Centre for Policy Studies

University College Cork
National University of Ireland

Working Paper Series

CPS WP: 18-001

Big data & official statistics

Steve MacFeely
Head of Statistics and Information
United Nations Conference on Trade and Development
Geneva
Switzerland

This Paper constitutes original research by the authors. It does not necessarily reflect the positions of the Centre for Policy Studies or University College Cork. Short sections of text may be quoted without permission for educational or non-profit purposes provided proper citation is given. No use of this publication may be made for resale or other commercial purpose without the written permission of the authors.

Big data & official statistics

Steve MacFeely
Head of Statistics and Information, United Nations Conference on Trade and Development, Geneva, Switzerland
Adjunct Professor, Centre for Policy Studies, University College Cork, Cork, Ireland

Abstract

Over recent years the potential of big data for government, for business, for society has excited much comment, debate and even evangelism. But are big data really the panacea to all the challenges facing official statistics or is this just hype and hubris? The question facing official statisticians is whether big data are worth the investment and expose to risk? While the statistical possibilities appear to be theoretically endless, in practice big data also present enormous challenges and potential pits-falls: legal; ethical; technical; and reputational. This paper examines the opportunities and challenges presented by big data. Some particular governance issues for official statistics are also discussed.

Keywords

Governance, administrative data, national statistical offices, national statistical systems,

Acknowledgements

My sincere thanks to Aidan Condron and Padraig Dalton of CSO in Ireland and to Steven Vale of UNECE in Switzerland for their helpful insights and comments.

Introduction

Over recent years the potential of big data for government, for business, for society has excited much comment, debate and even evangelism. Described as the 'new science' with all the answers (Gelsinger, 2012) and a paradigm destroying phenomena of enormous potential (Seth Stephens-Davidowitz, 2017) big data are all the rage. Official statisticians, already with a long history of using non-survey data, which are often very large in terms of volume, must decide whether big data is really something new, or just more of the same, only more so. One the one hand, some argue that 'Big Data needs to be seen as an entirely new ecosystem comprising new data, new tools and methods, and new actors motivated by their own incentives, and should stir serious strategic rethinking and rewiring on the part of the official statistical community' (Letouzé and Jütting, 2015: 5) whereas others argue to the contrary that big data is just hype and that 'Big Data, Small Data, Little Data, Fast Data, and Smart Data are all "Just Data"' (Thamm, 2017: 2). Was psychologist Dan Ariely correct when he tweeted[1] hilariously in 2013 that 'Big Data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it?' The purpose of this chapter is to examine these questions from the perspective of official statistics and outline some of the opportunities, challenges and governance and privacy issues that big data present.

Big data is the by-product of a technological revolution. In simplistic terms, one can think of big data as the collective noun for all of the new digital data arising from our digital activities. Our increasing day-to-day dependence on technology is leaving 'digital footprints' everywhere. Those digital footprints or digital exhaust offers official statisticians rich and tantalizing opportunities to augment or supplant existing data sources or generate completely new statistics. These digital data can be shared, cross-referenced, and repurposed as never before opening up a myriad of new statistical possibilities. Big data also present enormous statistical and governance challenges and potential pits-falls: legal; ethical; technical; and reputational. Big data also present a significant expectations management challenge, as it seems many hold the misplaced belief that accessing big data is straight-forward and that their use will automatically and dramatically reduce the costs of producing statistical information. As yet the jury is out on whether big data will offer official statistics anything especially useful. Beyond the hype of big data, and hype it may well be[2], statisticians understand that big data are not always better data and that more data doesn't automatically mean more insight. In fact more data may simply mean more noise. As Boyd and Crawford (2012: 668) eloquently counsel 'Increasing the size of the haystack does not make the needle easier to find.'

To understand the opportunities, challenges and governance issues involved with big data from the unique perspective of official statistics, it is useful to first define what we mean by big data, administrative data and official statistics. Thereafter the chapter will look at sources of big data, access issues before examining the opportunities and some of the challenges presented by big data, in particular confidentiality and privacy. Before concluding, the chapter will briefly outline some of the governance structures that National Statistical Offices (NSOs) and International Organisations (IOs) may need to consider putting in place if they intend to harvest big data for the purposes of compiling official statistics.

Defining Big Data

What are big data? While some, such as, Stephens-Davidowitz argue that big data is 'an inherently vague concept' (2017: 15) it is nevertheless important to try and define it. This is important, if only, to explain to readers that big data are not simply 'lots of data' and that despite the name 'big data' size is not the defining feature. So if not size, what makes big data big? One of the challenges in trying to answer this question is that 'There is no rigorous definition of "big data"' (Mayer-Schonberger and Cukier 2013: 6).

[1] https://twitter.com/danariely/status/287952257926971392?lang=en

[2] Buytendijk (2014) argued that big data had passed the top of the 'Hype Cycle' and was moving towards the 'Trough of Disillusionment' and that now expectations regarding the use of big data would become more realistic.

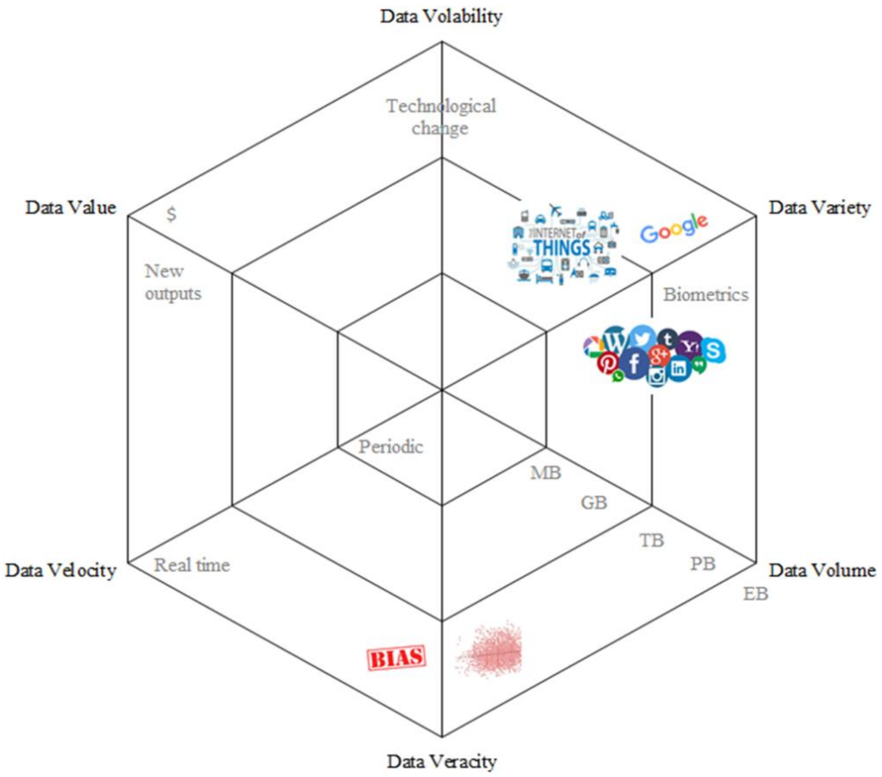
Gartner analyst Doug Laney provided what has become known as the three 'Vs' definition in 2001. He described big data as being high-volume, high-velocity, and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation. In other words, big data should be huge in terms of volume (i.e. at least terabytes), have high velocity (i.e. be created in or near real-time), and be varied in type (i.e. contain structured and unstructured data and span temporal and geographic planes). The European Commission (2014) definition of big data; 'large amounts of data produced very quickly by a high number of diverse sources' is essentially a summary of the 3V's definition. The Commission notes that big data can either be created by people or generated by machines, such as sensors gathering climate information, satellite imagery, digital pictures and videos, purchase transaction records, GPS signals and so forth.

It seemed that the 3Vs definition was generally accepted, within official statistics circles at least, with the United Nations Statistical Commission (2014: 2) adopting a very similar definition - 'data sources that can be described as; high volume, velocity and variety of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making.' But Tam and Clarke (2015: 3) have used a more general definition, simply describing big data as 'statistical data sources comprising both the traditional sources and new sources that are becoming available from the "web of everything".'

Over the intervening years, the '3Vs' has swollen to '10Vs.[3]' Perhaps more usefully, in 2017 Hammer et al. selected a 5V definition (the original 3Vs plus an additional two V's - volatility and veracity). Veracity refers to noise and bias in the data and volatility refers to the 'changing technology or business environments in which big data are produced, which could lead to invalid analyses and results, as well as to fragility in big data as a data source' (2017: 8). Volatility and veracity are extremely important additions, in particular for understanding the contribution that big data might make to compiling official statistics. Certainly the '5Vs' definition is more balanced and useful from an analytical perspective than the 3Vs as it flags some of the downside risks that prompted Borgman (2015: 129) to note that using big data is 'a path with trap doors, land mines, misdirection, and false clues.' But arguably a '6V' definition that includes 'value', where value means that something useful is derived from the data offers a better definition, introducing cost-benefit and yet striking a balance between parsimony and utility. The introduction of value is extremely important, as the costs of investing in big data must be carefully weighed up against what they might deliver in practical terms. See Figure 1.

[3] Volume, Velocity, Variety, Variability, Veracity, Validity, Vulnerability, Volatility, Visualisation and Value.

