Big data and knowledge

11th Association of Parliamentary Librarians of Asia and the Pacific Conference
National Assembly Library—Seoul—Korea 26-28 April 2017

Dr Fiona Allen
Executive Information Officer | Australian Parliamentary Library

Big data holds forth a promise that we can finally know with certainty. Proponents of big data write of its ability to solve complex social policy problems; contribute to breakthroughs in medical research; combat terrorism; and build better societies, amongst a host of other possibilities.

This paper begins with a definition of big data and examines its inherent characteristics before discussing the implications of the phenomenon for parliamentary libraries and research services, primarily in terms of access to big data, technology for storing large data sets, methodologies for examining data, and data curation. In so doing, it questions whether parliamentary libraries should seek to be, or even can be, repositories for, and manipulators of, big data.

The paper proceeds to ask significant questions about the nature of big data, by way of an examination of what is more broadly at stake in a research environment where big data methodologies and datasets are privileged to the exclusion of other approaches to knowledge. A more nuanced approach by parliamentary libraries and research services to the value and necessity of big data would seem advisable.

Big data—the physical and the conceptual

How big is big? In terms of data, of the big type, big is in excess of what we might be able to conceive—in size and complexity. With its derivation from the Greek for 'monster', big data comes in units measured with the prefix *tera*; a 1 followed by twelve zeroes. Big also reaches into *petabytes*; a 1 followed by fifteen zeroes, otherwise called a quadrillion bytes of data (roughly equivalent to 16 million photographs). And driven as it is by the growing capacity of technology, with the development of quantum processors, it will not stop there. Big is meaningful only in relation to that which the concept presupposes: small, itself a notion similarly with no fixed meaning except by way of comparison. Proceeding as such, a *gigabyte*, the unit of measurement that defines most consumer electronics is a 1 followed by a mere nine zeroes—or one thousand million bytes of data.[1] Size, even of the exponential type we are discussing here, though, is only one characteristic of big data.

Conceptually, big data encompasses a change in what is known about a particular phenomenon or object and thus makes claims to a complexity created by not just one single data source, but by the aggregation of multiple datasets that may hitherto have been considered too disparate for such an undertaking. These datasets are both structured (in the case of databases created by logs; RFID systems; financial transactions; information from wearable devices; geolocation information; medical tests; records from sensor-equipped infrastructure in manufacturing, energy, agriculture, and transportation; and other types of automated record keeping), and unstructured like email messages, social media posts, photographs, the

[1] L Gitelman & V Jackson, 'Introduction', in Lisa Gitelman (ed.), *Raw Data is an Oxymoron*, The MIT Press, Cambridge, 2013, p. 1.; D Bollier, 'The promise and peril of big data', *The Aspen Institute Communications and Society Program*, The Aspen Institute, Washington D.C. 2010, p. 1.; S González-Bailón, 'Social Science in the Era of Big Data', *Policy and Internet*, 5(2), 2013, pp. 147-160, p. 147.; A Gandomi & M Haider, 'Beyond the Hype: Big Data Concepts, Methods, and Analytics', *International Journal of Information Management*, 35, 2015, pp. 137-144, p. 138.

content of internet forums, or phone transcripts. Unstructured data constitutes as much as 95 per cent of big data. Thus, a key characteristic here is the imprecise and uncertain nature of much of this data—what is to be made of it; what does it mean? Often collected and stored by automated processes, big data is also characterised by the rate at which it is produced. By virtue of its size and complexity, big data is often beyond the capacity of traditional data processing tools; it requires an infrastructure of specialised storage devices and servers to facilitate large-scale analysis; and to elicit anything from it at all, big data demands the creation of new taxonomies, methodologies, and algorithms—the essence of which may be beyond our ability to understand, being a consequence of machine-to-machine interactions and the application of complicated mathematical processes too complex for human calculations absent artificial computing power.[2] The nature of this computational turn in thought—in its substantive and mathematical guises—is something to which I will return later in the paper.

Big data—the applications

Regardless of these challenges, big data's utility in resolving virtually any policy problem is predicted. It is being used now to explore patterns of crime; identify public health trends; examine economic behaviour; track communication dynamics; understand conflict and violence; predict and contest elections (with varying degrees of success); manage and control the physical world through machines, factories, and other infrastructure; deal with traffic congestion; and combat terrorism. Big data is a part of medical and scientific research; climate and weather analysis; the operation of financial markets; government intelligence gathering; online and traditional commerce; the planning of public infrastructure; and, the provision of medical, educational, and welfare services, amongst myriad other things. Its proponents make claims as to big data's ability to provide new insights with 'accuracy and objectivity'[3] to tackle longstanding problems previously believed too complex, for instance, how political movements originate and spread. For its proponents, big data makes it possible to study the 'collective intellectual space of the whole planet,'[4] to examine how ideas emerge, diffuse, burst, die, and interlink. The size of big data appears matched only by the size of its potential utility.[5]

---

[2] L Taylor & R Schroeder, 'Is bigger better? The emergence of big data as a tool for international development policy', *GeoJournal*, 80, 2014, p. 503/518.; P-C Huang, 'When big data gets small', *The International Journal of Organizational Innovation*, 8(2), 2015, pp. 100-117, p. 105.; J Kallinikos & ID Constantiou, 'Big data revisited: A rejoinder', *Journal of Information Technology*, 30, 2015, pp. 70-74, pp. 70-71.; M Bieraugel, 'Big data', *Keeping up with …* American Library Association, Chicago, accessed 4 January 2016.; D DeLyser & D Sui, 'Crossing the qualitative-quantitative chasm III: Enduring methods, open geography, participatory research, and the fourth paradigm', *Progress in Human Geography*, 38(2), 2014, pp. 294-307, pp. 300-301.; D Oprea, 'Big Questions on Big Data', *Review of Research and Social Intervention*, 55, 2016, pp. 112-126, p. 117.; C Croft, 'Of note: The limits of big data', *SAIS Review*, 34(1), 2014, pp. 117-120, p. 118.; GC Bowker, 'Data flakes: An afterword', in Lisa Gitelman (ed.), *Raw Data is an Oxymoron*, The MIT Press, Chicago, 2013, p. 169.; Burkholder (1992) in, D Boyd & K Crawford, 'Critical Questions for Big Data', *Information, Communication & Society*, 15(5), 2012, pp. 662-679, p. 665.; Gandomi & Haider, 'Beyond the Hype: Big Data Concepts, Methods, and Analytics', pp. 139, 143.

