Chapter 1
Big Data in Academic Research: Challenges, Pitfalls, and Opportunities

Jacques Raubenheimer

Abstract Big Data are a product of the computer era, enabling the knowledge economy, in which academic researchers are key players, although researchers have been slow to adopt Big Data as a source for academic enquiry. This may be in part because Big Data are curated by commercial or governmental entities, not by researchers. Big Data present several challenges to researchers, including those associated with the size of the data, the development and growth of data sources, and the temporal changes in large data sets. Further challenges are that Big Data are gathered for purposes other than research, making their fit-for-purpose problematic; that Big Data may easily lead to overfitting and spuriousness; and the biases inherent to Big Data. Linkage of data sets always remains problematic. Big Data results are hard to generalize, and working with Big Data may raise new ethical problems, even while obviating old ethical concerns. Nonetheless, Big Data offer many opportunities, allowing researchers to study previously inaccessible problems, with previously inconceivable sources of data. Although Big Data overcome some of the challenges of small data studies, Big Data studies will not supplant small data studies—these should work in concert, leading to real-world translation that can have a lasting impact.

Keywords Big data · Volume, velocity, variety · Dimensionality · Overfitting · Ethics · Google Trends · Facebook · Crimson Hexagon · Social media

J. Raubenheimer (✉)
University of Sydney, Sydney, Australia
e-mail: jacques.raubenheimer@sydney.edu.au

T. Prodromou (ed.), *Big Data in Education: Pedagogy and Research*, Policy Implications of Research in Education 13,
https://doi.org/10.1007/978-3-030-76841-6_1

Introduction

The concept of "Big Data" may very well be relative. Certainly, if the term had existed at the time, the library of Alexandria would have been described as a Big Data warehouse—it was purported to contain "all the books in the world" (Charles, 1913). In fact, its closest modern equivalent, the Library of Congress, has holdings for which the complete catalogue itself counts as a big data set even by today's standards, being in excess of 164 million items (Library of Congress, 2017).

Regardless of how we define *big data*, one thing is certain of the twenty-first century—we are no longer in an industrial era, where wealth and progress stem primarily from manufacturing. We are now solidly embedded in the knowledge era, a concept formalized and developed by Peter Drucker, as far back as the 1950s (see, for example, chapters 4 and 5 of Drucker, 1957). He realized that, in this new age, knowledge would become the new capital, and that education would undergo a complete revolution in terms of its role in society (Turriago-Hoyos et al., 2016; Wartzman, 2014). This concept may be extended, though, in that knowledge is always *about* something, and since knowledge is the skill of interpreting and using information, knowledge requires information. Or, more to the point, *knowledge requires data*. Thus, having data does not mean having knowledge, but one cannot gain knowledge without data. And so, the true capital of the twenty-first century is not knowledge, but data, and those with knowledge deal in data. Of course, industrial production—and hence manufacturing—will always remain. Just as industrialization did not obviate agriculture, but rather profoundly changed agriculture, especially through the introduction of mechanization to farming and the subsequent creation of super-farms, so also the knowledge era already has profoundly altered manufacturing, and will continue to do so (Thomas & McSharry, 2015).

The primary commodity of the twenty-first century is thus data. And the more data one has, the more one can benefit—provided, of course, the data can be used. Big Data, then, is what it is all about at the moment. And yet, the idea of exactly what constitutes *Big Data* remains a moving target. As will be pointed out, though, the necessity of a precise definition is becoming more of a moot point as time progresses—the real issue is no longer the size of a data set, but its scope and its utility. As Stephens-Davidowitz notes (Stephens-Davidowitz, 2017, p. 62): "The Big Data revolution is less about collecting more and more data. It is about collecting the right data."

At present, Big Data is a concept that appears to have much more traction in business than in academia. Early academic forays into the world of Big Data appeared to be mostly meta-study—for example, defining what constitutes Big Data, methods for working with Big Data, etc. (Youtie et al., 2017). It has taken longer for academics to come to the realization that Big Data itself can be used as a data source for answering their research questions, so much so that Stephens-Davidowitz could still lament in 2017 (p. 274) that "the majority of academics have ignored the data explosion caused by the digital age." Now, however, we are on the cusp of an explosion of research using Big Data (Harlow & Oswald, 2016; King,

2014; Lazer, Pentland, et al., 2014b). This chapter seeks to provide an assessment of the progress made, and the changes that are to come, showing how the wider world's embrace of Big Data impacts researchers and their research studies, across the world, and across disciplines.

A Brief Overview of the Historical Development of Big Data

Big Data would not be possible without computers, which have, thus far (Courtland, 2015), followed what is commonly called Moore's law (Moore, 1965)—that the computing power of silicone semiconductor chips being manufactured would double every 2 years—an exponential growth in computing power. Without these ever more powerful computers, we would not have the tools with which to do the massive number of calculations needed to analyse Big Data.

However, a complementary process, more often ignored, which is vital for the very concept of Big Data, is that before data can be analysed by a computer, they need to be stored electronically. Big Data would not have been possible without digitization. Thankfully, due to a slew of new technologies which have appeared over the years, the tools for data storage have grown in capacity. Figure 1.1 shows the storage capacity, in bytes, of various data storage tools over the years (using information from Anthony, 2011; Data Recovery Group, 2011; Pingdom, 2008). The lines show how select devices (e.g., hard drives) have grown in capacity since

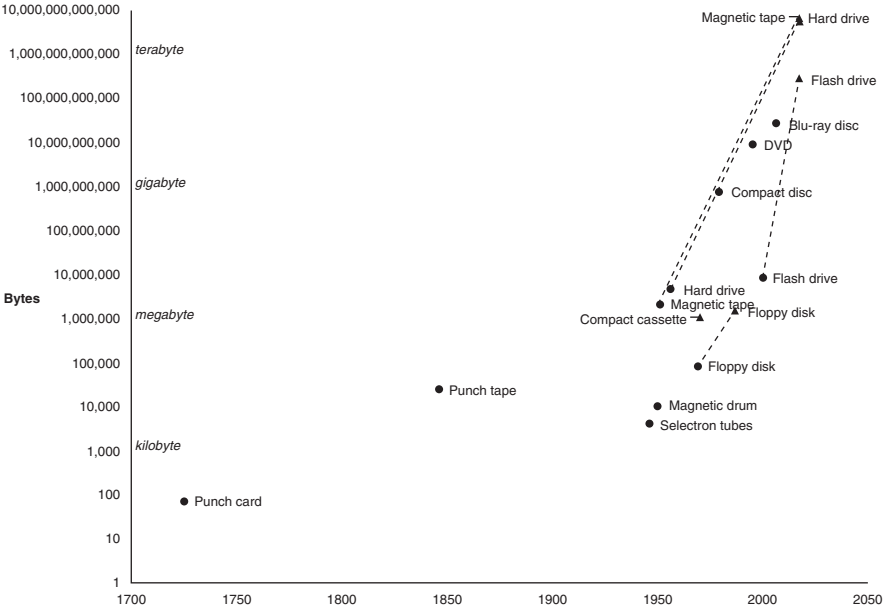

Fig. 1.1 Storage capacity of various data storage devices over time

their initial introduction. What this shows is that the storage capacity of individual digital storage devices has grown massively over the years, but what it fails to show is the rate at which these devices have been manufactured. While it is reported that IBM manufactured 10,000,000 punch cards daily as early as 1937 (Anthony, 2011), some of the more esoteric devices were never manufactured in huge quantities. Today, however, hard drives, flash drives, and optical disks are being manufactured by the hundreds of millions—estimates (e.g., The Statistics Portal, 2018) are for half a billion hard drives alone being shipped annually, although that covers devices over the full range of storage capacities, making estimates of the total storage difficult. Nonetheless, Coughlin (2015) has estimated that the annual storage volume created by hard drive manufacture alone is in the vicinity of 540 exabytes.

Hilbert has studied the rate at which the volume of digitized information is growing, and has made two important observations. First, he defines the current era as the digital era, estimating its inception in 2002, the year in which, by his calculations, more data were stored in digital format than in analog (Hilbert, 2012a, p. 9). Second, his calculation of the total amount of storage space (combining the missing element of the amount of storage devices with their capacity; Hilbert, 2012a, 2012b, 2014, Hilbert & López, 2011, 2012a, 2012b) shows that the amount of available storage is, at least for the present, and probably into the foreseeable future, growing faster than the amount of data being generated (see also Lesk, 1997).

It should also be noted that this growth in the amount of digitized information stems not only from the capture and storage of *new* digital information, but also from the digitization of *existing* analog data. Many examples can be given, but mentioning the digitization of performance recordings (e.g., video recordings, vinyl records) and the scanning and often the optical character recognition of print material (e.g., Google Books, academic journals digitizing old print issues) should suffice.

It is, then, this combination of the ever-increasing computing power, and the ever-expanding storage capacity, together with the increasing tendency to capture and store information, past and present, digitally, that has enabled the Big Data revolution.

However, if the Big Data revolution has been enabled by advances in computing power and storage capacity, researchers should realize that we have entered a new phase, viz., the Big Data explosion. The rapid proliferation of Internet-connected devices, together with the advent of cloud computing (i.e., data being sent directly to, and stored on, the Internet), giving rise to the Internet of Things (IoT), means that potentially anything, and perhaps more intimidatingly, everything, can be become a data-capture device. New companies are being started, offering services solely for the purpose of collecting data. Data itself has become a commodity, and the new economy is an economy of data—Big Data (Thomas & McSharry, 2015).