Figure 1 - The 6V's of Big Data for Official Statistics



In understanding big data from a statistical perspective, it is important to understand that like administrative data, big data is conceptually quite different to traditional survey data. 'Big Data sets are made available to us, rather than designed by us' (Dass et al., 2015: 256). They are a collection of by-product data rather than data designed by statisticians for a specific purpose. In other words the derivation of statistics is a secondary purpose. This difference is perhaps obvious but profoundly important. We find ourselves today in a situation, reminiscent of the storyline in 'The Hitchhikers Guide to the Galaxy' (Adams, 1979), where we now have the answers but we are still struggling to define the question. As big data is a by-product of our interactions with technologies that are evolving quickly, we must also accept that as a consequence big data are not a stable platform but a very dynamic one, and so any definition is likely to require further refinement over time.

Administrative Data

Many NSOs already make extensive use of administrative data. Blackwell (1985: 78) defined administrative or public sector data as ‘information which is collected as a matter of routine in the day-to-day management or supervision of a scheme or service or revenue collecting system.’ Similarly, UNECE (2011) defined administrative data ‘as collections of data held by other parts of government, collected and used for the purposes of administering taxes, benefits or services.’ In other words across public services, a huge volume of administrative records are collected, maintained and updated on a regular basis. These data pertain to the wide range of administrative functions in which the state is involved, ranging from individual and enterprise tax payments to social welfare claims or education or farming grants. Typically these administrative records are collected and maintained at the lowest level of aggregation, i.e. at transactions level. The interactions of individual taxpayers, applicants and recipients make these data very rich from an analytical perspective (MacFeely & Dunne, 2014). Brackstone (1987) identified four key distinguishing characteristics of administrative data: (1) the data are collected by an agency other than an NSO; (2) the methodology and processing are controlled by an agency other than an NSO; (3) the data were collected for non-statistical purposes; (4) the data have complete coverage of the target population.

Although administrative datasets are often very large, they are not typically considered big data, in the sense that they are not updated in real time (high velocity) and they tend to be relatively stable (low volatility). Although it is worth noting that in 1997, Eurostat, proposed a narrow and a broad definition of administrative data. The narrow view saw administrative data comprising of only public sector non-statistical sources whereas the wider definition included private sector sources (this would presumably include big data). The Conference of European Statisticians adopted a definition of administrative data consistent with this wider concept in 2000 (UNECE, 2000). Administrative and big data share some important characteristics: both are secondary data; both may suffer from problems with veracity; and neither is originally compiled for statistical purposes. But there are some important differences too: administrative data are typically national whereas big data are more likely to be supra-national or global[4]; big data are inherently more unstable and volatile than administrative data; and big data will in all likelihood compromise a greater variety of sources and types of data.

Defining Official Statistics

It is also useful to define official statistics. The purpose of official statistics are to provide 'an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honor citizens' entitlement to public information' (United Nations, 2014: Principle 1). This is a demanding and ever more complicated role, as in addition to measuring the traditional economic, social and environment dimensions, new ones such as, peace and security and welfare have emerged.

Official statistics can be national or international (or both). Official National Statistics are all statistics produced by the NSO in accordance with the Fundamental Principles of Official Statistics (United Nations, 2014), other than those explicitly stated by the NSO not to be official; and all statistics produced by the National Statistical System (NSS) i.e. by other national organisations that have been mandated by national government or certified by the head of the NSS to compile statistics for that specific domain. So in practice, if a NSO produces statistics they are, de facto, official unless stated otherwise. If another organisation within a NSS produces statistics, then they are typically also considered official.

[4] This is an important difference as a NSO may be able to influence the content or quality of national administrative data, whereas it is highly unlikely to be able to exert any influence over global datasets.

Official International statistics are statistics, indicators or aggregates produced by a UN agency or other international organisation in accordance with the Principles Governing International Statistical Activities formulated by the Committee for the Coordination of Statistical Activities (2014). It is often necessary for a UN agency, or other international organisation, to modify official national statistics that have been provided by an NSO or another organisation of a NSS, in order to harmonise statistics across countries or to correct evidently erroneous values. Furthermore, in the absence of an official national statistic, a UN agency or other international organisation may compile estimates. Thus, it is insufficient to define official international statistics as simply the reproduction of official national statistics.

Sources of Big Data

In a world where our day-to-day use of technology and applications are leaving significant ‘digital footprints’, it seems that just about everything we think or do is now potentially a source of data. Big data are being generated from a bewildering array of activities and transactions. Our spending and travel patterns, our online search queries, our reading habits, our television and movies choices, our social media posts - everything it seems now leaves a trail of data. Some examples – big data were generated by the 227.1 billion global credit/debit card purchase transactions made in 2015 (Nilson, 2018). The 7.7 billion mobile telephone subscribers in 2017 around the world (International Telecommunications Union, 2017) also unwittingly created big data every time they used their phone. In fact even when they didn’t use their phones, they were still generating data - according to Goodman (2015) mobile phones generate 600 billion unique data events every day. Every day we send 500 million tweets (Krikorian 2013), 8 billion snapchats (Aslam 2015), upload 1.8 billion images (Meeker, 2017) and conduct 3.5 billion google searches. Every minute of every day we upload 400 hours of video to YouTube (Taplin, 2017). Each one of these transactions leaves several digital footprints, from which new types of statistics can be compiled. In fact as Stephens-Davidowitz (2017: 103) explains, today 'Everything is data.' The torrent of by-product data being generated by our digital interactions is now so huge it has been described variously as a data deluge; data smog; info-glut or the original information overload. This deluge is also the result of an important behavioral change, where people now record and load content for free. Weigand (2009) described this phenomenon where people actively share or supply data directly to various social networks and product reviews and led to the evolution of the wiki model as a 'social data revolution'.

Not only have the sources changed, the very concept of data itself has changed - 'the days of structured, clean, simple, survey-based data are over. In this new age, the messy traces we leave as we go through life are becoming the primary source of data' (Stephens-Davidowitz, 2017: 97). Now data includes text, sound and images, not just neat columns and rows of numbers. Begging the question, in this digital age, how much data now exist. Definitional differences again make this a difficult question to answer, and consequently there are various estimates to choose from. Hilbert and Lopez (2012) estimated that 300 exabytes (or slightly less than one third of a zettabyte[5]) of data were stored in 2007. According to their 2017 Big Data factsheet, Waterford Technologies estimated that 2.7 zettabytes of digital data exist (Waterford Technologies, 2017)[6].

[5] A zettabyte is 10^{21} bytes (i.e. 1,000,000,000,000,000,000,000 bytes) or 1,000 exabytes or 1,000,000 petabytes

[6] It is not clear whether these estimates include data on the 'Deep Web' or 'Dark Net'. Goodman (2015) estimates that the Deep Web is 500 larger than the google-able 'Surface Web'.

Goodbody (2018) states that 16 zettabytes of data are produced globally every year and that by 2025 it is predicted that that estimate will have risen to 160 zettabytes annually. IBM now estimates we create an additional 2.5 quintillion bytes[7] of data every day (IBM, 2017).

Although the definitions and consequent estimates differ, it is clear that, a massive volume of digital data now exists. But as Harkness (2017: 17) wisely counsels, the 'proliferation of data is deceptive; we're just recording the same things in more detail'. Nor are all of these data necessarily accessible or of good quality. As Borgman (2015: 131) warns, big data must be treated with caution. A threat 'to the validity of tweets as indicators of social activity is the evolution in how online services are being used. A growing proportion of twitter accounts consists of social robots used to influence public communication….As few as 35 percent of twitter followers may be real people, and as much as 10 percent of activity is social networks may be generated by robotic accounts.' Furthermore, Goodman (2015) states that 25% of reviews on Yelp are bogus. Facebook themselves have admitted that 3% of accounts are fake and an additional 6% are clones or duplicates, the equivalent of 270 million accounts (Kulp, 2017). Taplin (2017) also states that 11% of display ads, almost 25% of video ads, and 50% of publisher traffic are viewed by bots[8] not people - 'fake clicks.'

There are also issues of coverage, as sizeable digital divides exist. For example, the International Telecommunication Union (2017) estimates that global Internet penetration is only 48% and global mobile broadband subscriptions 56%, although they are as high as 97% in the developed world. Although global coverage is improving rapidly, it still means that of 2017 almost half of the world's population does not use the web. Limited access and connectivity to the web or mobile phones is creating a data divide. Anyone excluded from web access or mobile phones will not have a digital footprint or at best, a rather limited one. Even within countries, digital divides exist arising from a range of access barriers: social; gender; geographic; or economic strata. This may lead to important cohorts being excluded, with obvious bias implications for statistics (see Struijs and Daas 2014) - hence the importance of 'Veracity.' The digital divide is creating a big data divide. To quote William Gibson (2003) 'The future is already here - it's just not very evenly distributed.' The question for NSOs is whether these data can be safely accessed and whether they are representative and stable enough to be used to compile official statistics.

Accessing and Using Big Data

Many big data are proprietary i.e. data that are commercially or privately-owned and not publically available. For example, data generated from the use of credit cards, search engines, social media, mobile phones and store loyalty cards are all proprietary and may not be available for use. Even if these data were publically accessible, sensitivities around their repurposing to compile official statistics must be carefully considered. 'Even if there are no legal impediments, public perception is a factor that must be taken into account' warn Daas et al. (2015: 257).

