ESRI.” Geoinformatics, March 2013. http://fluidbook.geoinformatics.com/GEO-Informatics_2_2013/#/22/.

[18] J. Lee. “Biometrics combined with big data will impact the healthcare industry,” BiometricUpdate.Com, March 17, 2015. http://www.biometricupdate.com/201503/biometrics-combined-with-big-data-will-impact-the-healthcare-industry.

[19] M. Walsh, “Rising to the “Big Data” Challenge in Biometrics,” Global Strategies Group Blog, March 2015. http://www.globalgroup.com/newsblog/blog/rising-to-the-big-data-challenge-in-biometrics/.

[20] J. Sadowski, “Biometrics are coming for you,” Al Jazeera America, July 6, 2014. http://america.aljazeera.com/opinions/2014/7/biometrics-big-datamining.html.

Annex A Big Data vis a vis CSSP Portfolios

	Portfolio	Relevance to Portfolio	Past/Current Related Projects	Future Direction/Projects
Community Safety Directorate	Psychosocial	One of the portfolio's strategic priorities	• Statscan survey of Resilience (CSSP-2013-TI-1048) • Open Source Intelligence (CSSP-2015-TI-2143)	• Exploring big data's role in understanding psycho-social indications and resilience; • Online/offline social dynamics, social media analysis.
	Fire	Emerging, but not yet a priority. (Focus on 'small' data')	• National Fire Database (CSSP-2015-CP-2107)	• Linking fire data sets to social and crime stats to analyze for correlation and target prevention activities;
	Law Enforcement	Emerging	• Body Worn Video (CSSP-2014-TI-2031 and CSSP-2015-TI-2197) • Evidentiary Data, Cloud and Analytics (CSSP-2015-CP-2106) • HUB Community mobilizing (CSSP-2015-TI-2181) • Open Source Intelligence (CSSP-2015-TI-2143)	• Predictive analytics from current data sources (e.g., BWV, evidence in the cloud); • Automated image recognition (e.g., license plate, facial).
	Paramedic	Emerging, but not yet a priority	• Paramedic "Mega" Database – proof of concept (CPRC 02-3049) • Point of Care Testing (CSSP-2015-CP-2110)	• Remote patient monitoring via wearable technology; • Data to extract 'intelligence' on economics of community paramedicine.
	Emergency Management and Disaster Resilience	Emerging	• Social Media in Emergency Management (CSSP-2013-TI-1034)	• Ecosystem approach to big data; • Next-gen 911 (including video) analytics; • Smart sensors; • Analytical engines to support real-time analytics.
	Wireless	Current/Emerging	• Next Generation 9-1-1 over Public Safety LTE	• Transmitting and processing data from

			(CSSP-2015-TI-2187) • Public Safety Response Broadband (CSSP-2015-CP-2103)	Internet of Things devices, sensors; • Extracting consumable metadata; • Analytics for performance management and operational understanding.
	Communications Interoperability	Emerging	• Next Generation 911 (CSSP-2013-CP-1016, CSSP-2013-CP-1021)	• Data security, access and identity management; • Extracting information to make it actionable or intelligible for situational awareness.
Security Directorate	eSecurity	Very relevant	• Countering Security Threats Using Natural Language Technology (CSSP-2013-CP-1031)	• Metadata analysis and deep packet inspection.
	Natural Hazards	Very relevant	• Canadian EO Satellite Based Flood Monitoring System (CSSP-2015-CP-2100) • Severe Wildfire Risk Prediction (CSSP-2015-CP-2112)	• Seismic predictions.
	Explosives	Not directly relevant	None	• Can be utilized to uncover intent of individual/groups through semantic analysis of social media.
	Radiological and Nuclear	Very relevant	• Canadian National Nuclear Forensics Capability Pilot (CSSP-2012-TI-1119) • Canada-U.S. Interop for Airborne Gamma Surveying (CSSP-2013-CD-1129) • Atmospheric Dispersion from RDDs (CSSP-2013-CD-1131) • Field validation of novel algorithms for imaging (CSSP-2013-CD-1133) • High-fidelity, multi-scale atmospheric dispersion modeling of natural, accidental or malicious releases of toxic agents in the atmosphere (CSSP-2013-CP-1013) • A Mission-Ready Compton Gamma Imager for	• Analyzing social media and traditional media streams to track indications of non-proliferation compliance. • Pervasive radiation detection involving the integration of millions of monitors. • Cargo screening. • Modelling nuclear accidents.

		Safety and Security (CSSP-2015-CP-2096)	
Border and Transportation Security	Emerging	• Tailorable Operational Picture System (CSSP-2015-CP-2091) • Arctic Maritime Awareness for Safety and Security (CSSP-2015-CP-2092)	• Automated targeting.
Chemical	Not relevant	N/A	N/A
Biological	Relevant and potentially extremely useful	• Centre for Emerging Zoonotic Disease Integrated Intelligence and Response (CSSP-2013-CP-1022) • Understanding Antimicrobial Resistance Using a Complex Adaptive Systems Approach (CSSP-2015-CP-2098) • Canadian Network for Public Health Intelligence (CSSP-2015-CP-2099) • The Development of International Best Practices for Microbial Forensics (CSSP-2015-TI-2153) • Integrated Microbiology Testing Laboratory Network (CSSP-2015-TI-2157)	• Identifying biological strains using large databases of relevant information. • Sequencing genomes of hazardous strains.
Critical Infrastructure	Relevant	None	N/A
Security, Intelligence and Interdiction	Relevant	None	N/A