Massive data sets are literally being generated everywhere, and even academic researchers themselves are involved in creating a massive set of data through their scholarly publications. The questions of how many academic publications are in existence and how many are published per year are complicated by the fact that there are numerous academic publishers involved, and several large (and

competing) cataloguing services. Nonetheless, various estimates (Björk et al., 2009; Jinha, 2010; Larsen & von Ins, 2010; Mabe & Amin, 2001; Ware & Mabe, 2009, 2012, 2015) put the annual growth rate in publications at 3% or above (even as high as 8% for certain fields), so that it was estimated that at the end of 2009, there were over 50 million academic publications in existence (Jinha, 2010). Using these estimates, and an absolute minimum growth rate of 3%, we can safely estimate that at the very latest, 2019 will be the year in which two million new journal articles are published, although Plume and van Weijen (2014) estimated 2.4 million new journal articles in 2013 already.

Where Did the Concept of "Big Data" Originate?

The coining of the phrase "Big Data" is commonly (e.g., Lohr, 2013) attributed to Mashey (cf. Mashey, 1998), who used the term in various presentations and discussions in the 1990s (but see also Dontha, 2017). The concept itself was perhaps first succinctly articulated by Laney (2001). Although he did not use the term "Big Data," per se, he discussed the data challenges large enterprises would face going into the twenty-first century, and described the three dimensions of data as volume, velocity, and variety. These so-called "3 Vs" of Big Data have become a catchphrase. These concepts are still important, especially to academic researchers wishing to work with Big Data, and will be elaborated in the section on the challenges of Big Data to academic research below.

In essence, however, an awareness had long been arising that information workers (including researchers) were being exposed to, and required to work with, ever more voluminous sets of data, and that, apart from the "tools" (the aforementioned computing power and storage capacity) required, special techniques and approaches would be required to make sense of this information (cf. Press, 2013 for a good historical overview).

Interestingly, one of the exciting new sources of Big Data for academic researchers which will be discussed below, Google Trends, can give us a good understanding of how interest in the topic has grown. Figure 1.2 shows a plot of 2040 samples (giving a margin of error of 2.17%) of worldwide search volume from the Google Trends Extended Health API, using a custom-designed data extraction tool developed by this researcher (Raubenheimer, 2021). Curiously, despite the term first being used in the 1990s, and appearing in an academic publication in 1997 (Cox & Ellsworth, 1997), the topic only really gained interest, in terms of it being something people started searching for on the Internet, towards the middle of 2011. This is further corroborated by Youtie, Porter, and Huang's (2017, p. 67) review which "did not find a substantial number of social science Big Data articles until 2012." From there, however, it showed a sharp rise in interest, and has reached something of a plateau in the last 2 years, so that it can accurately be called a buzzword.

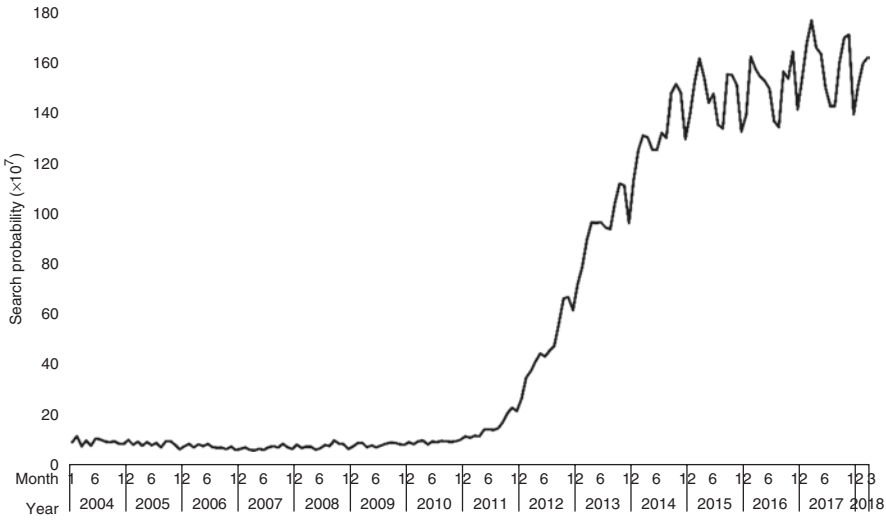

Fig. 1.2 Google Trends searches for Big Data (Topic), 2004–2017

The Custodians of Big Data

To date, the bulk of work on the analysis of Big Data has been conducted in the commercial/industrial sector, where companies gather data and then analyse it for some financial advantage. It is no secret that Google and Facebook, as well as other large technology companies like Apple and Microsoft, use their huge volumes of data to create immense profits. These companies often offer a service for free, with their reward being the data provided by users. They are then able to leverage that data to generate immense profits, most commonly through the sale of targeted advertising (Hachman, 2015; Holland, 2017; Monnappa, 2017; Mullin, 2012). But the list of companies using their data in this way is by no means confined to these more well-established tech companies. Further names in different industries, like Uber, Airbnb, etc., are all alike in their use of large volumes of data for generating profit. And, of course, governments themselves have established processes by which they can amass large volumes of data on their citizens and their behaviour, sometimes contentiously so, under the premise that these data can be used for the betterment of the citizenry. Transport for London, as an example, has shared its London Tube data (gathered via its Oyster card system) with researchers, so that it could use the results of their analyses to increase its own efficiency (cf. Badger, 2014; Chiang et al., 2017; "London's Tube and Big Data: Underground movement," 2013; Sari Aslama et al., 2015). In a similar fashion, Strava, a commercial company which will be discussed below in the section on ethical considerations, has released its Metro data (Gordon, 2014) to "[partner] with departments of transportation and city planning groups to plan, measure and improve infrastructure for bicyclists and pedestrians." Large data sets are thus available, and although not purpose-built for academic research, these data can be used for academic research.

Big Data as a Source for Academic Research

Traditionally, researchers have had to gather their own data. Of course, researchers have long used pre-existing data (e.g., a health researcher using hospital records as a source of data) in their research as well, and so the concept of using data not initially gathered for the purposes of research is not new. However, the difference with Big Data is that while, in more conventional research, researchers might choose between collecting their own data or using already collected data (as a matter of necessity or of convenience), with Big Data, most researchers will, at least for the foreseeable future, have neither the means nor the resources to amass large data sets for their own purposes—they will necessarily be reliant on external sources for large data sets. (There are, of course, exceptions to this, such as the mapping of the human genome, or large volumes of data collected by massive telescopes, but for the majority of researchers, this holds true.)

Thankfully, several sources of Big Data are now being opened up to researchers, often in the hope that by allowing researchers to work with the data, the data owners may gain insights they might not otherwise have come by, perhaps as a consequence of the lack of motivation, time, or ability to extract such insights from their own data. In a sense, this is the best form of collaboration possible. Big companies and government institutions amass data, and share them with researchers, who then provide valuable information back to the data owners. A prime example of this is how Transport for London has shared its London Tube data with academic researchers mentioned above. Researchers are, then, to use Drucker's term, becoming some of the best knowledge workers. Their reward is provided up front—they are given access to data, and the product of their labour is knowledge, through which both they (in their academic careers) and the data custodians benefit.

Challenges of Big Data for Academic Research

At the best of times, research can be challenging. Working with Big Data is no different, although the domain does present some unique challenges.

The Big-Ness of Big Data

Perhaps the best place to start discussing the challenges which Big Data presents to academic researchers is to return to the 3 Vs which Laney (2001) listed as business enterprise challenges. These three points have come to be viewed as the "classical" defining characteristics of Big Data (e.g., Dave, 2013), even though others have sought to elaborate on the concept (e.g., DeVan, 2016; IBM, n.d.-a, n.d.-b; Jain, 2016; Marr, 2014; McNulty, 2014; "Why the 3V's are not sufficient to describe big

data," 2015), and as such, this chapter will retain its focus on the original three points. Given that data may be defined as "big" when they meet *at least* the conditions of the 3 Vs, it would do well to consider each of these as a challenge to researchers wishing to use Big Data for their studies.

Volume refers to the fact that data sets are becoming massive, even more so as data are seen as being inherently valuable, leading to the situation where new data are added, but old data are not discarded. Lesk (1997) noted that our capability to store information has outpaced the already massive amounts of information being stored, with the paradoxical outcome that no data will have to be lost (simply everything can be stored), while most data will not be used (at least not by humans). This presents a twofold challenge to researchers: Firstly, researchers simply will have to learn to use the technology required to deal with voluminous data. As an example, social media produces data sets that are truly massive (Moe & Schweidel, 2017)—Crimson Hexagon allows its clients (which includes researchers) access to over a trillion items of social media data (Crimson Hexagon, 2018; Moriarity, 2017).

The volume of Big Data present many practical challenges for analysis, as detailed by Jacobs (2009). This means researchers will have to use more powerful computers or networked high-performance computing facilities (Doctorow, 2008). At the very least, it necessitates working in a 64-bit, as opposed to a 32-bit environment (e.g., 32-bit programs cannot handle data files larger than 4 Gb, which can easily be found with voluminous data sets). It requires that researchers learn to modify their programming techniques to optimize performance with large data sets (e.g., when using an inefficient sorting method with a typical research data set, the result might be a time difference of a second or two, but with a massive data set, the result of an inefficient sorting algorithm could be hours of additional computing time). Stephens-Davidowitz (2017, p. 180) casually mentions that "one day I decided to download Wikipedia. (You can do that sort of thing nowadays.)" Given that Wikipedia (2018) itself estimates the size of its text-only dumps at 58 Gb (when uncompressed—the actual compressed download file is a mere 14 Gb!), this is a clear example of the volume of data that one may have to work with.