The current proprietary status of some data may change in the future as people around the world realise that their data are being used and traded. But for the moment many datasets are not currently accessible by NSOs, either because costs are prohibitive or proprietary ownership

[7] A quintillion bytes is 10^{18} bytes or 1 exabyte.

[8] Goodman (2015) refers to these bots as WMDs - Weapons of Mass Disruption.

makes it impossible. Changes to statistical legislation may be required to give NSOs or NSSs access to big data sources. MacFeely and Barnat (2007: 898) have argued that 'in order to future-proof statistical legislation, consideration should be given to mandatory access to all appropriate secondary data…where secondary data would be defined to include not only administrative or public sector data but also some important, commercially held data, such as for example, information on credit card transactions, information held by utilities or information regarding the movements of mobile phones.'

NSOs must be extremely careful not to damage their reputation and the public trust they enjoy. To do so, a NSO must ensure it does not break the law or stray too far outside the culturally acceptable boundaries or norms of their country. So an NSO must decide whether it is legally permissible, ethical or culturally acceptable to access and use big data. These are not always easy questions to answer. When it comes to accessing new data sources, the legal, ethical and cultural boundaries are not always clear-cut. In some cases NSOs may be forced to confront issues well before the law is clear or cultural norms have been established. Furthermore, given the speed with which the digital data world is changing, Rudder (2014: 250) notes, 'new laws will be drafted…but their letter will be outdated before the ink is dry.' This poses a challenge as public trust and reputation is fragile; hard won but easily lost. NSOs depend on the public to supply information to countless surveys and enquiries. If an NSO breaks that trust, they risk biting the hand that feeds them. Yet a progressive NSO must to some extent lead public opinion, meaning they must maintain a delicate balance, innovating and publishing new statistics that deal with sensitive public issues but without moving too far ahead of public opinion. For example, from a technical, statistical perspective the most logical and cost-effective method of deriving international travel and tourism statistics might be to use mobile phone data, but from a data protection and public opinion perspective using these type of data may not be acceptable.

This tension or trade-off does not appear to be well understood and is certainly not well reflected in many national and international policy documents. MacFeely (2017: 57-58) notes 'In an increasingly complex data protection environment, there is a growing but discernable mismatch between potential and actual, between expectations and reality.' He further notes ‘The rather fantastic talk of [big] data revolution does not seem to make any allowance for the complex legal and ethical issues that prevent access to many valuable data sources.' United Nations Economic Commission for Europe (2016) reflecting on their experiences, note 'High initial expectations about the opportunities of Big Data had to face the complexity of reality. The fact that data are produced in large amounts does not mean they are immediately and easily available for producing statistics. Data from mobile phones represent a notable example in this sense. It has been proved that such data can be exploited for a wide range of purposes, but they are still largely outside the reach of the majority of statistical organizations, due to the high sensitivity of the data.'

Opportunities for official statistics

There will almost certainly be opportunities in the future to compile official statistics in new and exciting ways. Assuming access problems can be overcome then big data offers the potential to improve official statistics in a number of ways. Big data may be used in conjunction with or as a replacement for traditional data sources to improve, enhance and complement existing statistics. Big data may offer new cost-effective or efficient ways of compiling statistics, improve timeliness or offer some relief to survey fatigue and burden. Big data also offers the tantalizing potential of being able to generate more granular or disaggregated statistics, allowing for more segmented and bespoke analyses, or the possibility of generating completely new statistics. That said, Kitchin (2015) sounds a cautionary note with regard to generating new statistics, noting that it is important that NSOs don't allow mission drift where big data drives their direction i.e. falling into the trap of measuring what is easy rather than what needs to be measured. Notwithstanding this risk, the opportunities presented by big data for official statistics can be summarized as entirely or partially replacing existing data sources with new big data sources to compile existing statistics in a more efficient, timely or more precise way or compiling entirely new statistics altogether.

Big data also offer the potential to compile datasets that are linkable, offering enormous potential to undertake more cross-cutting and dynamic analyses that may help us to better understand causation, offering the potential for more policy-relevant, outcome-based statistics. One of the short comings of many existing official statistics is that each statistic is compiled discretely, and typically derived from a sample. While this bespoke approach offers many advantages regarding bias, accuracy and precision, it has the disadvantage that as discrete data, those data cannot be easily connected or linked (other than at aggregate level) with other data. It is not always possible subsequently to construct a comprehensive analyses or narrative for many complex phenomena. As big data sets are more likely to have full or universal coverage, then provided there are common identifiers, the potential to match those data with other datasets increases enormously, increasing the analytical power of the data hugely.

The possibility of improving timeliness by utilizing big data is enormously attractive. Policy makers require not only long-term structural information but they also require up-to-date, real time information. Official statistics has generally been very good at providing the former but rather more poor at the latter. This has been a long standing criticism of official statistics. In the words of the Data Revolution Group (2011: 22) 'Data delayed is data denied…The data cycle must match the decision cycle.' Big data offers the possibility of publishing very current indicators, using what Choi and Varian (2011: 1) describe as 'contemporaneous forecasting' or 'nowcasting.' This offers the possibility of identifying turning points much faster, which, from a public policy perspective could be very useful to making better decisions.

As noted above, many digital data are supra-national or global in scope. This globalized aspect of big data offers exciting, although strategically sensitive, opportunities to reconsider the national production models currently employed by NSOs and NSSs all around the world. Switching from a national to a collaborative international production model might make sense from the perspective of improving efficiency and international comparability, but it would be a dramatic change in approach, and possibly a step too far for many NSOs and governments. As MacFeely (2016: 801) notes 'current modernisation initiatives can be summarised as attempts to de-silo legacy production systems. However, in most cases, these attempts to de-silo are done within the constraints of national silos, that is, each country is attempting to de-silo independently.' Nevertheless, in the case of global digital data, the most logical and efficient approach might be to centralise statistical production in a single centre rather than replicating production many times over in individual countries. Obviously, this would not work for all domains, but for some new statistics being derived from globalized big data sets it would offer the chance of real international comparability. A good example of this would be agricultural, land use, maritime and fishery statistics derived from satellite imagery. Such an approach will however pose some difficult questions, not least legal, and the thorny question of who would compile the global statistics?

As noted above, across the world there exists not only a 'digital divide' but also a significant 'data divide.' For many developing countries, the provision of basic statistical information remains a real challenge. The Global Partnership for Sustainable Development Data (2016) note that much of the data that does exist is 'incomplete, inaccessible, or simply inaccurate.' In 2015, at the end of the fifteen year MDG life cycle, developing countries could populate, on average, only two thirds of the Millennium Development Goal (MDG) indicators (United Nations Conference on Trade and Development, 2016). If this is a barometer for data availability in developing countries, then it is clear, that despite significant progress, serious problems with data availability persist. Some (Long & Brindley, 2013; Korte, 2014; Ismail, 2016) have argued, that owing to the falling costs associated with technology, big data may offer developing countries opportunities to skip ahead, and compile next-generation statistics. Example, such as, the massive growth of M-Pesa mobile money services in countries like Kenya, where almost half of the population use it, lend some credence to this argument (Donkin, 2017). Nevertheless others (Mutuku, 2016; United Nations Conference for Trade and Development, 2016; MacFeely & Barnat, 2017; Runde, 2017) have cautioned that in order to do so, there will need to improved access to computers and internet, significant development in numeric and statistical literacy, and in basic data infrastructure. There are also concerns that as statistical legislation and data protection are often weak in many parts of the developing world, focusing on big data before addressing these fundamental issues might do more harm than good in the long term.

Big data may in some cases be better data than survey data. Seth Stephens-Davidowitz makes a compelling argument in his 2017 book 'Everybody Lies' that the content of social media posts, social media likes and dating profiles is no more (or less) accurate than what respondents report in social surveys. However, big data has other types of data available that are of much superior quality. He explains 'the trails we leave as we seek knowledge on the internet are tremendously revealing. In other words, people's search for information is, in itself, information' (2017: 4). He describes data generated from searches, views, clicks and swipes as 'digital truth.' So, Big data may be able to provide more honest data with greater *veracity* than can be achieved from survey data. Hand (2015) makes a similar argument, noting that as big data are transaction data they are closer to social reality than traditional survey and census data that are based on opinions and statements or rely on recall.

Finally, big data may offer NSSs and IOs an opportunity to exercise some leadership and regain some control over an increasingly congested and rapidly fragmenting information space (Cervera et al. 2014; Landefeld, 2014; Kitchin, 2015; MacFeely, 2016). 'Statistical agencies could consider new tasks, such as the accreditation or certification of data sets created by third parties or public or private sectors. By widening its mandate, it would help keep control of quality and limit the risk of private big data producers and users fabricating data sets that fail the test of transparency, proper quality, and sound methodology' (Hammer et al, 2017: 19). Accreditation with standards might help to make a 'multi stakeholder data ecosystem' a reality, offering big data holders an incentive to join that system. Such a move would not be without risks: legal, reputational; and equity. As Landefeld (2014) also points out, such a move might also face its share of resistance, based on ideological grounds challenging the right of government to impose more regulation.