[3] DeLyser & Sui, 'Crossing the qualitative-quantitative chasm III: Enduring methods, open geography, participatory research, and the fourth paradigm', p. 301.

[4] L Manovich, 'Trending: The promises and the challenges of big social data', *manovich*, Lev Manovich, New York, accessed 4 January 2017.

[5] C Poulin, 'Big data custodianship in a global society', *SAIS Review*, 34(1), 2014, pp. 109-116, pp. 109, 115.; G Marcus & E Davis, 'Eight (no, nine!) problems with big data', *The New York Times*, 7 April, 2014, p. A23.; Huang, 'When big data gets small', pp. 107-108.; Kallinikos & Constantiou, 'Big data revisited: A rejoinder', p. 71.; Taylor & Schroeder, 'Is bigger better? The emergence of big data as a tool for international development policy', pp. 503, 507.; H Margetts & D Sutcliffe, 'Addressing the policy challenges and opportunities

Big data—the questions for parliamentary libraries and research services

Having thus far determined the nature of big data and explored its potential, one is lead to consider what an era of big data means for parliamentary libraries and research services—particularly as it relates to the big question of the institutional platforms that will collect, manage, store, and analyse the 'tsunami' of big data, and for what purpose.[6] Further, parliamentary libraries and research services provide significant resources and research to inform the decisions of policymakers and how we grapple with the big data era, in all its facets, will have far-reaching implications. Although there is an absence of discussion around where parliamentary libraries and research services sit, there is a growing body of literature on the role of university libraries in big data, where university libraries are positioned as the logical repositories and curators for the data-intensive research carried out by their faculties, which then might be made available to researchers more broadly in the interests of public-benefit research. This directionality is one of libraries responding to the pre-existence of large datasets constructed in the process of intensive research activity—the ongoing viability, aggregation, and usability of which might not be possible with existing faculty resources—not one of libraries independently creating big data or of manipulating it. It is the management of research data, its storage (physically in servers and computationally in aggregated and relational databases), curation (creating and applying taxonomies, designating metadata standards, and systematising retrieval methods to ensure interoperability), preservation (including forward and backward compatibility), and dissemination that is considered a professional imperative for academic libraries and there is much in this literature that discusses the resources, skills, funding, and infrastructure necessary for libraries to become repositories for big data.[7]

Such is the interest that the International Federation of Library Associations and Institutions (IFLA) published a webinar in 2016, encompassing a number of issues including the big data opportunities available to libraries, the questions libraries have to ask and answer about their motivations and capacities, and the role of libraries in making data more accessible through visualisation products.[8] The organisation further devoted a special issue of the *IFLA Journal* to research data services.[9] In fact, there is much attention across the literature to libraries considering their ability to access necessary physical and

---

of 'big data'', *Policy and Internet*, 5(2), 2013, pp. 139-146, pp. 139-140.; Bollier, 'The promise and peril of big data', 2010, pp. 33, 40.; D Ribes & SJ Jackson, 'Data bite man: The work of sustaining a long-term study', in Lisa Gitelman (ed.), *Raw Data is an Oxymoron*, The MIT Press, Chicago, 2013.; H Grassegger & M Krogerus, 'The Data That Turned the World Upside Down', *Das Magazin*, accessed 31 January 2017.; S Mutula, 'Editorial Feature - Big data industry: Implications for the library and information sciences', *African Journal of Library, Archives and Information Science*, 26(2), 2016, pp. 93-96.; World Economic Forum, 'Industrial Internet of Things: Unleashing the Potential of Connected Products and Services', *Industry Agenda*, World Economic Forum, Geneva 2015.

[6] K Prewitt, 'The 2012 Morris Hansen Lecture: Thank you Morris, et al., for Westat, et al.', *Journal of Official Statistics*, 29(2), 2013, pp. 223-231, p. 229.

[7] M McLure, AV Level, CL Cranston, B Oehlerts, & M Culbertson, 'Data curation: A study of researcher practices and needs', *Libraries and the Academy*, 14(2), 2014, pp. 139-164.; C Schubert, Y Shorish, P Frankel, & K Giles, 'The evolution of research data: Strategies for curation and data management', *Library Hi Tech News*, 30(6), 2013, pp. 1-6.; M Steeleworthy, 'Research data management and the Canadian academic library: An organizational consideration of data management and data stewardship', *Partnership: Canadian Journal of Library and Information Practice and Research*, 9(1), 2014, pp. 1-11.; J Ray, 'The rise of digital curation and cyberinfrastructure: From experimentation to implementation and maybe integration', *Library Hi Tech*, 30(4), 2012, pp. 604-622.; Bieraugel, Big data, American Library Association.

[8] 'Big data: New roles and opportunities for new librarians', *IFLA New Professionals Special Interest Group Webinars*, International Federation of Library Associations and Institutions, accessed 21 March 2017.

[9] M Witt & W Horstmann (eds), 'Research Data Services', *IFLA Journal*, 43(1), 2017.

intellectual resources, identifying the needs of their clients, and to life-cycle planning, prior to pursuing big data. Effectively maintaining massive repositories of data requires a long chain of coordinated action.[10] These considerations are no less relevant for parliamentary libraries and research services, in particular the question about the data parliamentary libraries and research services might curate, for what purpose, and for whom.