Secondly, and more importantly, the true challenge of volume, given Lesk's conclusions mentioned above, is not the computational challenge of dealing with millions or even billions of records of data per se, but rather the challenge of finding the information we need amidst a growing amount of irrelevant and redundant data. The size of the needle is not changing, but the size of the haystack is growing exponentially (Taleb, 2013). Stephens-Davidowitz (2017, p. 21) notes that "The size of a dataset… is frequently overrated…. You don't always need a ton of data to find important insights. You need the right data…. Most important… *you have to ask the right questions*" (emphasis added). To meet this challenge, researchers will have to become innovative and learn to think outside of the box, to use another management catchphrase. As will be discussed below, working with Big Data requires rethinking even the questions we want to ask, so that we can ask the questions that are suited to what the data can give us, instead of banging our heads against a wall, trying to force Big Data to give us answers it cannot yield.

Laney's (2001) second concept of *velocity* refers to the fact that data sets are growing continuously. An analysis of data today will include information not available yesterday. A good example of this is Google Trends data: Firstly, the data are updated in near real time, so that an investigation into the popularity of searches will continuously provide new (never ending) information. As a side note, it should be pointed out, especially in further discussions of Google Trends below, that even Google scientists contended that the data may not be suitable for the forecasting of future events, but may be useful for contemporaneous forecasting (cf. Choi & Varian, 2009, 2012). Furthermore, even for historic data, Google samples 10–15% of all searches, and resamples it each day, so that repeating an analysis of the same search terms for the same time frame on successive days will provide a researcher with a sample of samplings for that query (Matsa et al., 2017). To meet this challenge, researchers will have to accept the time-boundedness and the immediacy of their work. Time is an enormous confounder in any research, and even more so with Big Data, because the data may change considerably over the longer time frames common in large data sets. More will be said about this when discussing the variety of data, but the challenge of velocity is that researchers will, in many instances, have to think about how their studies can be tailored to make use of new data as it arrives, often in pursuit of making predictions based on real-time data. This is no simple task, no matter how much data one has at one's disposal.

Laney (2001) was convinced, rightly, that the largest challenge to data management was the *variety* of data. Primarily, this refers to the fact that data come in so many different forms. Disparate data sets are "incompatible… non-aligned… and inconsistent." However, this extends beyond just incompatibilities between types of data, and even incompatibilities between data sets of the same type, to inconsistencies with the same sets of data over time—data definitions, variables captured, and even data location may change. A good example of this are the data provided by the United States National Highway Traffic Safety Administration's Fatality Analysis Reporting System (http://www.nhtsa.gov/FARS). The user guide (National Highway Traffic Safety Administration, 2016) explains how variables were added, removed, and changed, over the years (e.g., accidents due to cell phone usage were obviously not part of the data recorded in 1975). Also, the user guide (NHTSA, 2016, pp. 520–585) details the ways in which the very structure of the FARS data sets was changed profoundly in 2010 so as to improve its compatibility with the National Automotive Sampling System General Estimates System data.

Perhaps the best comment about the variety inherent in Big Data comes from Lazer, Kennedy, King, and Vespignani (Lazer, Kennedy, et al., 2014a, p. 1204), who, in discussing the failure of Google Flu Trends to maintain its predictive accuracy beyond the original model development (as discussed below), note that,

> A more likely culprit is changes made by Google's search algorithm itself. The Google search algorithm is not a static entity—the company is constantly testing and improving search. For example, the official Google search blog reported 86 changes in June and July 2012 alone…. Search patterns are the result of thousands of decisions made by the company's programmers in various subunits and by millions of consumers worldwide.

The variety inherent in Big Data means that researchers will not be able to work with Big Data and stay armchair theorists. They will have to get their hands dirty, and become data wranglers, able to beat unruly data into shape, as it were (of course, without altering or fabricating data!), struggling with incompatibilities in various data sets, and developing a clear understanding of the intricacies and nuances of their data, so that they can wield it appropriately.

Big Data are Not Research Data

One of the biggest problems faced by researchers wishing to use Big Data for their own research work stems from the fact that the data are often not gathered for the sake of research, but rather for the specific purposes of the data custodians. Of course, as has been mentioned, this is not a new challenge, but is faced by any researcher wishing to use externally collected data for their work. Big Data, however, exacerbates this problem. Essentially, Big Data provide us with masses of information but the information may not be ideally suited to our research questions. Lazer, Kennedy, King, and Vespignani (Lazer, Kennedy, et al., 2014a, p. 1203) note that "the core challenge is that most Big Data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis." Lazer et al. (Lazer, Pentland, et al., 2014b, p. 1203) then warn against "Big Data hubris" which they define as "the often implicit assumption that Big Data are a substitute for, rather than a supplement to, traditional data collection and analysis." Stephens-Davidowitz (2017), after noting earlier (p. 245) that there are things that cannot be done with Big Data, and things that ought not be done with Big Data, points out that Big Data do not supersede, but rather complement, conventional research. He notes (p. 256) that having access to Big Data for our research "… does not mean that we can just throw data at any question. And Big Data does not eliminate the need for all the other ways humans have developed over the millennia to understand the world. They complement each other." This means that Big Data do not spell the end of small data. In fact, many times, research questions can only be answered successfully when insights from Big Data are supplemented with insights from small data, such as combining Big Data analysis results with small survey results. This combination of disparate data sources must be done with care, however. The process is well illustrated by Mellon (2013) who demonstrated the use of survey data to validate Google Trends search data and concludes (pp. 289–290):

> Rather than replacing the traditional survey, using Internet data may require researchers to do more survey work, particularly in hard to reach areas. Paradoxically researchers may only be able to unlock the potential of the Internet data being generated today, in retrospect, when they have enough data generated with conventional survey tools to test the validity of search data.

An extension of this issue (that Big Data are not created for specific research problems) is the issue of dimensionality, or overfitting (Clarke & Chien, 2017). A problem may arise when our dependent variables are themselves not massive, such that many independent variables are used to predict dependent variables of limited scope (Stephens-Davidowitz, 2017, p. 247), confounding analysis and prediction. Big Data tend very easily to gather information about too many variables but not enough subjects. This may not be immediately apparent, as we naturally assume that Big Data contains information about many people, but when our dependent variable is a rare condition, even vast data sets may easily contain more variables than subjects (Salsburg, 2017, pp. 69–77; Taleb, 2013). A classic example of this is the Google Flu issue, as clearly indicated by Lazer, Pentland, et al. (2014b), p. 1203) when they explain that "essentially, the methodology was to find the best matches among 50 million search terms to fit 1152 data points…. The odds of finding search terms that match the propensity of the flu but are structurally unrelated, and so do not predict the future, were quite high." The history of Google Flu is as follows: Building on work of Johnson et al. (2004) and Eysenbach (2006), Polgreen et al. (2008) found that they could use Yahoo! search queries related to flu, to predict the incidence of flu. This was swiftly followed by Ginsberg et al., who, after an announcement on the Google blog (Ginsberg & Mohebbi, 2008), followed with a paper in *Nature* (Ginsberg et al., 2009) showing similar results from Google searches. This led to the creation of the Google Flu Trends web page. Initial results seemed to show that the algorithm worked reasonably well (Araz et al., 2014; Carneiro & Mylonakis, 2009; Dugas et al., 2012, 2013; Malik et al., 2011; Valdivia et al., 2010; Wilson et al., 2009). But over time, this ability degraded (Butler, 2013; Lazer, Kennedy, et al., 2014a; Olson et al., 2013), nor did it perform well when tested against retrospective data (Ortiz et al., 2011), requiring adjustments to the algorithm (Cook et al., 2011; Copeland et al., 2013). The project was labelled as a failure (Butler, 2013; Fung, 2014; Lazer & Kennedy, 2015; Salzberg, 2014; Walsh, 2014), and eventually the project was shuttered (The Flu Trends Team, 2015). The correlation they found for one season did not hold well across many seasons. This means that researchers using Big Data for forecasting would do well to use out-of-sample predictions to further validate their findings, and the insights from Big Data might best be understood when validated with small data. It should be noted, though, that right from the outset, and even noted by its critics, Google Flu Trends was best seen as a way to supplement, not supplant, the conventional epidemic surveillance systems (Lazer, Kennedy, et al., 2014a, p. 1204; Madrigal, 2014; Valdivia et al., 2010, p. 5; Wilson et al., 2009, p. 4).

A last implication of the fact that Big Data are not gathered explicitly for research purposes, is that the quality of the data may be a double-edged sword. On the one hand, the data may be surprisingly free of conventional biases that confound research, such as placebo effects, Hawthorne effects, and a litany of other biases researchers have to deal with. But at the same time, the data may present new biases of their own, biases, perhaps, that researchers are less accustomed to dealing with. So, for example, Stephens-Davidowitz (2017, p. 111) notes that "Google [data] can display a bias towards unseemly thoughts," and researchers using social media data,

need to be aware of, and account for, the biases inherent in these data (Moe & Schweidel, 2017). Most data gathered from the Internet contains an inherent bias in that it is restricted to Internet users, whose demographics still do not adequately represent the world population, or any country's population, even though Internet usage is growing, and Internet access is being considered as fundamental to the exercise of basic human rights (Human Rights Council of the United Nations, 2016; La Rue, 2011). But even more subtle in its effect, and hence more difficult to control for, is that different demographics of Internet users use the Internet in different ways, with, for example, different age groups showing preferences for different social media networks (Nanji, 2017). Furthermore, a variety of "socioeconomic, political, cultural, social, and technological factors" all help shape Internet access at the national level (Fuchs, 2008).