Big data also potentially offer a range of other benefits or opportunities. The *variety* offered by big data provides not only new data sources but the promise of new types of data. These alternative or substitute data sources may offer a mechanism to relieve survey fatigue and burden to households and businesses. Given the exhaustive nature or massive *volume* of big data, they also offer opportunities to improve existing registers (or develop completely new ones) that could improve sample selection and weighting for traditional statistical instruments. The sheer *volume* of data may also allow greater disaggregation of some statistics allowing greater segmentation or granular analyses. It also offers the chance to measure a much wider seelction of new statistics. As noted above, this is a double edged sword, and NSOs must be careful to measure what is important not just what is easy. Nevertheless, big data may be able to contribute to improving the quality of a number of existing statistics (for example, tourism expenditure, travel volumes…) while also offering new approaches to measuring difficult concepts like wellbeing.

Big data offer a wide range of potential opportunities: cost savings; improved timeliness; burden reduction; greater granularity; linkability and scalability; greater accuracy; improved international comparability; greater variety of indicators; and new dynamic indicators. Big data may offer solutions to data deficits in the developing world where traditional approaches have so far failed. Big data may also offer opportunities to rethink what official statistics means and re-position the role of official statistics vis-a-vis the wider data ecosystem. But of course, big data also presents risks and challenges for official statistics. These are examined in the next section.

Challenges for official statistics

Technology, the source of many big data, continues to rapidly evolve. This continuing and rapid evolution raises questions regarding the long-term stability or maturity of big data and their practicality as a data source for the compilation of official statistics. As Daas et al. (2015: 258) note 'The big data sources encountered so far seem subject to frequent modifications'. For example, social media may tweak their services to test alternative layouts, colours and design, which in turn may mutate the underlying data. Kitchin (2015: 9) warns 'the data created by such systems are therefore inconsistent across users and/or time'. The United Nations Economic Commission for Europe (2016) caution that official statisticians using big data will need to accept a general instability in the data. They note 'Wikipedia access statistics show a general drop in the overall number of accesses from the time the mobile version of Wikipedia was released. Similarly, Twitter had a significantly lower number of geo-located tweets after Apple changed the default options for its products.' Consequently they note that time series consistency would be affected by such events. Hence the importance of incorporating 'volatility' in to the definition of big data. Instability of some big data sources introduces risks to continuity of data supply itself. NSOs must decide whether together, access and maturity are sufficiently stable to justify making an investment in big data. Will this 'exhaust pipe' data be consistent over time? If it is not, then this will pose a challenge for official statistics, where a primary focus is to provide consistent time series over time to serve policy analyses. It is often said that data are the new oil. But data (just like crude oil) must be refined in order to produce useable statistics. And just like oil, if the quality and consistency of the raw input data (crude oil) keeps changing, it will be very costly and difficult to refine.

Ownership of source data is another issue of concern. As an NSO moves away from survey based data and becomes more reliant on administrative or other secondary data, such as big data, it surrenders control of its production system. The main input commodity, the source data, is dependent on external factors, exposing the NSO to the risk of exogenous shocks. Partnerships with third party data suppliers means, not only losing control of data generation, but perhaps also sampling and data processing (perhaps as a solution to overcome data protection concerns). Furthermore NSOs will have limited ability to shape the input data they rely upon (Landefeld, 2014; Kitchin, 2015). The technologies that produce 'tailpipe' data may change or become redundant, leading to changes in or disappearances of data. Changes in government social or tax policy may lead to alterations or termination of important administrative datasets. Changes in data protection law, if it does not take the concerns of official statistics into account, could retard the development of statistics for decades[9]. These are risks that a NSO must carefully consider when deciding whether or not to invest in secondary data (administrative or big). Reliance on external data sources also introduces new financial and reputational risks. If a NSO is paying to access a big data set, there is always the risk, that the data provider realizing the value of the data will increase the price. There are also reputational risks. The first is the public, learning that the NSO (and office of the State) is using or 'repurposing’ their social media, telephone, smart metering or credit card data without their consent may react negatively. There may also be concerns or perceptions of state driven 'big brother' surveillance or what Raley (2013) terms 'dataveillance.' So an NSO must consider carefully how it communicates with the public to try and mitigate negative public sentiment. The other reputational risk is that of association. If an NSO is using particular social media data for example, and that provider becomes embroiled in a public scandal, the reputation of the NSO may be adversely affected, through no fault of their own.

As noted earlier, big data are essentially re-purposed data and so, a lot of contextual knowledge of the original generating system is required before the data can be recycled and used for statistical purposes. Developing that knowledge can be difficult as frequently data owners have no incentive to be transparent. Both the data and the algorithms are typically proprietary and often of enormous commercial value. But accessing accurate metadata is vitally important to using any secondary data. For example, understanding how missing data have arisen, perhaps from server downtime or network outages, is essential to assessing the quality of data and then using the data (Daas et al., 2015). Furthermore, as big data can be gamed or contain fake data

[9] For example, within the statistical community of the European Union there are concerns that the new General Data Protection Regulation (GDPR) has not fully taken the particular needs of official statistics into consideration. If this is the case then new legislation may retard significantly the development of official statistics in that region.

(Kitchin, 2015; MacFeely, 2016) it is important to understand the vulnerabilities in the data. There may also be challenges with regard to the representativity and accuracy of many big data. There may be age, gender, language, disability, social class, regional and cultural biases. There are also concerns too that many social media are simply echo-chambers cultivating less than rigorous debate and leading to cyber-cascading, where a belief (either correct or incorrect) rapidly gains currency as a 'fact' as it is passed around the web (Weinberger, 2014). There are also concerns for veracity arising from the concentration of data owners. Reich (2015) notes that in 2010, the top ten websites in the United States accounted for 75 percent of all page views. According to Taplin (2017) Google has an 88 percent market share in online searches, Amazon has a 70 percent market share in ebook sales, Facebook has a 77 percent market share in mobile social media. Such concentration introduces obvious risks of abuse and manipulation, leaving serious questions for the continued veracity of any resultant data. The decision by the Federal Communications Commission (2017) in the United States in December 2017 to repeal Net Neutrality[10] raises a whole new set of concerns regarding the veracity of big data for statistical purposes. United Nations Conference for Trade and Development (2015) noted that ambiguities exist for a range of issues connected with net neutrality, including traffic management practices and their effects on quality of service, competition, innovation, investments, and diversity, online freedom, and protection of human rights. Tim Berners-Lee (2014) has warned against the loss of net neutrality and the increasing concentration within the web: both trends that are undermining the web as a public good.

The emergence of big data is changing the information world. NSOs and NSSs are no longer the single source of truth. The digital revolution and technological ubiquity that has created an abundance of data is challenging the dominant position enjoyed by official statistics for so long, to provide free, timely and high-quality statistics. Today's abundance of data, the basic input commodity or fuel for statistics, has reduced the cost of entry into the statistics compilation business. And so today, 'there is now underway a battle for the ownership of "facts" – a battle that perhaps the global statistical community has not taken sufficiently seriously' (MacFeely, 2017: 64). Today a variety of compilers are producing statistics - and although little is known about the quality of the input data or the compilation process, the allure of these statistics is seductive. The data deluge has contributed to what we call the 'post-truth age' where virtually all authoritative information sources can be challenged by 'alternative facts' or 'fake news' with a consequent diminution of trust and credibility of all sources. As Fukuyama (2017) warns 'In a world without gatekeepers, there is no reason to think that good information will win out over bad.' In fact there are mounting concerns at the weaponisation of data (O'Neill, 2016; Berners-Lee, 2018). Manjoo (2016) too argues there is widespread concern that Facebook and Twitter have hastened a decline in the relevance of facts. Davies (2017) believes official statistics is losing this battle, and argues ‘The declining authority of statistics is at the heart of the crisis that has become known as “post-truth” politics.’ Furthermore, these data are allowing new types of indicators and statistics to be

[10] Net Neutrality sets out the principles for equal treatment of Internet traffic, regardless of the type of service, the sender, or the receiver. In practice, however, the Internet service providers conduct a degree of appropriate traffic management aimed at avoiding congestion, and delivering a reliable quality of service. Concerns regarding the loss of net neutrality focus mainly on definitions of (in)appropriate and (un)reasonable management and discriminatory practices, especially those that are conducted for commercial (e.g. anti-competitive behaviour) or political reasons (e.g. censorship). Net neutrality has three important dimensions: (1) technical (impact on Internet infrastructure); (2) economic (influence on Internet business models); and (3) human rights (possible discrimination in the use of the Internet).

compiled. So not only is the primacy of NSOs and NSSs being challenged, the legitimacy of many traditional statistics, such as GDP or unemployment statistics is also being diminished. National level statistics, based on international agreed classifications, are increasingly viewed by many as overly reductionist and inflexible. Letouzé and Jütting (2014: 19) warn that the 'proliferation of alternative "official" statistics' produced by a variety of outlets are challenging the veracity and trustworthiness of those generated by NSSs. Davies (2017) warns that the privatization of truth will undermine liberalism, democracy and ultimately enlightenment.