More specifically, however, there are suggestions librarians may become data scientists themselves, capable of not only acquiring, curating, preserving, and disseminating, but also of manipulating big data itself.[11] Whilst the scenario of a librarian or researcher calmly wielding an array of analytical instruments to adeptly examine a mass of data before coming to a conclusion that can be neatly encapsulated and handed off to the nearest lawmaker may come to mind, a significant characteristic of big data is the absence of established analytical instruments with which it might be manipulated. Indeed, the very complexity upon which its value is said to rest means there can be no one analytical instrument. Different data require different questions to be asked, a situation that requires different analytical instruments; different instruments identify different patterns; and, different patterns and the researcher's chosen mode of their expression elicit different insights. Big data does not give one answer or exist to answer one problem. Indeed, a big data set, despite its size, may not have an answer for a particular problem. It is a mass of information, aggregated in complex ways, within which researchers seek to find meaningful patterns. Researchers have not yet written the complex algorithms to effectively examine large portions of the big data that currently exists and developing these tools is beyond the capacity of a single person working in isolation. Instead, teams of computer scientists, statisticians, economists, social scientists, anthropologists, psychologists, and more, are needed to build tools specific to the task at hand—to identify meaningful correlations, test theories, and progress to models. An exponentially increasing number and variety of analytic tools is necessary to keep pace with big data sets provided by technological advances. The inescapable conclusion is parliamentary libraries and research services, often resource-poor, will have to reconfigure themselves if they seek to "do" big data.[12]

Big data—the data speaks for itself or the interpretations pile up

It is not beyond the realms of possibility, given the resources necessary, that a research service configured to collect and manipulate big data would do so at the expense of other modes of inquiry. Before committing to expending the significant resources required to do big data, parliamentary libraries and research services should consider its essential nature. And it is to this that the paper now turns; in the first instance to examine the notion of raw data, as big data is often claimed to be, its quality, and the manner

[10] Ribes & Jackson, 'Data bite man: The work of sustaining a long-term study', p. 152.
[11] W Klapwijk, 'The Library (Big) Data Scientist', *Big Data: New Roles and Opportunities for New Librarians - IFLA New Professionals Speical Ineterest Group Webinars*, International Federation of Library Associations and Institutions, 2016.
[12] GM Allenby, ET Bradlow, EI George, J Leichty, & RE McCulloch, 'Perspectives on bayesian methods and big data', *Customer Needs and Solutions*, 1, 2014, pp. 169-175, pp. 174-175.; Manovich, 'Trending: The promises and the challenges of big social data', pp. 7-8.; Ribes & Jackson, 'Data bite man: The work of sustaining a long-term study', p. 151.; Bollier, 'The promise and peril of big data', 2010, pp. 6-9.

in which it is analysed. It will proceed then to examine the perspective of big data. Like any approach to research, big data takes a particular epistemological and ontological standpoint which creates blind-spots and biases. The intention is not to suggest big data is of no utility, but rather to encourage a clear-eyed analysis of its strengths and weaknesses. This is relevant even if libraries and research services choose not to pursue big data capabilities and instead rely upon the big data research and analysis conducted elsewhere to inform their own activities.

A common perception of research based on quantitative data analysis is that it is objective, a notion drawn at least partly from the idea that it is based on 'raw data'—that is, information that has not been manipulated or interpreted in any way. The data speaks for itself, it is possible to come to a research conclusion without any kind of human intervention.[13] On a number of levels, this is a misguided assumption and in so far as it is the basis for claiming big data's value, it is misleading. Data does not simply exist, it has to be generated and it is here, before anything happens, that interpretation enters the equation, so to speak. Every discipline has its norms, classifications, and standards that create and structure data; that determine what counts as data and how it is presented. It is accurate to say 'Data are always already "cooked" and never entirely raw.'"[14] That data comes in a 'set' means some elements are included and others are considered irrelevant and excluded. Once collected (or even in some cases before it is collected), data is always cleaned or scrubbed—a process whereby the researcher decides which variables are important and which can be ignored, and these decisions are inevitably theory laden and perspectival, whether acknowledged or not. Data cannot generate value independently; researchers choose the data by deciding which to integrate and manipulate.[15] There is always an element of homogenization in this process, whereby alternative perspectives and outliers are discounted. The instance of social media is illustrative of a further issue—that data collected is subject to prior organisation according to a number of stylized and limited options, for example, tagging, sharing, following, and liking. This allows the recording of user participation so it can be counted, aggregated, and processed, but it also necessarily limits what counts as data and the options users have.[16] At each of these steps, interpretation is involved, even before the data is processed.

The challenges to the notion of objectivity do not cease there. Big data implies that data simply stack up in an aggregative sense; the more data, the better. The ability to analyse large amounts of data, rather than a handful of interviews or case studies, is said to be one of big data's key attributes.[17] Yet, size poses a

[13] C Anderson, 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete', *Wired*, accessed 25 January 2017.; L Rainie & J Anderson, 'Code-Dependent: Pros and cons of the algorithm age', *Internet, Science & Tech*, Pew Research Center, 2017, p. 49.
[14] Geoffrey Bowker in, Gitelman & Jackson, 'Introduction', p. 2.; Bowker, 'Data flakes: An afterword'.
[15] Bollier, 'The promise and peril of big data', 2010, p. 13.; Huang, 'When big data gets small', p. 114.; TD Williams, 'Procrustean Marxism and subjective rigor: Early modern arithmetic and its readers', in Lisa Gitelman (ed.), *Raw Data is an Oxymoron*, The MIT Press, Chicago, 2013, p. 41.; A Mantelero, 'Social Control, Transparency, and Participation in the Big Data World', *Journal of Internet Law*, April, 2014, pp. 23-29, p. 24.; Boyd & Crawford, 'Critical Questions for Big Data', pp. 666-669.; T Oliphant, 'A Case for Critical Data Studies in Library and Information Studies', *Journal of Critical Library and Information Studies*, (1), 2017, pp. 1-24, p. 8.
[16] Kallinikos & Constantiou, 'Big data revisited: A rejoinder', p. 73.
[17] Gitelman & Jackson, 'Introduction', p. 8.; A Halevy, P Norvig, & F Pereira, 'The unreasonable effectiveness of big data', *IEEE Intelligent Systems*, (March/April), 2009, pp. 8-12, p. 8.