Data Linkage

Although this could have been mentioned when discussing the variety of Big Data above, this topic is complex enough to warrant a brief discussion. A surprisingly common problem encountered when working with Big Data stems from the previously discussed fact that large data sets are most commonly created for commercial reasons, and are not bespoke research data sets. When researchers plan a conventional study, they go to great pains to ensure that their data collection encompasses all the variables needed to answer the research questions. However, it is not unusual for researchers using Big Data to require variables from multiple, disparate data sets in order to answer their research questions. The only way this can be done is to link the data in some way, so that every record for each individual in one data set is successfully matched with every record for that individual in each of the other data sets. This quickly becomes a very complex exercise, as, for example, the identifying information for each individual may differ in the various data sets. A variety of data linkage methodologies have arisen to deal with this problem (Bohensky et al., 2010; Churches & Christen, 2004; Clark, 2004; Dusetzina et al., 2014; Fellegi & Sunter, 1969; Ong et al., 2014; Tromp et al., 2011; Winkler, 1993, 2014), but the level of specialisation required means that the exercise is often out of reach of the individual researcher. The obstacles are many: Data have to be obtained from multiple custodians, each of which may have their own ethical requirements as to the usage of their data. There are ethical issues around the linkage of data (e.g., personal identification may be impossible in the individual data sets, but could become possible in the linked data; or, conversely, de-identification for ethical purposes may make the data difficult to link). The exercise is typically laborious, time-consuming, costly, and requires significant computing power—even when specialist service providers are used to link the data, time delays are significant, and costs are high. While government institutions (such as census bureaus) may have the personnel and skills to perform these functions themselves, academic institutions typically do not. Thankfully, collaborative efforts may often arise, which can result in the creation of

data linkage services, such as the NSW Centre for Health Record Linkage (CHeReL) in Australia. Examining their guide (Centre for Health Record Linkage, n.d.) gives one a clear understanding of the process involved. The implication for researchers, though, is that the linking of data requires a considerable amount of time and ethical approval, and involves significant cost. This is, however, the reality of attempting to work with data of this nature.

Furthermore, working to analyse these linked data involves a significant step up in terms of the complexity of the statistical programming required. Small data studies very commonly use data sets where the entire study utilizes only a single data set with a 1:1 participant-to-record ratio. The most complex extension of that is longitudinal studies where multiple records are collected for each individual, but still, typically, within a single data set, or with separate data sets for each time point, but with each time-bound data set containing a single record for each participant. Working with these data remains a fairly simple matter. Big Data studies, however, are very often longitudinal (it takes time to amass so much data), and routinely entail working with multiple 1-to-many or many-to-many participant-to-record data sets. The complexity of the way in which records from these data sets are linked, then, grows in order of magnitude for each additional data set, and techniques are needed for determining temporal start and end points for each individual, extracting first and/or last records for each individual from any number of data sets, aggregating records within, and even between, data sets, and identifying individuals with more or less than a certain number of records. The problem of missing data and its underlying reasons also grows in complexity, and the statistical techniques needed to analyse these data also become more complex.

Big Data May Still Not Deliver Generalizable Results

Even though it is hoped that the results we derive from Big Data studies will lead to translational research, the very nature of Big Data presents a fundamental obstacle to this. The ideal flow of events with conventional medical studies is that when researchers perform clinical trials that show the benefits of certain interventions (e.g., health screenings), and these results are confirmed in further studies, the evidence would be sufficient to recommend the intervention to everyone in the risk category, even though we know that not all at-risk individuals will truly be afflicted over time. But the results of these studies are generalized to the population nonetheless.

However, generalizability is one of the largest challenges of Big Data. Big Data analyses are observational, and can, as an example, indicate increased odds that certain individuals might display a large degree of suicidal ideation. But on data sets of millions of individuals, even with highly significant odds, that would leave a large number (in real terms) of individuals identified as false positives. Even if Big Data analysis results are generalizable to populations at large, they are not specifiable to individuals in particular. The point is that while Big Data insights can help

us understand dynamics in populations at large, we should still refrain from making inferences about individuals in particular. As Agatha Christie's character Sir Henry Clithering notes in "The Four Suspects" (Christie, 1932): "It isn't really guilt that is important—it's innocence. That's the thing that nobody will realize…. Some people may go through a lifetime crushed by the weight of a suspicion that is really unjustified." This can perhaps best be understood by means of an illustration: In the short span of August 2017 to January 2018, two women, one in Australia (cf. Chillingworth, 2018) and one in the U.S. (cf. Hitt, 2018) searched the Internet for information on how to murder their partners. While it is apparent, from these, and other cases (e.g., Deery, 2013; also mentioned in Stephens-Davidowitz, 2017, p. 266) that would-be murderers actually are brazen enough to search for tips on the Internet, and while Big Data allow us to immediately flag any individual carrying out such searches, it would be wrong to immediately prosecute, or even place under surveillance, any individual carrying out such searches (à la *Minority Report* [Spielberg, 2002], as Stephens-Davidowitz, 2017, p. 266] intimates). If we ever were to do that, then fiction novelists (researching their next murder mystery), journalists (researching articles about murderers who google their intended methods), and even academic researchers (investigating the uses of Big Data) also might be unfairly accused.

Spurious Correlations, Sub-Populations, and Other Confounders

The results from the analyses of Big Data are, simply by nature of their size (and thus the large diversity of their underlying populations), particularly susceptible to the two extremes of delivering spurious correlations on the one hand (as already discussed on the issue of dimensionality above), and being bedevilled by numerous confounders on the other. Academics working with Big Data will need a good level of statistical literacy in order to make the right inferences (and only the right inferences) from their data (Prodromou & Dunne, 2017). These skills, as well as the necessary software programming skills for analysing these data, need to be incorporated into the curricula of university degrees (Puang-ngern et al., 2017).

The study of suicide and the Internet serves as a good example. Several studies appear to have found evidence of a correlation between suicide rates and Internet searches related to suicide and self-harm (mostly analysed using Google Trends): Yang et al. (2011) in Taipei; Hagihara et al. (2012) in Japan; Gunn and Lester (2013) and Ma-Kellams et al. (2016) in the US; Arora, Stuckler, and McKee (2016, in certain subgroups of the population), and Kristoufek et al. (2016) in the UK; and Song et al. (2014) in South Korea. Other ancillary findings also support this idea: Ayers et al. (2013) found evidence of seasonality in searches related to suicide (and other conditions). Biddle et al. (2012, 2016) and Gunnell et al. (2012) examined the growth of suicide-related information on the Internet. Bragazzi (2013) found a strong correlation between Google Trends searches and non-suicidal self-injury.

However, the results are not as clear-cut as they may seem at first. McCarthy (2010) found that the correlation between completed suicides and Internet searches is inverse for the population as a whole, but strongly positive for youth, suggesting that Google Trends may only serve as a predictor for certain demographics. Even one of the aforementioned studies—Ma-Kellams et al. (2016, p. 483, emphasis added)—noted the following:

> By capturing ecological variation in suicide-directed Internet behavior, Google search data can be more useful than random-survey-sampling techniques because Google search data capture a broader proportion of the population in an unobtrusive way. This system is *not meant to be a replacement for traditional surveillance mechanisms* or to supplant the need for other forms of suicide surveillance. Rather, data on notable increases in suicide search activity can be used to prevent suicide by indicating where the needs are for closer suicide surveillance… or by reinforcing the need… for more sensitive programs among particular cultural groups.

Furthermore, Fond et al. (2015) found a relationship between suicide searches and depression (but not bipolar) searches, but noted that search volumes could be influenced heavily by external factors, such as media reports. They concluded (p. 917) that "Google Trends® cannot be used as an effective tool for monitoring potential suicide epidemic peaks, because of the lack of precision of existing tools and data provided. However, this tool can be further developed and offers great potential." The strong influence of media reports on search volume was corroborated by Page et al. (2011) in Australia, who noted that Google Trends could not serve as a predictor on its own. They did note the difficulty of comparing the online data, which is available in near-real time, to official mortality data, which has a significant delay. Additionally, Solano et al. (2016) could not find any predictive ability of Internet searches for completed suicides in Italy (by contrast, the searches tended to lag behind completed suicides), nor could Sueki (2011) in Japan. Bruckner et al. (2014) also reported mixed results, and concluded (it should be noted that their comment also touches on the previously mentioned issue of extrapolating from the general to the individual):

> We caution against inferring an individual's Internet search behavior, depressive symptoms, and suicide risk from our population-level tests. Temporal fluctuations in Google searches are indicators of a population-level interest, which cannot shed light on individual differences in suicide risk. In addition, our findings of a positive relation between Internet search volume for "suicide and depression" and completed suicide in the same month cannot definitively establish that suicide-related searches precede suicide completion. Nevertheless, our findings indicate that current surveillance efforts for suicide may benefit from real-time use of Google-based Internet searches for depression.

Mars et al. (2015) performed a very instructive small data study. Using survey responses in the UK, they found that adolescents with a history of self-harm were significantly more likely to have searched for information on self-harm on the Internet (although no causality could be inferred). But their study also showed that suicide-searchers additionally use the Internet to search for help, and thus making inferences based on searching for information about suicide alone fails to account adequately for the reality of a very complex issue.