The challenges of Privacy and Confidentiality

For official statistics, safeguarding the confidentiality of individual data is sacrosanct and is enshrined in Principle 6 of the United Nations Fundamental Principles of Official Statistics (United Nations, 2014), which states 'Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.' The UN Handbook of Statistical Organization (United Nations, 2003), too, 'underscores repeatedly the requirement that the information that statistical agencies collect should remain confidential and inviolate. The Scheveningen Memorandum (European Commission, 2013)[11] prepared by the Directors General of NSOs in the European Union identified the need to adapt statistical legislation in order to use big data - both to secure access but also protect privacy. The failure to treat individual information as a trust would prevent the statistical agency from functioning effectively. For a NSS to function, confidentiality of the persons and entities for which it holds individual data must be protected i.e. a guarantee to protect the identities and information supplied by all persons, enterprises or other entities. In short, everyone who supplies data for statistical purposes does so with the reasonable presumption that their confidentiality will be respected and protected[12]. In most countries, safeguarding confidentiality is enshrined in national statistical legislation. But with the increased volumes of big data being generated, and the potential to match those data, greater attention must be paid to data suppression techniques to ensure confidentiality can be safeguarded.

The emergence of big data is forcing many challenging questions to be asked, not least with regard to privacy and confidentiality. Mark Zuckerberg, the founder of Facebook, famously claimed that the age of privacy is over (Kirkpatrick 2010). Scott McNealy, CEO of Sun Microsystems, too famously asserted that concerns over privacy are a 'red herring' as we 'have zero privacy' (Noyes, 2015). Many disagree and have voiced concerns over loss of privacy (see Pearson, 2013; Payton and Claypoole, 2015). Fry (2017) has likened developments with regard to

[11] Para 3 - Recognise that the implications of Big Data for legislation especially with regard to data protection and personal rights (e.g. access to Big Data sources held by third parties) should be properly addressed as a matter of priority in a coordinated manner.

[12] In effect this means that only aggregate data can be published for general release by official statistical compilers and those aggregates will have been tested for primary and secondary disclosure. Data that cannot be published due to the risk of statistical disclosure are referred to as confidential data. Primary confidentiality disclosure arises when dissemination of data provides direct identification of an individual person or entity. This usually arises when there are insufficient records in a cell to mask individuals or when one or two records are dominant and so their identity remains evident despite many records (this is a recurring challenge for business statistics where 'hiding' the identity of large multinational enterprises can be very difficult). Secondary disclosure may arise when data that have been protected for primary disclosure nevertheless reveal individual information when cross-tabulated with other data.

big data and the loss of privacy to the opening of Pandoras Box - Pandora 5.0. The introduction in Europe of the new General Data Protection Regulation which comes into effect in 2018, reinforcing citizen’s data-protection rights, including among other things the right ‘to be forgotten’, suggests that privacy is still a real concern (European Parliament 2016) - at least in some regions of the world. By contrast, in the United States, users who provide information under the 'third-party doctrine' i.e. to utilities, banks, social networks etc. should have 'no reasonable expectation of privacy.'

This introduces two new challenges for official statistics: one technical and one of perception. The technical challenge arises from the availability of large, linkable datasets which present a problem thought to have been solved in traditional statistics – anonymisation. But big data, combined with the enormous computing power available today, it is clear that simply removing personal identifiers and aggregating individual data is not a sufficient safeguard. A paper by Ohm in 2010 outlining the consequences of failing to adequately anonymise data graphically illustrates why there is no room for complacency. Thus a problem that had been solved in the context of traditional official statistics must now be solved again, in the context of a richer and more varied data ecosystem. The changing nature of perception is arguably a trickier problem. What if Zuckerberg and McNealy are correct and future generations are less concerned about privacy? There appears to be some evidence to suggest that they may be correct. It seems there are clear inter-generational differences in opinion vis-a-vis privacy and confidentiality, where those 'born digital' (roughly those born since 1990) are less concerned about disclosing personal information than older generations (European Commission, 2011). Taplin (2017: 157) ponders this, musing 'It very well may be that privacy is a hopelessly outdated notion and that Mark Zuckerberg's belief that privacy is no longer a social norm has won the day.' If this is so, what are the implications for official statistics and anonymisation? If other statistical providers, not governed by the UN fundamental principles, take a looser approach to confidentiality and privacy, it may leave official statistics in a relatively anachronistic and disadvantaged position vis-a-vis other data providers. But moving away from or discarding principle 6 of the UN Fundamental Principles for Official Statistics would seem to be a very risky move, given the importance of public trust for NSOs.

A related and emerging challenge for official statistics is that of open data, or more specifically, the asymmetry in openness expected of private and public sector data. Many of the 'open data' initiatives are in fact drives to open government data[13]. This of course makes sense, in that tax payers should to some extent own the data they have paid for, and so those data should be public, within sensible limits. But arguably people also own much of the data being held by search

[13] For example: the OECD Open Government Data (OGD) is a philosophy- and increasingly a set of policies - that promotes transparency, accountability and value creation by making government data available to all - see: http://www.oecd.org/gov/digital-government/open-government-data.htm. In the United States, Data.gov aims to make government more open and accountable. Opening government data increases citizen participation in government, creates opportunities for economic development, and informs decision making in both the private and public sectors - see: https://www.data.gov/open-gov/. In the European Union, there is a legal framework promoting the re-use of public sector information - Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information. See - http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32013L0037&from=FR

engines, payments systems and telecommunication providers too. So why is there an exclusive focus on public or government data? Letouzé and Jütting (2014: 10) have highlighted this issue, remarking that 'Official statisticians express an acute and understandable sense of frustration over pressure to open up their data to private-sector actors, while these same actors are increasingly locking away what they and many consider to be "their" data.' Official Statistics, as a public good, should of course be open. But the philosophy of open data should be more evenly applied to avoid asymmetrical conditions. This is a complex challenge, as to some extent it feeds off poor understanding of privacy issues, statistical literacy and the data wars that are underway at the moment. Rudder (2014: 241) notes that 'because so much happens with so little public notice, the lay understanding of data is inevitably many steps behind the reality.'

Taplin (2017: 157) argues that we trade our privacy with corporations in return for innovation or benefits, 'but it is one thing to forfeit our privacy as individuals to a company that we believe is delivering a needed service and another to open our personal lives to the federal government.' MacFeely (2016) has warned that if the privacy-benefit trade-off of official statistics is insufficiently clear to the public or policy makers, then it leaves official statistics vulnerable, and possibly facing a precarious and bleak future. Rudder (2014: 242) highlights this challenge too noting that 'the fundamental question in any discussion of privacy is the trade-off - what you get for losing it.' Like Taplin, Rudder also argues that the trade-off benefit with the private sector is clear - better targeted ads! He argues that 'what we get in return for the government's intrusion is less straightforward.' McNealy too, who seems unconcerned about the lack of privacy in the private sector, takes a very different attitude when it comes to government, saying 'It scares me to death when the NSA or the IRS know things about my personal life and how I vote…Every American ought to be very afraid of big government' (Noyes, 2015). A challenge for official statistics, an arm of the state, is how to put clear blue water between the NSS and the other institutions of government from the perspective of data sharing, but highlight the common benefits of official statistics as a public good. To some extent there is ideology at play here, where a neo-liberal agenda is pushing to minimise the role of the public sector, but it also illustrates the challenge facing governments and their agencies generally where their contribution to the wellbeing of economies and societies is poorly understood.

Thus, while big data may offer opportunities, they also present some real challenges for NSOs and NSSs. To some extent, these challenges are magnified versions of problems that exist with other data sources, such as, uncertainty over the quality or veracity of data and dealing with a range of potential biases. Access to external secondary sources, such as, administrative data can already be challenging, and is not unique to big data. But big data do appear to present some rather unique challenges with regard to rapidly evolving and unstable data, ownership of data, data protection and safeguarding confidentiality. These are some of the issues that NSOs and IOs will need to carefully consider before committing resources to any big data projects.

Governance Issues

As outlined above, big data presents a range of challenges for NSOs, NSSs and IOs. In considering whether big data provides a viable option, statistical offices and systems must carefully decide what governance systems will be required to ensure the official statistics brand is not compromised. For the purposes of this chapter, governance systems can be defined as the policies, rules and monitoring mechanisms that allow the management of a NSO or IO to direct and control the activities of the office. That governance system should help decision makers to balance the often competing needs of new statistical demands with the rights of data owners and ensure public accountability.

At a global level, questions naturally arise as to whether some sort of global governance framework for the treatment of big data will be required or whether ad-hoc or bespoke national or regional agreements can work. In a world where big data are being used more extensively, the multinational enterprises generating many of these massive global datasets will effectively be setting many future data standards. What will this mean for the global statistical system? What will it mean for the United Nations Fundamental Principles of Official Statistics? These massive new globalized data also challenge the justification for national or local compilation, raising a host of legal, security and organizational questions.

At individual NSO, NSS or IO level, there are also governance issues to be considered. The issues identified here are not exhaustive, but give a flavor of the issues that a NSO or IO may need to be consider:

Ethics - many big data are the exhaust from technology. Deriving statistics involves repurposing those data. The possibilities are exciting and may offer incredible opportunities. In the rush to compile new statistics, it may be easy to forget where those data came from. Thus it may be sensible to establish an ethics committee to consider whether the compilation of new statistics justify the potential 'intrusion' to citizens privacy. A board, not immediately involved in the compilation process, may be better able to weigh-up the pros and cons of a big data project and take a more balanced view on whether 'no harm' will be done. A NSO may wish to consider also, that in using a particular big data set, it may be inadvertently taking an ideological or philosophical stance on a range of emerging debates, including for example, the ownership of data.