number of problems. It does not equal representation. For instance, if you do not use social media, you will not exist in research that gains insights from social media use, research that may inform broader policymaking. There are approximately two billion facebook accounts, are these users representative of more than five billion non-users? But even within this data, does what one person consider a 'friend' on facebook represent what all people consider a friend? Are all social networks equal? Does everyone mean the same thing when they 'like' something? Epidemiological studies that make use of mobile phone data are hampered by the fact mobile phone usage can be highly differentiated by gender and income level. In many countries, subscriber populations for mobile phone companies are segmented by socio-economic status. It has yet to be established that Twitter users are more broadly representative of anything, or that there is commensurability between accounts.[18] In addition to questions about representation, size and complexity are significant challenges in and of themselves. All data is subject to unreliability, every source is error prone—data always contains gaps and mistakes and these are magnified when aggregated. Sometimes these errors go unnoticed as machines collect the data from other machines, aggregate it, and then pass it into complex algorithms for analysis. It can be difficult even to determine the quality of data, particularly if it is proprietary data, meaning the bias is not understood, quantified, or accounted for prior to analysis.[19]

The aggregation of complex data sets itself involves researchers making assumptions about 'sameness' within changing situations and circumstances. For instance, environmental, human, and infrastructure conditions are continually changing, researchers make decision on whether and how data is commensurate under these conditions. The process of aggregation is subject to opinion and perspective as researchers determine how to combine, for example, datasets containing social media posts, photographs, conversation streams, and geolocation data in order that they may be evaluated mathematically in some kind of meaningful way. The act of quantifying a blog post, for instance, is embedded in interpretation; interpretation that may take it out of its original context, assign an alternative meaning, or analyse it in a manner never envisioned by the author.[20] Researchers are inevitably making assumptions about the content of data before it even sees an algorithm, including the assumption that the aggregation can elicit something meaningful.

Once it reaches the processing stage, big data is subject to further interpretation with regard to the analytical tool with which it is to be aggregated and interrogated. Analytical tools are run by algorithms,

[18] Taylor & Schroeder, 'Is bigger better? The emergence of big data as a tool for international development policy', pp. 506, 510, 513.; Boyd & Crawford, 'Critical Questions for Big Data', p. 669.

[19] Prewitt notes the existence of significant research about errors in census or survey data, such as sampling error, respondent burden, cognitive bias, imputation, response rates, external validity, attrition, and panel studies, and statisticians have established methods to account for these errors. He suggests there is no such generally accepted understanding of what constitutes errors in big data, or of ways the effects of errors may be ameliorated. Prewitt, 'The 2012 Morris Hansen Lecture: Thank you Morris, et al., for Westat, et al.', p. 230.; DeLyser & Sui, 'Crossing the qualitative-quantitative chasm III: Enduring methods, open geography, participatory research, and the fourth paradigm', p. 301.; Marcus & Davis, 'Eight (no, nine!) problems with big data'.; Bollier, 'The promise and peril of big data', 2010, p. 13.; Taylor & Schroeder, 'Is bigger better? The emergence of big data as a tool for international development policy', p. 504.

[20] Ribes & Jackson, 'Data bite man: The work of sustaining a long-term study', p. 148.; Boyd & Crawford, 'Critical Questions for Big Data', p. 672.

which are complex mathematical equations that instruct a computer to perform certain calculations to solve a problem. Algorithms, in and of themselves, are neither bad nor good (though there are examples where algorithms have had unintended consequences), and were all algorithms to cease working, the world as we know it would also cease—they are embedded in every piece of equipment containing a processor; the internet runs on them; modern manufacturing and industry would be impossible without their existence. Complex logic is also employed to manipulate big data, this logic contains 'if' and 'else' statements and Boolean triggers. The issue at point is algorithms and complex logic are created, they reflect the biases of programmers and datasets, they contain embedded judgements, are generally subjective, and they are dependent on the limitations of the data they process. Algorithms are the consequence of and impose a certain type of logic-driven perspective which, taken to its extreme, undermines sophisticated human reasoning and the value of local intelligence. It is not unusual for algorithms to categorise human beings as 'inputs' to a process, rather than complex beings capable of an array of thoughts, feelings, desires, hopes and perspectives—a process of homogenization and dehumanization. No dataset can capture the complexity and fullness of a human life.[21]

The instance of big data predictive policing goes some way to providing a practical example of the interpretation endemic to the creation and analysis of big data, demonstrating the fallacy of the claim that correlation supersedes causation because the data can speak for itself.[22] Predictive policing, of significant interest to some law and policymakers, is a collective term encompassing a variety of analytical tools and subsequent law enforcement practices that claim the ability to identify where crime will occur and direct the deployment of law enforcement officers accordingly. The underlying model is based upon technical, organisational, social, and ethical assumptions across the entire breadth of the predictive policing cycle. Within the data collection phase, for instance, it is assumed that the data is an accurate representation of the criminal activity in a specific area. However, there is always a gap between crimes committed and reported and this is not random but systemic and linked to, amongst other things, the type of crime and the characteristics of the victim, and the existing bias of law enforcement officers. As such, crimes omitted from data collection will go ignored or marginalised into the future. Further, how crimes are categorised or classified, or even if law enforcement officers decide they are sufficiently serious to be recorded varies—this issue is amplified when data sets are combined. There is also a feedback loop where more police being sent to a particular area is likely to increase the reporting of crime. The algorithm itself can thus become implicated in the process of prediction. Because the subsequent actions of law

[21] Taylor & Schroeder, 'Is bigger better? The emergence of big data as a tool for international development policy', p. 509.; Bollier, 'The promise and peril of big data', 2010, pp. 11-12.; Rainie & Anderson, 'Code-Dependent: Pros and cons of the algorithm age', 2017, pp. 4, 9-11.;S Frank, 'Power and Paranoia in Silicon Valley - Come With Us If You Want To Live: Among the Apocalyptic Libertarians of Silicon Valley', *Harper's Magazine*, January, 2015, pp. 26-36.