Perhaps the most instructive application of Google Trends suicide data as an example of the utility of Big Data, is this: Noting the association of Internet searches with suicide, both Kristoufek et al. (2016) and Parker et al. (2017) found that they could use Google Trends data to predict suicide rates accurately enough for policy decision-making purposes, while conventional data are slow to be made available (e.g., coroner's statistics that could be up to 2 years behind). But this indicates a limited scope for the application of this data. Finally, in the most rigorous of the studies, Tran et al. (2017) performed a multi-country investigation with strict methodology to account for spurious effects in the time series analysis, and found that the associations found did not follow any discernible pattern, and were in line with the Type-I error rate for the expected number of significant results, indicating that Google Trends data may well be too afflicted by spurious correlations to accurately predict suicide rates.

Results from the analysis of Big Data, then, need to be interpreted with caution. While Big Data may provide useful insights, the application of these results should not be overstated and the susceptibility to spurious effects which may be present in the data purely because of the size and nature of the data set, must always be acknowledged, and used to temper hubris about what researchers find.

Big Data Gathering Is Not Passive

Reference will be made below to the fact that Big Data are sometimes gathered in the form of non-consensual (or at least, unconsciously consensual) randomized trials. What is relevant here is to raise the very large temporal confounder of how Big Data gathering actually changes the behaviour of individuals. Vespignani (2009, p. 428) discusses the implications of this:

> An interesting and ethically challenging aspect of predicting and managing the unfolding of catastrophic events in techno-social networks is the system's adaptation to predictions when they are made publicly available. Social behaviors react and adapt to knowledge of predictions. Contrary to what happens in physical systems, the predictions themselves are part of the system dynamic. In addition, predictions may point to unethical control and anticipation strategies favoring specific demographic sectors of the society. Finally, the risk of erroneous predictions may lead to costly or unethical social control mechanisms with no actual benefits.

On a more practical level, the implication for researchers is that the nature of their data may change over time. Lazer and Kennedy (2015), discussing Google Flu Trends, point out how changes over time to the way the Google search page provides additional search terms (e.g., in its autosuggest feature) and recommends certain searches, may cause people to favour the suggested terms instead of the term they had intended entering, which results in temporal changes to the actual data collected. Google itself, ironically, may have been the greatest culprit for the failure of Google Flu Trends, because Google has not only allowed people to search the Internet for content more efficiently, it has also changed the very way in which we

search (and, it can be argued, continues to do so). On a different level, it can additionally be argued (although less easy to prove substantively), that in the relatively short lifespan of Internet searching—although Google was not the first search engine, none can argue that it came to define Internet searching, and the recent 20-year anniversary of the incorporation of Google (September 4, 1998, see Brin & Page, 1998)—the number of people conducting searches, the things people were willing to search for, and the very way in which people searched the Internet has changed, making any comparisons over longer time periods fraught with difficulty.

Ethical Considerations with Big Data

Big Data offers unique ethical complications to the researcher. The issue it appears to resolve is anonymity: When data from millions of individuals are examined in bulk, no finding can be traced back to a specific individual. But Big Data are sometimes not truly anonymous. A typical example of this is found when researchers analyse social media posts, where identifying information may still abound (as users typically do not adequately control their privacy settings). Thus Big Data may contain very personal, very identifying information. And this remains an ethical problem. A clear instance of this has been demonstrated recently in the Strava global heat map. Strava is a software company offering a mobile app and a website interface; it manufactures no physical products. The company gathers geolocation information from users of its training app on their mobile devices, a service that exemplifies data collection in the Big Data era. While the app offers a free service to millions of people worldwide (or enhanced services for a fee), there can be no doubt that the value Strava derives is from the data it harvests from the app users. Strava has used those data, amongst others to release a global heat map (https://labs.strava.com/heatmap)—a very high-tech global visualization of the exercise routes and patterns of their users (Robb, 2017). And while, in theory, the map does not allow identification, in practice, it does. As a hypothetical example, if only one person in a rural area has used Strava, then anyone viewing that area on the heat map would be able to identify that person's routes. This may be less problematic than the fact that a student of the International Security Studies programme at the Australian National University was able to identify a surprising amount of detail about US military bases in Syria and other conflict zones, using the Strava heat map (Bogle, 2018). Strava has since corrected this, and allowed users to opt out of having their data included in the heat map, and has suppressed very low volume activities. Thus, academics working with data provided by a commercial data custodian may easily run into ethical issues not foreseen at the inception of their studies.

A further ethical problem is that research is meant to be translational. We do research with the intention that it will, at the very least, make the world a slightly better place. But the results of Big Data analyses may lead to unintended consequences. This is especially so when attempts are made to relate the general findings derived from a Big Data analysis to individuals, something which, it was pointed

out above, should not be done, based purely on the nature of the analysis and the data, let alone the ethical considerations. Nonetheless, Stephens-Davidowitz (2017, pp. 260–1) discusses the possibilities of a world in which inferences like this might be attempted, and notes that "better prediction can lead to subtler and more nefarious discrimination." In the world of online business, this could easily lead to data wars, where businesses use Big Data to extract maximum profits from consumers using strategies like targeted marketing and differential pricing, while consumers will fight back using Big Data for price comparison, service review, and information sharing. This is a struggle in which academic researchers using those data could easily become embroiled.

A further ethical issue with Big Data stems, once again, from the fact that Big Data are ordinarily not gathered for the purposes of research: The individuals whose data are contained in the data set were not given the opportunity to consent for their data to be used in the research study—such as the people who searched for information on committing suicide discussed above, or people who proved to be thought leaders in fanning the flames of xenophobia in the Citizen Research Centre (2017) study. This is indeed a problematic area. The Google Trends data on suicide have been anonymized, and cannot lead to individual identification. Social media data, on the other hand, contain whatever identifying information is allowed by the individual's privacy settings, which means, in practice, that many people who either do not understand how to, or do not care to, set privacy settings have their personal information available to any entity (research or corporate) who scrapes that information from the social networking site. Admittedly, the terms of service of the various social networking sites—which, it should be noted, most people never read (Bakos et al., 2014; Obar & Oelof-Hirsch, 2016)—work along the general principle that whatever is not set to be private is considered to be public domain. And it is also so that the trade-off for using a social media site for free is that the site owners are given access to one's data (cf. Hachman, 2015), to do with pretty much as they please (within reasonable limits), which may include sharing that data with researchers, or using the data for research of their own. And so, the analysis of both Google Trends data and social media data is well within what is permitted by the terms of service of the sites on which the data were gathered. But around the fringes, some of this research definitely traverses into ethically grey areas.

What is even more frightening, though, and is definitely on the fringe of the grey ethical areas, and sometimes well beyond, is that users of these sites often find themselves being actively experimented upon. The simplest form of this is what is known in the information technology world as A/B testing. Essentially, A/B tests are elementary randomized controlled trials, and they are being performed on users of Internet sites continuously, so much so that Facebook reportedly performs over 1000 A/B tests *daily* (Stephens-Davidowitz, 2017, p. 211). Given Facebook's user base of over a billion people, this means that Facebook is actively experimenting on at least tens of millions, if not hundreds of millions of people, daily. It must be said that many of the tests are innocuous, such as determining which colour or font of a control users respond to best (of course, it should be noted that the ultimate aim of much of this testing is to maximize user engagement with the site, which, when

stripped of its technobabble, means making the site as addictive as possible). However, this experimentation can take on a decidedly darker and more sinister colour: After only two studies (Coviello et al., 2014; Ugander et al., 2012), in which they did not account fully for the issue of dimensionality in their data, Facebook engineers believed they had sufficient evidence of the way interactions on their platform affected the emotional states of users. On the basis of this belief, they embarked on actual and purposeful experimental manipulation of the emotional state of millions of their users (Kramer et al., 2014), totally unbeknownst to those users, and without any consent other than the generic consent given in the Facebook terms of use. This elicited a storm of criticism (e.g., Blue, 2014; Vaughan-Nichols, 2014), and eventually garnered an editorial expression of concern (Verma, 2014). The truly terrifying element from this is that the editorial statement notes that "because this experiment was conducted by Facebook, Inc. for internal purposes, the Cornell University IRB [Institutional Review Board] determined that the project did not fall under Cornell's Human Research Protection Program." What this means, for all practical purposes, is that as a private company, Facebook staffers may still be pursuing research of this nature, although they are likely to be less transparent about it and less likely to publish their findings in academic journals, because they feel that it "may not have justified all of this anxiety" (Kramer, 2014).

Recent events from the first half of 2018 have revealed just how easy it can be for academic researchers to become embroiled in the ethics minefield surrounding big data. Against the backdrop of an ongoing investigation into how the Facebook platform was possibly manipulated to influence the outcome of the US election (Entous et al., 2018), a Cambridge University psychology researcher established an outfit named "Global Science Research" through which he developed an application that conducted an online personality test. He purportedly received funding from a company known as Cambridge Analytica to pay Facebook users a nominal fee for taking the personality test as part of his data gathering strategy (Lewis & Carrie Wong, 2018; Ortutay et al., 2018). Unbeknownst, however, to those Facebook users, was the fact that his application was also harvesting both their data, and all the data of their friends, so that an eventually estimated 87 million or more Facebook users' data was harvested (Hern, 2018; Ortutay, 2018), and turned over to Cambridge Analytica as part of the funding agreement, while Facebook had reportedly turned a blind eye to this kind of data harvesting from its platform (Lewis, 2018). The researcher claimed that he had used the full $800,000 paid by Cambridge Analytica to pay the initial 270,000 test takers (from whose wider Facebook networks the remaining data were harvested), indicating that he had not profited at all, but that his "motivation was to get a dataset [he] could do research on" (Ortutay et al., 2018). Cambridge Analytica, for its part, had apparently used these data in providing marketing services for politicians in a number of elections, including the aforementioned US election (Ortutay et al., 2018).