Legal - there will be many legal issues to be unpicked in the years to come with regard to big data. For example, can a NSO or IO access data sources, such as, credit card expenditure information or mobile phone location data without breaching data protection, statistical or other legislation? It will probably be necessary (or, at the very least sensible) to establish a board of specialist legal experts who can adjudicate on these complex issues and provide comprehensive legal opinion to the management board of the NSO or IO. The correspondence between statistical and data protection legislation will be of paramount importance in the coming years.

Oversight and Confidentiality - there will most likely be a growing need for a committee that deals specifically with the confidentiality and oversight of access to data held by a NSO or IO. Storing big data will present governance challenges. Who has access to those data and why? Who decides who should have access and using what criterion? How is confidentiality of published data being safeguarded? This is a mixture of statistical methodology and broader governance. This board might also play a useful role, in coordination with ethics committee in deciding whether certain data sets should be linked, and if they are, what are the likely implications for protecting confidentiality?

IT and Security - storing large volumes of data, and providing sufficient processing power and memory, will present technical challenges too. Obviously sufficient space will be required. But new cyber-security protocols will also be required. 'Any data collected will invariably leak' - so warns Goodman (2015: 153). What does globalized data mean for storage location – does it make sense that NSOs, NSSs or IOs continue with the old paradigm where data are stored, locally, in-house? If it is stored locally, will the data be quarantined and stored offline (so that it cannot be hacked or corrupted). If not, will the NSO require some types of randomized identifiers to suppress identities? But does storing global data and re-processing the same data many times over in different locations make sense? Would it be more efficient to store the data at source, or in some central location (in the cloud?). But how then will the data be integrated with other data sources stored in different locations. The movement and transfer of data will require secure pipeline systems and sophisticated encryption.

Quality Assurance - Assessing the quality of big data is not the same as assessing traditional datasets. Firstly, quality must be defined from the perspective of big data and clear criterion for how these can be measured must be developed. United Nations Economic Commission for Europe (2016) note that using big data may mean accepting 'different notions of quality.' Owing to new quality issues, for example, disorganized data management, more time and effort may be required to organize and properly manage data. Gao et al. (2016) identify a number of quality parameters unique to cleaning and organizing big data. They are: determining quality assurance; dealing with data management and data organization; and the particular challenges of data scalability; and transformation and conversion. Using big data may require an extended quality framework for official statistics. Such a framework might put greater emphasis on risk management than that currently used.

Continuous Professional Development & Training - using big data will require a blend of different skills to that of the traditional statistician, with more emphasis on data mining and analytics. Given the demand for mathematically skilled graduates today, it will be necessary to retrain some existing statisticians. This should be an on-going process in any event for professional statisticians. Nevertheless, big data may be the catalyst for some NSOs or IOs to consider establishing formal training or a Continuous Professional Development (CPD) programme. It may also provide an impetus to consider new partnerships and collaborations in order to bring in new skills.

Strategic Partnerships - as noted above, using big data presents a range of technical challenges that may require new strategic partnerships. The decision for NSOs and IOs is whether it makes sense to try and develop all of the skills in-house or whether it will be better to partner with other entities that have the required skills. These will be critical decisions, both in terms of costs and efficiencies, but also for legal and reputational reasons. In making these decisions, NSOs and IOs must ensure they do not compromise the Fundamental Principles of Official Statistics.

Communications and dissemination - any NSO planning to use big data in the day-to-day compilation process should prepare carefully a communications strategy. How will repurposing be explained and communicated to the public? Will the NSO publish an inventory of administrative and big data being accessed, stored and used by the NSO? What is the plan, when and if, some scandal arises that embroils the NSO in a negative media story? NSOs and IOs must also carefully consider how to make new statistics available - in particular how to use technology to make the experience more interactive and user friendly for users.

Conclusion

It is not clear, as yet, whether big data offers anything special. But is seems likely that Tamm is correct and big data are just more data: another phase in the evolution of data rather than a revolution. That said there are some unique aspects to big data. Perhaps the most unusual is the source – many new big data are created or taken from people who are not necessarily aware that their data are being re-used. This raises some important ethical and perhaps philosophical questions regarding the ownership of the data. It is likely that in the future, the argument that by signing a social media ‘terms of service’ (that takes a week to read) means a citizen has signed over their data ownership rights will be tested in court. What that will mean for official statistics is unclear at this juncture.

Official statistics in the future must operate in a very different environment to the one it has survived in the past. Ubiquitous technology has created a deluge of digital data that are now being used to compile a variety of new informational indicators and statistics. NSOs, NSSs and IOs must grapple with this competition, and must also decide whether and how big data can be harnessed to compile official statistics. What makes big data so intriguing is the fact that they simultaneously present both threats and opportunities for official statistics. From a governance perspective, the challenge for official statistics is to identify and mitigate the threats while seizing the opportunities.

The purpose of official statistics is to provide high quality, independent, impartial and timely information that allow citizens to challenge stereotypes, governments, public bodies and private enterprises and hold them to account. Despite the abundance of information available today, the need for official statistics has arguably never been greater. In a world awash with 'fake news' official statistics must shine like a beacon of truth. Official statisticians should be gatekeepers, providing what Weinberger (2014) terms 'stopping points' i.e. information that can be taken as fact and used as the final resort in the case of disagreement.

In relative terms, big data are still new. At the turn of the century, Scott Cook, the CEO of Intuit mused 'we're still in the first minutes of the first day of the Internet revolution' (Levington, 2000). Even today we are probably only in the first hour. Many norms and standards are yet to evolve. But it does not take a huge leap of imagination to foresee that in the not too distant future, the misuse of big data will be at the heart of a serious human rights abuse scandal. Official statistics must take the ethical dimension seriously. Just because something can be measured doesn't mean it should be. In assessing whether to, and how to use big data, NSOs must begin to carefully consider the human rights of citizens in this digital age.

Big data, if they can be harnessed properly, would appear to offer some tantalizing opportunities - not least improved timeliness and the chance to better align the availability of statistics with policy needs. Perhaps in some cases they can improve accuracy. The possibilities of matching different digital data sets may also allow us to dramatically improve our understanding of complex, cross-cutting issues, such as, gender inequality or the challenges of being disabled.

Developments, such as, the Internet of Things[14], biometrics and behaviometrics will all surely present opportunities to develop new and useful statistics. As yet, the implications of this 'big data bang' for statistics is not immediately clear, but one can envisage a whole host of new ways to measure and understand the human condition.

These developments will bring a myriad of new challenges too, not least the growth of unreliable information. It is already clear that big data will not be 'a panacea for statistical agencies confronting demands for more, better, and faster data with fewer resources' (Landefeld, 2014: p19). This may not be universally understood and so managing expectations will be an ongoing challenge for official statisticians. Challenges regarding how best to determine the quality and veracity of big data from a statistical perspective remain. The growing centralization or monopolization of the internet, the threat to net neutrality, and the growing volumes of 'bot' traffic are just some of the issues that may compromise the quality and impartiality of any resultant statistics. There are concerns too, that many social media channels are polarising social exchange and promoting 'echo chambers' and cyber-cascading. As David Eggers, in his wonderful book *The Circle* remarked, social media has 'elevated gossip, hearsay and conjecture to the level of valid, mainstream communication' (Eggers, 2013: 132). Official statisticians must ensure they can filter the wheat from the chaff.

There is a new gold rush underway - a data rush. In that rush, NSOs and IOs are feeling the pressure to be seen to utilize big data. But as outlined above, it will be a bumpy road with many challenges along the way. It is of course often easier to see problems than opportunities, so NSO's and IOs must carefully weigh-up the likely costs and benefits of using big data, both now and in the future. In making that decision, they must not lose sight of their mission and their mandates.

References:

Adams, D. (1979). 'The Hitchhikers Guide to the Galaxy.' Pan, London.

Aslam, S. (2015). 'Snapchat by the Numbers: Stats, Demographics and Fun Facts.' *Omnicore*, October 7, 2015. Retrieved from: http://www.omnicoreagency.com/snapchatstatistics/ [26 September, 2016].

Berners-Lee, T. (2014). 'Tim Berners-Lee on the Web at 25: the past, present and future'. *Wired*, 23 August, 2014. Retrieved from: http://www.wired.co.uk/article/tim-berners-lee [19 March, 2018].

Berners-Lee, T. (2018). 'The web is under threat. Join us and fight for it.' *World Wide Web Foundation*. March 12, 2018. Retrieved from: https://webfoundation.org/2018/03/web-birthday-29/ [19 March, 2018].

[14] In 2006 there were some 2 billion 'smart devices' connected to each other. By 2020 it is projected that this 'internet of things' will compromise of somewhere between 30 and 50 billion devices (Nordrum, 2016). Goodman (2015) notes the result will be 2.5 sextillion potential networked object-to-object interactions.

Blackwell, J. (1985). 'Information for policy.' *National and Economic Social Council*, Report no. 78. Dublin: NESC. Retrieved from: http://files.nesc.ie/nesc_reports/en/NESC_78_1985.pdf [18 January, 2018].

Borgman, C.L. (2015). 'Big Data, Little Data, No Data - Scholarship in the Networked World.' Cambridge, MA: MIT Press.

Boyd, J and K. Crawford (2012). 'Critical Questions for Big Data - Provocations for a cultural, technological, and scholarly phenomenon.' *Information, Communication & Society*, Vol 15, No.5, pp. 662 - 679, DOI:10.1080/1369118X.2012.678878.