[22] R Kitchin & TP Lauriault, 'Small data in the era of big data', *GeoJournal*, 80, 2015, pp. 463-475, p. 471.; Anderson, The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.

enforcement officers affects that which is being measured, crime, the predictive accuracy of the program and its effectiveness cannot be measured concurrently.[23]

In the data analysis phase, all predictive policing software is based upon a particular model of crime—for instance, one where crime rates in different areas at different times are assumed to be constant. Further, the variables assumed relevant depend upon the information typically collected about crime, which in turn is dependent upon how much is known about a particular crime. Some variables are omitted because they are expensive or difficult legally to collect. The algorithm used to process the data, based as it is upon a particular theory of policing or assumed model of crime, will identify correlations between a particular feature and a probability of offending—between a person demonstrating a particular characteristic, or between communities located in certain neighbourhoods. Policing based upon profiling has been shown in many instances to be counterproductive and ineffective; big data analysis cannot identify why a person will commit a crime. Understanding the causes of crime in a particular area or the causes of particular crimes is a process fraught with interpretation, particularly if one is relying upon police-collected and other aggregated data sets to the exclusion of less easily quantified types of information and interaction. And the assumptions continue through the police operations phase and the criminal response phase. This small example though, is sufficient to illustrate that the data cannot speak for itself and to show how, through the complex interrelation between assumptions, data, and algorithms, that it is possible, indeed probable, that big data calculations have the capacity to contribute to or create a situation they claim to be merely predicting.[24]

Big data—the correlations

It is particularly important here to emphasise that the assumption-based analytical software that processes big data does not identify causes; it identifies correlations, or patterns where particular variables can be observed concurrently. However, this process is fraught with the difficulties associated with identifying correlations, and determining if they are meaningful; bearing in mind an initial assumption in big data is that the attributes of previously unrelated datasets are comparable, and can reveal something of significance about a particular issue. In a study examining the use of big data in political campaigns, it was argued that the ability to uncover correlations was highly dependent on the talent of the particular data analyst and their familiarity with the properties of the datasets. Researchers admit each time they analyse a collection of big data, they will find different patterns.[25] However, there is no guarantee these patterns or correlations are meaningful. If a researcher looks one hundred times for correlations between two variables, they will risk, according to one analyst, finding, by chance, five bogus ones that appear

[23] LB Moses & J Chan, 'Algorithmic Prediction in Policing: Assumptions, Evaluation, and Accountability', *Policing and Society*, November, 2016, pp. 1-17.
[24] Moses & Chan, 'Algorithmic Prediction in Policing: Assumptions, Evaluation, and Accountability'.; CK Citron & FA Pasquale (2014) in, Moses & Chan, 'Algorithmic Prediction in Policing: Assumptions, Evaluation, and Accountability', p. 14.
[25] Marcus & Davis, 'Eight (no, nine!) problems with big data'.; DW Nickerson & T Rogers, 'Political Campaigns and Big Data', *Journal of Economic Perspectives*, 28(2), 2014, pp. 51-74, p. 59.; Manovich, 'Trending: The promises and the challenges of big social data', pp. 7-8.

statistically significant. The massive size of datasets itself results in a greater number of correlations.[26] Instances of spurious correlations can be amusing: the number of people who trip over their own two feet and die correlates with the number of lawyers in the US state of Nevada; the sharp increase in diagnoses of autism is correlated with increased sales of organic food; and in the United States, the murder rate between 2006 and 2011 was correlated with the market share of the internet browsing software, Internet Explorer.[27] It is indeed the case that analysis of large data sets can identify correlations that would be otherwise missed, but they cannot determine if these correlations are meaningful.

Big data—the questions not asked

A general reliance on the predictive capability of big data means there is less attention paid to difference, to alternative conceptions and systems of meaning and signification, to deep nuance, to the idea that meaning is not always simple or stable, things are not always self-evidently so. Further, when used in a law or policymaking context; such an orientation encourages the quantification of policymaking—a reliance on indicators, ranking schemes, performance metrics, and accountability measures which themselves are perspectival. How much can you glean from meeting a quantitative indicator concerned with a physical amount of work about the quality of the work and the broader consequences of conceiving quality in this manner.[28]

The nature of big data is such that certain questions cannot be asked or answered. There are two points to be made with regard to this: the first goes to the nature of the data itself; the second to the underlying presuppositions made in quantitative approaches to research. With relation to the nature of the data, two cases illustrate the limitations of big data analysis, particularly with regard to nuance. Chris Poulin examined publicly published tweets to understand the dynamics of protests that began in Egypt in early 2011—part of the so-called Arab Spring. Through a process of keyword analysis, Poulin suggested researchers could predict dissent and regime change. However, Christine Croft countered that Poulin was overstating the capacity of big data and looking for answers in places where none existed. She suggested it was overly simplistic to reduce the dynamic of events in Egypt to measuring how often certain words were used on Twitter. Keyword analysis cannot account for the dialogue fuelled by each tweet; it cannot account for the cumulative effect of a movement; it cannot perceive intensity of emotion; it cannot predict free will.[29] Moving beyond Croft's criticism, an analysis of Twitter cannot account for the conditions of possibility—what was happening on the ground, what were the complex factors feeding into the discontent, what was the diversity of opinion, what were the myriad factors motivating each individual. An analysis of tweets cannot understand the importance of contextual factors like government

[26] J Fan & J Lv (2008) in, Gandomi & Haider, 'Beyond the Hype: Big Data Concepts, Methods, and Analytics', p. 143.
[27] T Vigen, 'Number of People Who Tripped Over Their Own Two Feet and Died', *Spurious Correlations,* Spurious Media, accessed 16 April 2017.; Marcus & Davis, 'Eight (no, nine!) problems with big data'.
[28] Croft, 'Of note: The limits of big data', p. 120.; Taylor & Schroeder, 'Is bigger better? The emergence of big data as a tool for international development policy', p. 504.; Bowker, 'Data flakes: An afterword', p. 168.; Prewitt, 'The 2012 Morris Hansen Lecture: Thank you Morris, et al., for Westat, et al.', p. 225.
[29] Poulin, 'Big data custodianship in a global society'.; Croft, 'Of note: The limits of big data'.

policies and actions, the state of the economy, the weather, the support of broader societal groupings, and more. Further, it is not clear how much of the substance of the twitter feed was influenced by the medium—the requirement to limit characters, to use hashtags, to gather followers, to truncate language and emotion. Nuance, when one is seeking certainty, muddies the waters.