The matter caused even graver concern when it was revealed that Facebook itself was, in fact, harvesting not only the data of its users, but even the data of non-users (who obviously had not consented to such data being collected) who, for example,

had visited pages on other websites (i.e., not belonging to Facebook) but which contained Facebook Like or Share buttons (Ingram, 2018).

Cambridge Analytica and Global Science Research eventually shut down over the controversy (Reuters, 2018). Facebook has, under pressure from various directions, committed itself to labelling politically-paid advertising content (Leathern, 2018), and removed the "Trending news" feature from its pages in June 2018. However, the biggest outcome of this, for academic researchers, is that Facebook announced a new research initiative through which it hopes to empower academic researchers to independently set the agenda for, and work on solutions towards, misinformation, political interference, and other abuse of its platform (Schrage & Ginsberg, 2018). Importantly, the initiative contains stringent guidelines to counter "the threat presented by the recent misuse of Facebook data, including by an academic associated with Cambridge Analytica" and is built on King and Persily's (2018) model for industry-academic collaboration, an important work for any academic researcher wishing to work with social media Big Data. As will be seen below, Gary King is at the very forefront of creating opportunities for academic researchers to work with Big Data in the social sciences.

In summary, while Big Data may present wonderful opportunities, researchers would do well to consider carefully the full ethical implications of the work they do and the data they work with. A consideration of all the challenges posed by Big Data for academic research may prove disheartening, and lead one to wonder whether working with Big Data is worth the effort. It would do well, then, to consider some of the opportunities provided by Big Data.

Opportunities of Big Data for Academic Research

Big Data present us with many opportunities. We should avoid the "Big Data hubris" (Lazer, Pentland, et al., 2014b) which could beguile us into thinking that Big Data research could supplant existing "small data" research, but neither should we devalue insights gained from Big Data. Big Data will allow us information that can either supplement traditional survey results, providing additional nuance, or information that may contradict survey results. Even this is good, because it will lead us to the important question of which results should be believed, the deliberation of which should lead us to a clearer understanding of the reality we are attempting to uncover. In their article discussing one of the most widely reported failures of Big Data analysis (Google Flu Trends), Lazer, Pentland, et al. (2014b, p. 1205) conclude by pointing out that "Big Data offer enormous possibilities for understanding human interactions at a societal scale, with rich spatial and temporal dynamics, and for detecting complex interactions and nonlinearities among variables." Despite all the complications surrounding Big Data, Big Data still offer incredible opportunities that deserve mention.

First and foremost, new frontiers are opening up for academic researchers. Big Data often allow us to discover heretofore unknown things. In fact, Lazer, Pentland,

et al. (2014b, p. 1205) encourage scientists to "use Big Data to understand the unknown." It should be noted that one of Lazer's co-authors is Gary King, of the Institute for Quantitative Social Science (IQSS) at Harvard University, and also one of the founders of Crimson Hexagon (now Brandwatch), an exciting company for researchers—especially researchers in the social sciences—wanting to explore the Big Data world of social media. The work done at the IQSS (e.g., Hopkins & King, 2010) was spun off into a commercial entity—Crimson Hexagon (Breese, 2015; King, 2011)—which now holds several patents for text mining and sentiment analysis (e.g., US20160189171 A1, 2016, US 9483544 B2, 2012, US20160070748 A1, 2017). The company monitors all major social media and blogging networks, continuously harvesting their data into a database that now contains over a trillion social media posts (Crimson Hexagon, 2018; Moriarity, 2017), which can be analysed using their ForSight™ tool. The company offers its services for a fee, and its major clients are corporate conglomerates who use its real-time data gathering capabilities for brand management and campaign tracking. However, the tool is also being used for academic research (Breese, 2015), and when one considers the cost of figuratively reinventing the wheel (King, 2014) and trying to access the same information, let alone develop an alternative to their patented methods, the fee pales in comparison (although there are alternatives—Chan et al., 2017). The tool has been used in a number of compelling studies, such as a study of how people deal with crisis events (Bair, 2016), public discourse on nanotechnology (Runge et al., 2013), encryption (Addawood et al., 2017), net neutrality (Faris et al., 2015), nuclear energy (Kim et al., 2016; Li et al., 2016), xenophobia (Citizen Research Centre, 2017), food prices (UN Global Pulse, 2014), fracking (Hopke & Simis, 2017a, b), and even as a surveillance tool for eye disease (Deiner et al., 2016) and suicide (Ueda et al., 2017). Big Data, and Crimson Hexagon (now Brandwatch) in particular, are opening up the frontier of social media to the world of the academic researcher.

These new frontiers indicate that Big Data allow us—compel us—to study heretofore inaccessible research questions. Big Data can perhaps best be employed to answer questions where conventional methods are failing (although the caveat may be that the failure of conventional methods may only become apparent when their results are contrasted with those obtained from the analysis of Big Data). However, it is not only that Big Data bring new insights to old questions. Far rather, academics wishing to employ Big Data in their research should be fundamentally rethinking the questions they need to be studying with those data (e.g., Vespignani, 2009). An insight into this comes from one of the early pioneers of academic research using Big Data, Raj Chetty. He notes (in Stephens-Davidowitz, 2017, p. 173): "Big Data is not just about doing the same thing you would have done with surveys, except with more data. Big Data really should allow you to use completely different designs than what you would have with a survey." Big Data have the ability to upend even very established modes of research. Even medicine, with its strong focus on clinical trials as a gold standard, and evidence-based medicine (but see Smith & Pell, 2003 and the ensuing debate), will see changes resulting from Big Data. Ayers et al. (2014) provide compelling arguments for ways in which behavioural medicine,

powered by Internet Big Data, can open new frontiers in medical science (see Emery et al., 2014 for a good example).

However, Big Data are not only redefining our research questions, but are also redefining what constitutes our research data, opening up endless new possibilities. Stephens-Davidowitz (2017, p. 103) notes that "These days, a data scientist must not limit herself to a narrow or traditional view of data…. Everything is data!" An example of a surprising use of data is the attempt to predict economic growth with night light visible from space (Henderson et al., 2008). In fact, it may be this very redefinition of what we are using as data that compels us to revise the questions we should be asking of that data. But, this redefinition of data also extends to other areas. Stephens-Davidowitz refers to Big Data as "honest data" (2017, pp. 54, 105–163). His enthusiasm may have been dampened if he had googled "Fitbit cheat," but the truth of the matter is that conventional survey methodology can be (although is not necessarily) significantly biased (e.g., Fish & Russell, 2017; Kamenetz, 2014; Przybylski, 2016; Robinson-Cimpian, 2014). However, while Big Data can give surprisingly unpretentious data, it may also be affected by large amounts of noise and bias (as Stephens-Davidowitz does acknowledge when he discusses social media data, p.150–3). The task of the researcher is to bear this in mind when considering the nature of the data. For example, Fitbit general motion data may be more biased than Fitbit exercise data, and data gathered from devices only activated and used during exercise (e.g., Garmin fitness watches or Strava apps) may be even more unbiased.

Although it was mentioned that research translation can be a challenge when using Big Data, it should be mentioned that when the right data are used to investigate the right questions, Big Data may precisely present opportunities for translation that were not available through small data studies. Big Data *can* give us insights that lead to translation. Stephens-Davidowitz (2017, p. 163) notes that "collecting rich data on the world's problems is the first step towards fixing them." So it is hoped that some of the most intangible problems which have always failed analysis by conventional research methods, might find new solutions through the use of Big Data. The examples of how Big Data are being used to improve and transform transportation networks, mentioned earlier, show how Big Data can lead to real-world changes.

One of the biggest advantages of Big Data is that they finally overcome one of the Achilles' heels of so many research studies (but only in some areas of research, it warrants adding): Sample size. The sheer volume of data means that, if dealing carefully with the issue of dimensionality mentioned above, researchers may still find sufficient data points in Big Data to study even exceptionally rare phenomena. Stephens-Davidowitz (2017, p. 171) makes the simple observation that "you need a lot of observations in a dataset in order to be able to zoom in with clarity on one small subset of that data," and then goes on to contend (p. 197) that "the world is complicated…. These connections and relationships… cannot be traced with tiny surveys or traditional data methods. The world, quite simply, is too complex and too rich for little data." Big Data, by the very nature of their volume, contain a depth of cases and a breadth of variables that allow researchers to tackle problems they

would ordinarily have struggled to do with conventional data gathering. Inasmuch as Big Data do not allow us to investigate certain problems, which are best left to conventional small data studies, so also Big Data open up new questions and new areas of research, simply by virtue of their volume, which small data studies cannot investigate.

Following from the issue of large samples, Big Data allow us to find enough data on people with rare conditions or in rare situations, and allow us to find enough individuals who can act as controls matched to those subjects on a wide variety of criteria. When one considers that each additional matching criterion exponentially reduces the number of available controls, it quickly becomes clear why Big Data allow matching when small data fail. It must be said, however, that researchers have devised techniques to compensate for this problem in small data studies (D'agostino, 1998; Fogarty & Small, 2016; Rassen et al., 2011; Roberts et al., 2015; Rosenbaum, 1987, 1989; Schneeweiss, 2006), and Big Data do not make these techniques obsolete, but can complement them beautifully.