Brackstone, G. J. (1987). 'Statistical Issues of Administrative Data: Issues and Challenges'. Survey Methodology, Vol. 13, No. 1, pp. 29 – 43.

Buytendijk, F. (2014). 'Hype Cycle for Big Data, 2014.' *Gartner*. Retrieved from: https:// www.gartner.com/doc/2814517/hype-cycle-big-data- [11 August, 2015].

Coordination Committee for Statistical Activities (2014). 'Principles Governing International Statistical Activities.' Retrieved from: https://unstats.un.org/unsd/accsub-public/principles_stat_activities.htm [21 February, 2018].

Cervera J.L., P. Votta, D. Fazio, M. Scannapieco, R. Brennenraedts and T, van der Vorst (2014). 'Big Data in Official Statistics.' *Eurostat ESS Big Data Event*, Rome2014 – Technical Event Report. Retrieved from:
https://ec.europa.eu/eurostat/cros/system/files/Big%20Data%20Event%202014%20-%20Technical%20Final%20Report%20-finalV01_0.pdf [18 January, 2018].

Choi, H. and H. Varian (2011). 'Predicting the present with Google Trends.' Retrieved from: http://people.ischool.berkeley.edu/~hal/Papers/2011/ptp.pdf [17 January, 2018].

Dass, P.J.H., M.J. Puts, B. Buelens and P.A.M. van den Hurk (2015). 'Big Data as a Source for Official Statistics.' *Journal of Official Statistics*, Vol. 31, No. 2, pp. 249 - 262.

Data Revolution Group (2011). 'A World that Counts: Mobilizing the Data Revolution for Sustainable Development.' Report prepared at the request of the United Nations Secretary-General, by the Independent Expert Advisory Group on a Data Revolution for Sustainable Development. November 2014. Retrieved from:
http://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf [17 January, 2018].

Davies, W. (2017). 'How statistics lost their power – and why we should fear what comes next.' Retrieved from: https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy?CMP=share_btn_link [19 January 2018].

Donkin, C. (2017). 'M-Pesa continues to dominate Kenyan market.' *Mobile World Live*, January 25, 2017. Retrieved from: https://www.mobileworldlive.com/money/analysis-money/m-pesa-continues-to-dominate-kenyan-market/ [20 February, 2018]

Eggers, D. (2013). 'The Circle.' Penguin Books, London

European Commission (2011). 'Attitudes on Data Protection and Electronic Identity in the European Union.' *Special Eurobarometer* No. 359, Wave 74.3 - TNS Opinion and Social. Published June 2011. Retrieved from: http://ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs_359_en.pdf [16 January, 2018].

European Commission (2013). 'Scheveningen Memorandum on "Big Data and Official Statistics".' Adopted by the European Statistical System Committee on 27 September 2013. Retrieved from: https://ec.europa.eu/eurostat/cros/content/scheveningen-memorandum_en [25 January, 2018].

European Commission (2014). 'Big Data.' *Digital Single Market Policies*. Retrieved from: https://ec.europa.eu/digital-single-market/en/policies/big-data [31 January, 2018].

European Parliament (2016). 'Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).' Retrieved from: http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf [16 Jan, 2018].

Federal Communications Commission (2017). 'Restoring Internet Freedom.' Retrieved from: https://www.fcc.gov/restoring-internet-freedom [24 January, 2018].

Fry, S. (2017). 'The Way Ahead'. Lecture delivered on the 28th May 2017, Hay Festival, Hay-on-Wye. Retrieved from: http://www.stephenfry.com/2017/05/the-way-ahead/ [19 March, 2018].

Fukuyama, F. (2017). 'The Emergence of a Post Fact World.' *Project Syndicate*, August 21, 2017. Retrieved from: https://www.project-syndicate.org/onpoint/the-emergence-of-a-post-fact-world-by-francis-fukuyama-2017-01 [4 January, 2018].

Gelsinger, P. in Whatsthebig data? (2012). 'Big Data quotes of the week'. Retrieved from: https://whatsthebigdata.com/2012/06/29/big-data-quotes-of-the-week-11/ [19 March, 2018].

Gibson, W. in The Economist (2001). 'Broadband Blues - Why has broadband Internet access taken off in some countries but not in others?' The Economist, June 21, 2001. Retrieved from: https://www.economist.com/node/666610 [19 March, 2018].

Global Partnership for Sustainable Development Data (2016). 'The Data Ecosystem and the Global Partnership.' Retrieved from: http://gpsdd.squarespace.com/who-we-are/ [19 January, 2018].

Goa,J., C. Xie and C. Tao (2016). 'Big Data Validation and Quality Assurance - Issuses, Challenges, and Needs.' 2016 *IEEE Symposium on Service-Oriented System Engineering* (SOSE), Issue Date: March 29 2016-April 2 2016. Retrieved from: http://ieeexplore.ieee.org/xpls/icp.jsp?arnumber=7473058 [20 February, 2018].

Goodbody, W (2018). 'Waterford researchers develop new method to store data in DNA.' *RTE News*, January 20, 2018. Retrieved from: https://www.rte.ie/news/ireland/2018/0219/941956-dna-data/ [20 January, 2018].

Goodman, M. (2015). 'Future Crimes - Inside the Digital Underground and the Battle for Our Connected World.' Anchor Books, New York.

Hammer, C.L., D.C. Kostroch, G. Quiros and STA Internal Group (2017). 'Big Data: Potential, Challenges, and Statistical Implications.' *IMF Staff Discussion Note*, SDN/17/06, September 2017. Retrieved from: http://www.imf.org/en/Publications/SPROLLs/Staff-Discussion-Notes [11 January, 2018].

Hand, D. J. (2015). 'Official Statistics in the New Data Ecosystem.' presented at the *New Techniques and Technologies in Statistics conference*, Brussels, March 10-12, 2015. Retrieved from: https://ec.europa.eu/eurostat/cros/system/files/Presentation%20S20AP2%20%20Hand%20-%20Slides%20NTTS%202015.pdf [23 January, 2018].

Harkness, T. (2017). 'Big Data: Does size matter?' Bloomsbury Sigma, London, UK.

Hilbert, M. and P. Lopez (2012). 'How to Measure the World's Technological Capacity to Store, Communicate and Compute Information.' *International Journal of Communication*, Vol 6, pp. 956–979.

IBM (2017). '10 Key Marketing Trends for 2017 and Ideas for Exceeding Customer Expectations.' *IBM Marketing Cloud*. Retrieved from: https://public.dhe.ibm.com/common/ssi/ecm/wr/en/wrl12345usen/watson-customer-engagement-watson-marketing-wr-other-papers-and-reports-wrl12345usen-20170719.pdf [25 January, 2018].

Ismail, N. (2016). 'Big Data in the developing world.' Information Age, 8 September, 2016. Retrieved from: http://www.information-age.com/big-data-developing-world-123461996/ [19 January, 2018]

International Telecommunications Union (2017). 'ITU Key 2005 - 2017 ICT data.' Retrieved from: https://idp.nz/Global-Rankings/ITU-Key-ICT-Indicators/6mef-ytg6 [12 Jan, 2018].

Kirkpatrick, M. (2010). 'Facebook's Zuckerberg Says the Age of Privacy is Over.' *Readwrite.Com*, 9 January 2010. Retrieved from: https://readwrite.com/2010/01/09/facebooks_zuckerberg_says_the_age_of_privacy_is_ov/ [21 February, 2018].

Kitchin, R. (2015). 'The opportunities, challenges and risks of big data for official statistics.' *Statistical Journal of the International Association of Official Statistics*, Vol. 31, No. 3, pp. 471-481.

Korte, T. (2014). 'How Data and Analytics Can Help the Developing World.' *Huffington Post - The Blog*. 21 September, 2014. Retrieved from: https://www.huffingtonpost.com/travis-korte/how-data-and-analytics-ca_b_5609411.html [19 January, 2018].
Krikorian, R. (2013). 'New Tweets per Second Record, and How!' *Engineering Blog*. Retrieved from: https://blog.twitter.com/2013/new-tweets-per-second-record-and-how [26 September, 2016].

Kulp, P. (2017). 'Facebook quietly admits to as many as 270 million fake or clone accounts.' *Mashable*, November 3, 2017. Retrieved from: https://mashable.com/2017/11/02/facebook-phony-accounts-admission/#UyvC2aOAmPqo [20 February, 2018].

Landefeld, S. (2014). 'Uses of Big Data for Official Statistics: Privacy, Incentives, Statistical Challenges, and Other Issues.' Discussion paper presented at the United Nations Global Working Group on Big Data for Official Statistics, Beijing, China 31 October, 2014. Retrieved from: https://unstats.un.org/unsd/trade/events/2014/beijing/Steve%20Landefeld%20-%20Uses%20of%20Big%20Data%20for%20official%20statistics.pdf [January 18, 2018].

Laney, D. (2001). '3D Data Management: Controlling data volume, velocity and variety.' Meta Group, File 949, 6 February, 2001. Retrieved from: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf [11 January, 2018].

Letouzé, E. and J. Jütting (2015). 'Official Statistics, Big Data and Human Development.' Data Pop Alliance, *White Paper Series*, Retrieved from: https://www.paris21.org/sites/default/files/WPS_OfficialStatistics_June2015.pdf [16 Jan, 2018].