A more straightforward example is provided by Danah Boyd—mobile phone data suggests workers spend more time with their colleagues than their partners. What one makes of this statistic is unclear; it does not necessarily follow that colleagues are more important than spouses because mobile phone data cannot provide an insight into the quality of the relationship between a person and their work colleagues and their spouse; it cannot account for the nature of the interactions in each context.[30] As noted above, big data can find subtle correlations that may have previously escaped the notice of researchers, but it cannot evaluate whether they are significant or understand the causal mechanism.[31]

Big data—the presuppositions of the assumptions

I want to now speak to an issue that comes prior to big data—and that is the presuppositions with relation to ontology, epistemology, and methodology that inform all approaches to research, big data and otherwise. It is an issue that will not be resolved here, it is endlessly debated in research departments, and it is the basis of schisms in the academy, particularly in the social sciences in politics and international relations where my education was embedded. However, appreciating the issue is important for how we conduct research, evaluate our own work, and write analysis for lawmakers and other policymakers. Big data is underpinned by what might be called a positivist epistemology—that is, it begins with a supposition that there are objects, processes or phenomena in the world that can be perceived as they exist; the central claim of which is that we can have a collection of facts about them that is objectively true. The research process starts with ontology (the phenomena that exist in our world), proceeds through epistemology (how we formulate and evaluate statements about phenomena in the world in order that we build knowledge about them), and ends with methodology (the tools we use to ensure our knowledge is accurate). By proceeding in this manner it gives precedence to the phenomena or objects we are studying—that is, it is the intrinsic nature of these objects or phenomena that determines what is possible to know about them.[32] This seems to be quite a common sense proposition.

For instance, we want to know about the phenomena of radicalisation, so we observe it to perceive its characteristics. Epistemologically, to do this observation, we assume a process whereby we formulate a theory and establish an hypothesis about the type of people who become radicalised and collect data to test this hypothesis. This data might be information on the characteristics of those who are radicalised and include, age, socio-economic status, residence, friends, religion, their views and opinions, their behaviours, their frustrations, and so on. Methodologically, we might conduct a particular statistical

[30] Boyd & Crawford, 'Critical Questions for Big Data', p. 671.
[31] Croft, 'Of note: The limits of big data', p. 119.; Marcus & Davis, 'Eight (no, nine!) problems with big data'.
[32] PT Jackson, *The Conduct of Inquiry in International Relations: Philosophy of Science and Its Implications for the Study of World Politics*, Routledge, Oxon, 2011, pp. 26-27, 31.

calculation on this data to test the hypothesis before coming to conclusions we might claim are scientifically robust and sound about the types of behaviour that correlate with radicalisation. At its foundation, this approach is based upon an assumption that knowledge production (or research) is separate from phenomena existing in the world and through our research we bridge the gap between what we think and what exists, in a verifiable way, to come out with a sound understanding of radicalisation.

The separation between empirical and theoretical propositions, and the actual character of phenomena that exist in the world, is a central presupposition to research methodologies for big data, because research allows us to bridge the gap between what we know and what is existentially true. It would be fair to say this is a common perception in the big data sector. A high-level round table discussion moderated by The Aspen Institute with participants from McKinsey & Company, Harvard Law School, Deloitte, Google, Creative Commons, and IBM, amongst others, showed participants saw research tools as the means to bridge the gap between what is in the world (referred to in this instance as 'ontologies') and what we objectively know about it.[33]

However, these presuppositions are not uncontested, and neither are they the only way to approach research—an alternative definition of ontology leads a researcher to take an entirely different approach; it is one that holds that no such separation is ever possible between what is in the world, and what we think about things in the world. The world is endogenous to (or already embedded within) the social practice of knowledge production, and theory, rather than being applied after the fact, is implicated in the process of understanding. What we think is 'already and inevitably' intertwined with what we perceive in the world. We cannot get at the world as it really exists because there is no way to see the world absent this entire apparatus—there is no such thing as a view from no-where because theory is practice.[34]

I want to now return to my example of radicalisation earlier to illustrate how this perspective changes what we find through research. Big data analysis might be able to suggest a likelihood that someone holding certain views, demonstrating certain behaviours, and with certain experiences will become radicalised. This analysis begins by assuming these actions are meaningful and relevant and that they can be interpreted in a non-problematic way. Yet, it can only be possible to think about radicalisation in this way if there is a whole system of meanings that make it a sensible proposition, for instance, that there is a notion of the legitimate use of force; that there is such a thing as a mainstream perspective, and further, that it is not radical; that the 'rule of law' is a concept with a fixed meaning experienced in the same way by everyone; that state sovereignty is a legitimate notion; that 'normal' people behave in a certain way; and even that democracy is all that it is claimed to be.

[33] Bollier, 'The promise and peril of big data', 2010, pp. 4-9.
[34] Jackson, The conduct of inquiry, pp. 35-36.; J George, *Discourses of Global Politics: A Critical (re)Introduction to International Relations*, Lynne Rienner, Boulder, 1994, pp. 3, 104, 177.; WE Connolly, 'Method, Problem, Faith', in Ian Shapiro, Rogers M Smith, & Tarek E Masoud (eds.), *Problems and Methods in the Study of Politics*, Cambridge University Press, Cambridge, 2004.

Rather than identifying behaviours, assigning numerical values, and running an algorithm, an alternative way to look at radicalisation is to examine the way radicalisation is understood by lawmakers and by so-called 'radicalised' people. Through what kind of perspective do these groups see the world, how do they interpret events, what meanings do they attribute to certain phenomena. This kind of approach might allow you to understand 'why' rather than solely 'what'. It might allow you to understand why, when so many people might hold similar views, only a small number choose to act differently. Or in the case of the Arab Spring, instead of being limited to tracing the flow of tweets, you might be able to ask why certain messages gained traction, or how was the vast array of information processed by individual protestors.[35] This kind of approach does not excuse violence, neither does it suggest you can make up the meaning of anything you want. It is too simplistic to say 'one man's terrorist is another man's freedom fighter,' or suggest people are ignorant of the facts. Neither is it accurate to suggest one is just choosing their own facts. What an approach such as this suggests is that we cannot perceive phenomena in the world absent the mind, the view from which is inevitably coloured by myriad factors and experiences, but even prior to this, what we think is only possible by what is thinkable.