Thus, while, as discussed above, many Internet companies (for better or worse) have turned their users into non-consenting (or—considering that those users probably did not read the terms and conditions as mentioned above—unconsciously consenting) research participants in the Internet equivalent of randomized trials (A/B testing), Big Data, because they so often incorporate time components, do allow for the examination of ex post facto conditions, where the size of the resulting groups in the data set are sufficiently powered to compensate for the design. This is a huge boon for academic researchers, where such data are extremely difficult to obtain otherwise.

Some Examples of the Application of Big Data

Many examples have already been provided in this chapter of the application of Big Data. However, some final examples will be provided, not to produce an encyclopaedia of Big Data sources, but rather as examples to stimulate readers to find their own.

The first example involves deep-vein thrombosis (DVT). Although clinical trials can be used to study factors like the risk for, and possible protective measures against, DVT associated with long-haul flights (Scurr et al., 2001), Big Data studies using record linkage (Kelman et al., 2003) can also provide similar insights into this risk for DVT.

The second example relates to prescription data. Many countries have started making routinely collected prescription data available for research, and the Nordic countries are definitely at the forefront of this (Furu et al., 2010; Lester, 2009; Wettermark et al., 2013). However, the Australian government has also made a sample of the de-identified prescription data of 10% of its population available to researchers, subject to a number of controls and provisions (Mellish et al., 2015; Paige et al., 2015; Pearson et al., 2015). Using this data, researchers were able to

show the massive impact on the medicine adherence of the general population brought about by a single television documentary program (Schaffer et al., 2015).

Openly Available Sources of Data

Perhaps, rather than saying that "Everything is data!" (Stephens-Davidowitz, 2017, p. 103), we should say that "Data are everywhere!" Researchers can look to many sources for data, some of which have been mentioned in this chapter already. But many, many more exist. Petabytes of data are being generated and stored, and much of that can be accessed for academic research. The United States government has committed to releasing massive amounts of data through its open data initiative (https://www.data.gov), and the European Union (http://data.europa.eu/euodp/en/data) and the UK (http://data.gov.uk) have similar initiatives. For example, the US has released 125 years of health data (https://www.healthdata.gov/content/about) for researchers and others to use in the interests of improving health outcomes for all, and the United Kingdom's NHS also provides similar data (https://digital.nhs.uk/data-and-information/data-collections-and-data-sets). Many countries make large amounts of census data available (e.g., https://www.census.gov/data.html). Data really can be found everywhere. All we need do is look, find, and ask.

In closing, two large portals for *finding* data that need to be mentioned are those provided by Google (https://www.google.com/publicdata/directory) and Amazon (https://aws.amazon.com/data sets), where researchers can search for a wide array of different existing data sources which they might profitably employ in their research work.

Conclusion

This chapter has demonstrated that the world of Big Data is not purely the preserve of industry, but that academic researchers can, and should, be using Big Data for their studies. This is not to say that researchers have not, but rather, that a critical mass of researchers extending the possibilities of what can be done with Big Data has not yet been accumulated. This may partly be ascribed to the challenges presented by working with Big Data. But these challenges are now better understood, and thus easier to account for. Additionally, the obstacles presented by these challenges are also being eroded as the dual forces of technology and methodology become more robust at working with voluminous data sets that may require some form of linking. It was also shown that one of the most important characteristics of Big Data used for academic research is precisely the fact that Big Data sets are typically not gathered for academic research, but obtained via a data custodian who has accumulated the data for other purposes. This impacts every aspect of working with those data, from accessing it, to linking and analysing it, through to the ethical

implications of working with such data. Furthermore, the opportunities which Big Data present to academic research are vast. Researchers should realise that Big Data will not replace traditional small data research methods, and that the opportunities presented by Big Data are dependent on the context of the researcher, the population and the data—in some contexts, Big Data may offer great advantage, but not in others. Ideally, Big Data and small data will become complementary tools in the armouries of academic researchers, which they may draw upon variably, and in concert if need be, as the situation demands.

Acknowledgments This project is partially funded by the National Health and Medical Research Council (NHMRC) through the Translational Australian Clinical Toxicology Program (TACT) (grant ID1055176).

References

Addawood, A., Schneider, J., & Bashier, M. (2017). Stance classification of twitter debates: The encryption debate as a use case. In *Proceedings of the 8th international conference on Social Media & Society*. Association for Computing Machinery. https://doi.org/10.1145/3097286.3097288

Anthony, S. (2011). *The history of computer storage (slideshow)*. Retrieved January 19, 2011, from https://www.extremetech.com/computing/90156-the-history-of-computer-storage-slideshow

Araz, O. M., Bentley, D., & Muelleman, R. L. (2014). Using Google Flu Trends data in forecasting influenza-like-illness related ED visits in Omaha, Nebraska. *American Journal of Emergency Medicine, 32*(9), 1016–1023. https://doi.org/10.1016/j.ajem.2014.05.052

Arora, V. S., Stuckler, D., & McKee, M. (2016). Tracking search engine queries for suicide in the United Kingdom, 2004–2013. *Public Health, 137*, 147–153. https://doi.org/10.1016/j.puhe.2015.10.015

Ayers, J. W., Althouse, B. M., Allem, J. P., Rosenquist, J. N., & Ford, D. E. (2013). Seasonality in seeking mental health information on Google. *American Journal of Preventive Medicine, 44*(5). https://doi.org/10.1016/j.amepre.2013.01.012

Ayers, J. W., Althouse, B. M., & Dredze, M. (2014). Could behavioral medicine lead the web data revolution? *Journal of the American Medical Association, 311*(14), 1399–1400. https://doi.org/10.1001/jama.2014.1505

Badger, E. (2014). *10 Years of London Underground ridership data in one map*. Retrieved January 26, 2018, from https://www.citylab.com/transportation/2014/01/10-years-london-underground-ridership-data-one-map/8156/

Bair, A. R. (2016). *From crisis to crisis: A Big Data, antenarrative analysis of how social media users make meaning during and after crisis events*. Utah State University. http://digitalcommons.usu.edu/etd/5045/

Bakos, Y., Marotta-Wurgler, F., & Trossen, D. R. (2014). Does anyone read the fine print? Consumer attention to standard form contracts. *The Journal of Legal Studies, 43*(1). https://www.journals.uchicago.edu/doi/abs/10.1086/674424

Biddle, L., Derges, J., Mars, B., Heron, J., Donovan, J. L., Potokar, J., Piper, M., Wyllie, C., & Gunnell, D. (2016). Suicide and the Internet: Changes in the accessibility of suicide-related information between 2007 and 2014. *Journal of Affective Disorders, 190*, 370–375. https://doi.org/10.1016/j.jad.2015.10.028

Biddle, L., Gunnell, D., Owen-Smith, A., Potokar, J., Longson, D., Hawton, K., Kapur, N., & Donovan, J. (2012). Information sources used by the suicidal to inform choice of method. *Journal of Affective Disorders, 136*(3), 702–709. https://doi.org/10.1016/j.jad.2011.10.004

Björk, B.-C., Roos, A., & Lauri, M. (2009). Scientific journal publishing: Yearly volume and open access availability. *Information Research, 14*(1). http://informationr.net/ir/14-1/paper391.html

Blue, V. (2014). *Facebook: Unethical, untrustworthy, and now downright harmful*. Retrieved July 1, 2014, from http://www.zdnet.com/facebook-unethical-untrustworthy-and-now-downright-harmful

Bogle, A. (2018). *Strava just published details about secret military bases, and an Australian was the first to know*. Retrieved January 29, 2018, from http://www.abc.net.au/news/science/2018-01-29/strava-heat-map-shows-military-bases-and-supply-routes/9369490

Bohensky, M. A., Jolley, D., Sundararajan, V., Evans, S., Pilcher, D. V., Scott, I., & Brand, C. A. (2010). Data linkage: A powerful research tool with potential problems. *BMC Health Services Research*, 10, 1–7. https://doi.org/10.1186/1472-6963-10-346.

Bragazzi, N. L. (2013). A Google trends-based approach for monitoring NSSI. *Psychology Research and Behavior Management*, 7, 1–8. https://doi.org/10.2147/PRBM.S44084.

Breese, E. B. (2015). When marketers and academics share a research platform: The story of crimson hexagon. *Journal of Applied Social Science, 10*(1), 3–7. https://doi.org/10.1177/1936724415569953

Brin, S., & Page, L. (1998). The anatomy of a large scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1/7), 107–117. https://doi.org/10.1.1.109.4049.

Bruckner, T. A., McClure, C., & Kim, Y. (2014). Google searches for suicide and risk of suicide. *Psychiatric Services, 65*(2), 271–272. https://doi.org/10.1176/appi.ps.201300211.

Butler, D. (2013). When Google got flu wrong. *Nature, 494*(7436), 155–156. https://doi.org/10.1038/494155a.

Carneiro, H. A., & Mylonakis, E. (2009). Google trends: A web-based tool for real-time surveillance of disease outbreaks. Clinical Infectious Diseases, 49(10), 1557–1564. https://doi.org/10.1086/630200.

Centre for Health Record Linkage. (n.d.). *Guide to health record linkage services*. Sydney. http://www.nss.gov.au/nss/home.nsf/0/e2d861c453d7b7f6ca25756700191b53/$FILE/CHeReL_Guide_version 1.3.pdf.

Chan, M. S., Morales, A., Farhadloo, M., Palmer, R. P., & Albarracín, D. (2017). Harvesting and harnessing social media data for psychological research. In H. Blanton (Ed.), *Social psychological research methods*. Taylor & Francis.