Levington, S. (2000). 'Internet Entrepreneurs Are Upbeat Despite Market's Rough Ride.' *The New York Times*, May 24, 2000. Retrieved from: http://www.nytimes.com/2000/05/24/business/worldbusiness/internet-entrepreneurs-are-upbeat-despite-markets.html [20 February, 2018].

Long, J. and W. Brindley (2013). 'The role of big data and analytics in the developing world: Insights into the role of technology in addressing development challenges.' Accenture Development Partnerships. Retrieved from: https://www.accenture.com/us-en/~/media/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Strategy_5/Accenture-ADP-Role-Big-Data-And-Analytics-Developing-World.pdf [19 January, 2018]

MacFeely, S. and J. Dunne (2014). 'Joining up public service information: The rationale for a national data infrastructure.' *Administration*, Vol.61, No.4, pp. 93–107, MacFeely, S. (2016). 'The Continuing Evolution of Official Statistics: Some Challenges and Opportunities.' *Journal of Official Statistics*, Vol. 32, No. 4, 2016, pp. 789–810.

MacFeely, S. (2017). 'Measuring the Sustainable Development Goals: What does it mean for Ireland?' *Administration*, Vol.65, No.4, pp. 41 - 71.

MacFeely, S. and N. Barnat (2017). 'Statistical capacity building for sustainable development: Developing the fundamental pillars necessary for modern national statistical systems.' *Statistical Journal of the International Association of Official Statistics*, Vol.33, No. 4, pp. 895 - 909.

Manjoo, F. (2016). 'Silicon Valley Reels After Trump's Election.' *The New York Times* - State of the Art, November 9, 2016. Retrieved from: https://www.nytimes.com/2016/11/10/technology/trump-election-silicon-valley-reels.html [26 January, 2018].

Mayer-Schonberger, V. and K. Cukier (2013). 'Big Data: A Revolution That Will Transform How We Live, Work and Think.' London: John Murray.

Meeker, M. (2017). 'Internet Trends 2017.' Presented at the *Code Conference*, Rancho Palos Verdes, California, May 31, 2017. Retrieved from: http://www.kpcb.com/internet-trends [12 Jan, 2018].

Mutuku, L. (2016) in Serra, C. (2016). 'The big data challenge for developing countries.' *The world academy of sciences*, 2 September, 2016. Retrieved from: https://twas.org/article/big-data-challenge-developing-countries [19 January, 2018].

Nilson (2018). 'Global Cards - 2015: Special Report.' *The Nilson report*, Jan 12, 2018. Retrieved from: https://www.nilsonreport.com/publication_special_feature_article.php [12 January, 2018].

Nordrum, A. (2016). 'Popular Internet of Things Forecast of 50 Billion Devices by 2020 Is Outdated.' *IEEE Spectrum*, August 18, 2016. Retrieved from: https://spectrum.ieee.org/tech-talk/telecom/internet/popular-internet-of-things-forecast-of-50-billion-devices-by-2020-is-outdated [21 February, 2018].

Noyes, K. (2015). 'Scott McNealy on privacy: You still don't have any.' *PC World*, IDG News Service, June 25, 2015. Retrieved from: https://www.pcworld.com/article/2941052/scott-mcnealy-on-privacy-you-still-dont-have-any.html [29 January, 2018].

Ohm, P. (2010). 'Broken promises of privacy: Responding to the surprising failure of anonymization.' UCLA Law Review, 2010, Vol. 57, pp.1701-1777. Retrieved from: http://www.uclalawreview.org/pdf/57-6-3.pdf [19 March, 2018].

O'Neill, C. (2016). 'Weapons of Math Destruction - How big data increases inequality and threatens democracy.' Allen Lane, London.

Payton, T. and T. Claypoole (2015). 'Privacy in the Age of Big Data - Recognising the Threats Defending Your Rights and Protecting Your Family.' Lanham, MD: Rowman & Littlefield.

Pearson, E. (2013). 'Growing Up Digital.' Presentation to the *OSS Statistics System Seminar Big Data and Statistics New Zealand*: A seminar for Statistics NZ staff, Wellington, 24 May 2013. Retrieved from: https://www.youtube.com/watch?v=lRgEMSqcKXA [19 December, 2017].

Raley, R. (2013). 'Dataveillance and countervailance' in Gitelman, l. (Ed) '"Raw Data" is an Oxymoron.' MIT Press, Cambridge.

Reich, R. (2015). 'Saving Capitalism: For the Many, Not the Few.' London: Icon Books Ltd.

Rudder, C. (2014). 'Dataclysm: What our online lives tell us about our offline selves.' 4th Estate, London.

Runde, D. (2017). 'The Data Revolution in Developing Countries Has a Long Way to Go.' *Forbes*, February 25, 2017. Retrieved from: https://www.forbes.com/sites/danielrunde/2017/02/25/the-data-revolution-in-developing-countries-has-a-long-way-to-go/2/#3a48f53e482f [19 January, 2018].

Stephens-Davidowitz, S. (2017). 'Everybody lies - What the internet can tell us about who we really are.' Bloomsbury, London, UK.

Struijs, P., B. Braaksma and P.J.H. Daas (2014). 'Official statistics and Big Data.' *Big Data & Society*, April–June 2014: 1–6, DOI: 10.1177/2053951714538417. Retrieved from: http://journals.sagepub.com/doi/pdf/10.1177/2053951714538417 [21 January, 2018].

Tam, S. and F. Clarke (2015). 'Big Data, Official Statistics and Some Initiatives by the Australian Bureau of Statistics.' *International Statistical Review*, doi: 10.1111/insr.12105. Retrieved from: https://www.researchgate.net/publication/280972848_Big_Data_Official_Statistics_and_Some_Initiatives_by_the_Australian_Bureau_of_Statistics [11 January, 2018].

Taplin, J. (2017), 'Move Fast and Break things - How Facebook, Google and Amazon cornered culture and undermined democracy.' Little, Brown and Company, New York.

Thamm, A. (2017). 'Big Data is dead.' LinkedIn, November 23, 2017. Retrieved from: https://www.linkedin.com/pulse/big-data-dead-just-regardless-quantity-structure-speed-thamm/ [16 January, 2018].

United Nations (2003). 'Handbook of Statistical Organization – 3rd Edition: The Operation and Organization of a Statistical Agency.' Department of Economic and Social Affairs Statistics Division Studies in Methods Series F No. 88. United Nations, New York, 2003. Retrieved from: https://www.paris21.org/sites/default/files/654.pdf [25 January, 2018].

United Nations (2014). 'Resolution adopted by the General Assembly on 29 January 2014 - Fundamental Principles of Official Statistics.' *General Assembly*, A/RES/68/261. Retrieved from: http://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf [26 September, 2016].

United Nations (2015). 'United Nations Fundamental Principles of Official Statistics Implementation Guidelines, 2015.' Retrieved from: https://unstats.un.org/unsd/dnss/gp/Implementation_Guidelines_FINAL_without_edit.pdf [15 January, 2018].

United Nations Conference for Trade and Development (2015). 'Mapping of international Internet public policy issues.' E/CN.16/2015/CRP.2, *Commission on Science and Technology for Development*, Eighteenth session, Geneva, 4-8 May 2015. Retrieved from: http://unctad.org/meetings/en/SessionalDocuments/ecn162015crp2_en.pdf [24 January, 2018].

United Nations Conference for Trade and Development (2016). 'Development and Globalization: Facts and Figures 2016.' http://stats.unctad.org/Dgff2016/ [19 January, 2018].

United Nations Economic Commission for Europe (2011). 'Using Administrative and Secondary Sources for Official Statistics - A Handbook of Principles and Practices'. Retrieved from: https://unstats.un.org/unsd/EconStatKB/KnowledgebaseArticle10349.aspx [26 February, 2018].

United Nations Economic Commission for Europe (2000). 'Terminology on Statistical Metadata.' Conference of European Statisticians Statistical Standards and Studies, No.53. Retrieved from: http://ec.europa.eu/eurostat/ramon/coded_files/UNECE_TERMINOLOGY_STAT_METADATA_2000_EN.pdf [26 February, 2018].

United Nations Economic Commission for Europe (2016). 'Outcomes of the UNECE Project on Using Big Data for Official Statistics.' Retrieved from: https://statswiki.unece.org/display/bigdata/Big+Data+in+Official+Statistics [15 February, 2018].

United Nations Statistical Commission (2014). 'Big data and modernization of statistical systems; Report of the Secretary-General.' E/CN.3.2014/11 of the forty-fifth session of UNSC 4-7 March 2014. Retrieved from: https://unstats.un.org/unsd/statcom/doc14/2014-11-BigData-E.pdf [11 January, 2018].

Waterford Technologies (2017). 'Big Data Statistics & Facts for 2017.' Posted on February 22, 2017: Retrieved from: https://www.waterfordtechnologies.com/big-data-interesting-facts/ [3 Jan, 2018].

Weigand, A. (2009). 'The Social Data Revolution(s).' *Harvard Business Review*, May 20, 2009. Retrieved from: https://hbr.org/2009/05/the-social-data-revolution.html [24 April, 2017].

Weinberger D. (2014). 'Too Big to Know.' Basic Books, New York