This approach is not amenable to big data calculations. In a simplistic sense, big data might be able to identify 'what' but it has difficulty identifying 'why.' More specifically, big data is capable of identifying 'what' within a certain framework of meaning, that established by the researcher and within the limitations and biases of the data set. Big data never gets at the world objectively, it gets at the world through a perspective, as all research does. What we claim to know with certainty is always contingent, every perspective is political.[36] This is an important consideration for parliamentary research services, beyond the issue of big data, because it challenges the idea that embracing the facts makes one objective, or steering a straight course between the Left and the Right, is an act of apolitical impartiality. Further, it has implications for policymaking—the questions asked and the answers proffered are the basis upon which decisions are made. There are many ways to conduct research, to answer what, why, and how—diversity in research methods and a reflection on the implications of knowledge production is essential for any research service. In the race to embrace big data, this should not be forgotten.

Big data—the way to where?

So where does all this leave parliamentary libraries and research services? Big data is an encompassing phenomenon that permeates every aspect of our daily existence. It is certainly the case big data has been at the core of scientific, technological, and social breakthroughs, but amongst the celebrations, there is little indication of widespread recognition of its very real limitations and biases, or even more broadly a

[35] An example of the analysis of tweets where the outcome is limited to the tracing of information flows, rather than the provision of any substantive understanding of why certain messages gained traction is provided by: G Lotan, E Graeff, M Ananny, D Gaffney, I Pearce, & D Boyd, 'The Revolutions Were Tweeted: Information Flows During the 2011 Tunisian and Egyptian Revolutions', *International Journal of Communications*, 5, 2011, pp. 1375-1405.

[36] RK Ashley, 'Living on Border Lines: Man, Poststructuralism, and War', in James Der Derian & Michael J Shapiro (eds.), *International/Intertextual Relations: Postmodern Readings of World Politics*, Lexington Books, Lexington, 1989.; Oliphant, 'A Case for Critical Data Studies in Library and Information Studies', p. 16.

questioning of its appropriateness as an analytical tool in every instance. The answer cannot always be found in big data; there may not in fact be an answer to some things; or there may not be an answer that can be succinctly elucidated and expressed, implying as it would an underlying existential simplicity and order to the social world for which is there is no evidence. It is certainly the case that parliamentary libraries and research services need to examine where they sit on becoming involved in big data. If libraries choose to become custodians and manipulators of big data, they will need to reorganise themselves to provide the infrastructure, systems and processes, and skilled personnel, and this will likely come at a cost of diversity. As noted by Danah Boyd, big data reframes what we think of as knowledge, the process of research, how we engage with information, and the categorization of reality.[37]

It is more likely that parliamentary libraries and research services will be users of the products of big data—we will use studies, academic articles, and reports that are produced through big data analysis conducted by others. This makes a critical perspective even more crucial: one cannot evaluate the quality of research unless one understands the substance of the assumptions and methods by which it was produced. Any method, any theory, any approach to understanding the social world, is a partial view from somewhere, and has very real implications for the type of world in which we live, particularly if it is believed to be one that can guarantee certainty. None of this is to suggest research based on big data is of no utility, libraries and research services should examine big data research, but from the perspective of an enquiring mind that is aware of the endless interpretations, the blind spots, the questions not asked or capable of being asked. This is true not only of big data research but any research.

[37] Boyd & Crawford, 'Critical Questions for Big Data', p. 665.

Bibliography

Allenby, Greg M, Eric T Bradlow, Edward I George, John Leichty, & Robert E McCulloch, 'Perspectives on Bayesian Methods and Big Data', *Customer Needs and Solutions*, 1, 2014, 169-175.

Anderson, Chris, 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete', *Wired*, accessed 25 January 2017.

Ashley, Richard K, 'Living on Border Lines: Man, Poststructuralism, and War', in J Der Derian and MJ Shapiro (eds.), *International/Intertextual Relations: Postmodern Readings of World Politics*, Lexington Books, Lexington, 1989.

Bieraugel, Mark, 'Big Data', *Keeping up with ...* American Library Association, Chicago, accessed 4 January 2016.

'Big Data: New Roles and Opportunities for New Librarians', *IFLA New Professionals Special Interest Group Webinars,* International Federation of Library Associations and Institutions, accessed 21 March 2017.

Bollier, David, 'The Promise and Peril of Big Data', *The Aspen Institute Communications and Society Program*, The Aspen Institute, Washington D.C., 2010.

Bowker, Geoffrey C, 'Data Flakes: An Afterword', in L Gitelman (ed.), *Raw Data Is an Oxymoron*, The MIT Press, Chicago, 2013, pp. 167-171.

Boyd, Danah & Kate Crawford, 'Critical Questions for Big Data', *Information, Communication & Society*, 15(5), 2012, 662-679.

Connolly, William E, 'Method, Problem, Faith', in I Shapiro, RM Smith and TE Masoud (eds.), *Problems and Methods in the Study of Politics*, Cambridge University Press, Cambridge, 2004, pp. 332-349.

Croft, Christine, 'Of Note: The Limits of Big Data', *SAIS Review*, 34(1), 2014, 117-120.

DeLyser, Dydia & Daniel Sui, 'Crossing the Qualitative-Quantitative Chasm III: Enduring Methods, Open Geography, Participatory Research, and the Fourth Paradigm', *Progress in Human Geography*, 38(2), 2014, 294-307.

Frank, Sam, 'Power and Paranoia in Silicon Valley - Come with Us If You Want to Live: Among the Apocalyptic Libertarians of Silicon Valley', *Harper's Magazine*, January, 2015, 26-36.

Gandomi, Amir & Murtaza Haider, 'Beyond the Hype: Big Data Concepts, Methods, and Analytics', *International Journal of Information Management*, 35, 2015, 137-144.

George, Jim, *Discourses of Global Politics: A Critical (Re)Introduction to International Relations*, Lynne Rienner, Boulder, 1994.

Gitelman, Lisa & Virginia Jackson, 'Introduction', in L Gitelman (ed.), *Raw Data Is an Oxymoron*, The MIT Press, Cambridge, 2013, pp. 1-14.

González-Bailón, Sandra, 'Social Science in the Era of Big Data', *Policy and Internet*, 5(2), 2013, 147-160.

Grassegger, Hannes & Mikael Krogerus, 'The Data That Turned the World Upside Down', *Das Magazin*, accessed 31 January 2017.

Halevy, Alon, Peter Norvig, & Fernando Pereira, 'The Unreasonable Effectiveness of Big Data', *IEEE Intelligent Systems*, (March/April), 2009, 8-12.

Huang, Po-Chieh, 'When Big Data Gets Small', *The International Journal of Organizational Innovation*, 8(2), 2015, 100-117.

Jackson, Patrick Thaddeus, *The Conduct of Inquiry in International Relations: Philosophy of Science and Its Implications for the Study of World Politics*, Routledge, Oxon, 2011.

Kallinikos, Jannis & Ioanna D Constantiou, 'Big Data Revisited: A Rejoinder', *Journal of Information Technology*, 30, 2015, 70-74.

Kitchin, Rob & Tracey P Lauriault, 'Small Data in the Era of Big Data', *GeoJournal*, 80, 2015, 463-475.

Klapwijk, Wouter, 'The Library (Big) Data Scientist', *Big Data: New Roles and Opportunities for New Librarians - IFLA New Professionals Speical Ineterest Group Webinars*, International Federation of Library Associations and Institutions, 2016.

Lotan, Gilad, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, & Danah Boyd, 'The Revolutions Were Tweeted: Information Flows During the 2011 Tunisian and Egyptian Revolutions', *International Journal of Communications*, 5, 2011, 1375-1405.

Manovich, Len, 'Trending: The Promises and the Challenges of Big Social Data', *manovich,* Lev Manovich, New York, accessed 4 January 2017.

Mantelero, Alessandro, 'Social Control, Transparency, and Participation in the Big Data World', *Journal of Internet Law*, April, 2014, 23-29.

Marcus, Gary & Ernest Davis, 'Eight (No, Nine!) Problems with Big Data', *The New York Times,* 7 April, 2014, p. A23.

Margetts, Helen & David Sutcliffe, 'Addressing the Policy Challenges and Opportunities of 'Big Data'', *Policy and Internet*, 5(2), 2013, 139-146.

McLure, Merinda, Allison V Level, Catherine L Cranston, Beth Oehlerts, & Mike Culbertson, 'Data Curation: A Study of Researcher Practices and Needs', *Libraries and the Academy*, 14(2), 2014, 139-164.

Moses, Lyria Bennett & Janet Chan, 'Algorithmic Prediction in Policing: Assumptions, Evaluation, and Accountability', *Policing and Society*, November, 2016, 1-17.

Mutula, Stephen, 'Editorial Feature - Big Data Industry: Implications for the Library and Information Sciences', *African Journal of Library, Archives and Information Science*, 26(2), 2016, 93-96.

Nickerson, David W & Todd Rogers, 'Political Campaigns and Big Data', *Journal of Economic Perspectives*, 28(2), 2014, 51-74.

Oliphant, Tami, 'A Case for Critical Data Studies in Library and Information Studies', *Journal of Critical Library and Information Studies*, (1), 2017, 1-24.

Oprea, Dumitru, 'Big Questions on Big Data', *Review of Research and Social Intervention*, 55, 2016, 112-126.

Poulin, Chris, 'Big Data Custodianship in a Global Society', *SAIS Review*, 34(1), 2014, 109-116.

Prewitt, Kenneth, 'The 2012 Morris Hansen Lecture: Thank You Morris, Et Al., for Westat, Et Al.', *Journal of Official Statistics*, 29(2), 2013, 223-231.

Rainie, Lee & Janna Anderson, 'Code-Dependent: Pros and Cons of the Algorithm Age', *Internet, Science & Tech*, Pew Research Center, 2017.

Ray, Joyce, 'The Rise of Digital Curation and Cyberinfrastructure: From Experimentation to Implementation and Maybe Integration', *Library Hi Tech*, 30(4), 2012, 604-622.

Ribes, David & Steven J Jackson, 'Data Bite Man: The Work of Sustaining a Long-Term Study', in L Gitelman (ed.), *Raw Data Is an Oxymoron*, The MIT Press, Chicago, 2013, pp. 147-166.

Schubert, Carolyn, Yasmeen Shorish, Paul Frankel, & Kelly Giles, 'The Evolution of Research Data: Strategies for Curation and Data Management', *Library Hi Tech News*, 30(6), 2013, 1-6.

Steeleworthy, Michael, 'Research Data Management and the Canadian Academic Library: An Organizational Consideration of Data Management and Data Stewardship', *Partnership: Canadian Journal of Library and Information Practice and Research*, 9(1), 2014, 1-11.

Taylor, Linnet & Ralph Schroeder, 'Is Bigger Better? The Emergence of Big Data as a Tool for International Development Policy', *GeoJournal*, 80, 2014, 503/518.

Vigen, Tyler, 'Number of People Who Tripped over Their Own Two Feet and Died', *Spurious Correlations,* Spurious Media, accessed 16 April 2017.

Williams, Travis D, 'Procrustean Marxism and Subjective Rigor: Early Modern Arithmetic and Its Readers', in L Gitelman (ed.), *Raw Data Is an Oxymoron*, The MIT Press, Chicago, 2013, pp. 41-59.

Witt, Michael & Wolfram Horstmann (eds), 'Research Data Services', *IFLA Journal*, 43(1), 2017.

World Economic Forum, 'Industrial Internet of Things: Unleashing the Potential of Connected Products and Services', *Industry Agenda*, World Economic Forum, Geneva, 2015.