Charles, R. H. (1913). *The letter of Aristeas to Philocrates*. Retrieved January 17, 2018, from http://www.attalus.org/translate/aristeas1.html#9

Chiang, L.-Y., Crockett, R., Johnson, I., & O'Keefe, A. (2017). Passenger flow in the tube. Worcester Polytechnic Institute, . http://wp.wpi.edu/london/files/2017/06/Tube_IQP_E17_Final.pdf

Chillingworth, B. (2018). *Woman charged over death of NSW farmer allegedly searched "murder" before his death*. Retrieved January 26, 2018, from http://www.smh.com.au/nsw/woman-charged-over-death-of-nsw-farmer-allegedly-searched-murder-before-his-death-20180124-h0o1ts.html

Choi, H., & Varian, H. (2009). *Official Google research blog: Predicting the present with Google Trends*. http://googleresearch.blogspot.com/2009/04/predicting-present-with-google-trends.html

Choi, H., & Varian, H. (2012). Predicting the present with Google trends. *Economic Record, 88*(SUPPL.1), 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x.

Christie, A. (1932). *The thirteen problems*. Ulverscroft.

Churches, T., & Christen, P. (2004). Some methods for blindfolded record linkage. *BMC Medical Informatics and Decision Making*, 4, 1–17. https://doi.org/10.1186/1472-6947-4-9.

Citizen Research Centre. (2017). *Xenophobia on Social Media in SA, 2011–2017: Anatomy of an Incident: Violence in Gauteng and the "March against Immigrants."* http://citizenresearchcentre.org/2017/03/15/xenophobia-on-social-media-in-south-africa/

Clark, D. E. (2004). Practical introduction to record linkage for injury research. *Injury Prevention, 10*(3), 186–191. https://doi.org/10.1136/ip.2003.004580.

Clarke, F., & Chien, C.-H. (2017). Visualising Big Data for official statistics: The ABS experience. In T. Prodromou (Ed.), *Data visualization and statistical literacy for open and Big Data* (pp. 224–252). : IGI Global. https://doi.org/10.4018/978-1-5225-2512-7.ch009.

Cook, S., Conrad, C., Fowlkes, A. L., & Mohebbi, M. H. (2011). Assessing Google Flu Trends performance in the United States during the 2009 influenza virus a (H1N1) pandemic. PLoS One, 6(8), 1–8. https://doi.org/10.1371/journal.pone.0023610.

Copeland, P., Romano, R., Zhang, T., Hecht, G., Zigmond, D., & Stefansen, C. (2013). Google Disease Trends: An update. http://research.google.com/pubs/archive/41763.pdf

Coughlin, T. (2015). *HDD annual unit shipments increase in 2014*. Retrieved January 22, 2018, from https://www.forbes.com/sites/tomcoughlin/2015/01/29/hdd-annual-unit-shipments-increase-in-2014

Courtland, R. (2015). *Gordon Moore: The man whose name means progress: The visionary engineer reflects on 50 years of Moore's Law*. Retrieved September 15, 2016, from http://spectrum.ieee.org/computing/hardware/gordon-moore-the-man-whose-name-means-progress

Coviello, L., Sohn, Y., Kramer, A. D. I., Marlow, C., Franceschetti, M., Christakis, N. A., & Fowler, J. H. (2014). Detecting emotional contagion in massive social networks. *PLoS One, 9*(3), 1–6. https://doi.org/10.1371/journal.pone.0090315.

Cox, M., & Ellsworth, D. (1997). Application-controlled demand paging for out-of-core visualization. In *Proceedings of the 8th IEEE Visualization '97 Conference* (pp. 235–244).

Crimson Hexagon. (2018). *Data library: A trillion posts can answer a lot of questions*. Retrieved January 17, 2018, from https://www.crimsonhexagon.com/data-library/

D'agostino, R. B. (1998). Tutorial in biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*, *17*, 2265–2281. https://doi.org/10.1002/(SICI)1097-0258(19981015)17:19<2265::AID-SIM918>3.0.CO;2-B

Data Recovery Group. (2011). *Data storage history and future*. Retrieved January 19, 2018, from http://www.datarecoverygroup.com/articles/data-storage-history-and-future

Dave, P. (2013). *Big Data – what is Big Data – 3 Vs of Big Data – volume, velocity and variety – day 2 of 21*. Retrieved January 25, 2018, from https://blog.sqlauthority.com/2013/10/02/big-data-what-is-big-data-3-vs-of-big-data-volume-velocity-and-variety-day-2-of-21/

Deery, S. (2013). *Ex-girlfriend Adriana Donato's murder planned for weeks by boyfriend using Google searches*. Retrieved January 29, 2018, from http://www.heraldsun.com.au/exgirlfriend-adriana-donatos-murder-planned-for-weeks-by-boyfriend-using-google-searches/news-story/4ce0d80e6a5582c3782befbd601508cc

Deiner, M. S., Lietman, T. M., McLeod, S. D., Chodosh, J., & Porco, T. C. (2016). Surveillance tools emerging from search engines and social media data for determining eye disease patterns. *JAMA Ophthalmology, 134*(9), 1024. https://doi.org/10.1001/jamaophthalmol.2016.2267.

DeVan, A. (2016). *The 7 V's of Big Data*. Retrieved January 25, 2018, from https://www.impactradius.com/blog/7-vs-big-data/

Doctorow, C. (2008). Big Data: Welcome to the petacentre. *Nature, 455*(7209), 16–21. https://doi.org/10.1038/455016a

Dontha, R. (2017). *The origins of Big Data*. Retrieved January 17, 2018, from https://www.kdnuggets.com/2017/02/origins-big-data.html

Drucker, P. (1957). *The landmarks of tomorrow*. Heinemann.

Dugas, A. F., Hsieh, Y. H., Levin, S. R., Pines, J. M., Mareiniss, D. P., Mohareb, A., Gaydos, C. A., Perl, T. M., & Rothman, R. E. (2012). Google Flu Trends: Correlation with emergency department influenza rates and crowding metrics. *Clinical Infectious Diseases, 54*(4), 463–469. https://doi.org/10.1093/cid/cir883.

Dugas, A. F., Jalalpour, M., Gel, Y., Levin, S., Torcaso, F., Igusa, T., & Rothman, R. E. (2013). Influenza forecasting with Google Flu Trends. *PLoS One, 8*(2). https://doi.org/10.1371/journal.pone.0056176.

Dusetzina, S. B., Tyree, S., Meyer, A.-M., Meyer, A., Green, L., & Carpenter, W. R. (2014). *Linking data for health services research: A framework and instructional guide*. Rockville. https://doi.org/AHRQ No.14-EHC033.

Emery, S. L., Szczypka, G., Abril, E. P., Kim, Y., & Vera, L. (2014). Are you scared yet? Evaluating fear appeal messages in tweets about the tips campaign. *Journal of Communication, 64*(2), 278–295. https://doi.org/10.1111/jcom.12083.

Entous, A., Dwoskin, E., & Timberg, C. (2018). *Obama tried to give Zuckerberg a wake-up call over fake news on Facebook*. Retrieved June 1, 2018, from https://www.washingtonpost.com/business/economy/obama-tried-to-give-zuckerberg-a-wake-up-call-over-fake-news-on-facebook/2017/09/24/15d19b12-ddac-4ad5-ac6e-ef909e1c1284_story.html

Eysenbach, G. (2006). Infodemiology: Tracking flu-related searches on the web for syndromic surveillance. In *Proceedings of the American Medical Informatics Association Annual Symposium 2006* (pp. 244–8). https://doi.org/PMC1839505

Faris, R., Roberts, H., Etling, B., Othman, D., & Benkler, Y. (2015). Score another one for the Internet? The role of the networked public sphere in the U.S. net neutrality policy debate. *SSRN Electronic Journal*, 1, 0–34. https://doi.org/10.2139/ssrn.2563761.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association, 64*(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049.

Fish, J. N., & Russell, S. T. (2017). Have mischievous responders misidentified sexual minority youth disparities in the National Longitudinal Study of adolescent to adult health? *Archives of Sexual Behavior*, 1–15. https://doi.org/10.1007/s10508-017-0993-6.

Fogarty, C. B., & Small, D. S. (2016). Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming. *Journal of the American Statistical Association, 111*(516), 1820–1830. https://doi.org/10.1080/01621459.2015.1120675.

Fond, G., Gaman, A., Brunel, L., Haffen, E., & Llorca, P. M. (2015). Google trends®: Ready for real-time suicide prevention or just a Zeta-Jones effect? An exploratory study. *Psychiatry Research, 228*(3), 913–917. https://doi.org/10.1016/j.psychres.2015.04.022.

Fuchs, C. (2008). The role of income inequality in a multivariate cross-national analysis of the digital divide. *Social Science Computer Review, 27*(1), 41–58. https://doi.org/10.1177/0894439308321628.

Fung, K. (2014). *Google Flu Trends' failure shows Good Data > Big Data*. Retrieved January 25, 2018, from https://hbr.org/2014/03/google-flu-trends-failure-shows-good-data-big-data

Furu, K., Wettermark, B., Andersen, M., Martikainen, J. E., Almarsdottir, A. B., & Sørensen, H. T. (2010). The Nordic countries as a cohort for pharmacoepidemiological research. *Basic & Clinical Pharmacology & Toxicology, 106*(2), 86–94. https://doi.org/10.1111/j.1742-7843.2009.00494.x.

Ginsberg, J., & Mohebbi, M. H. (2008). *Tracking Flu Trends*. Retrieved March 5, 2018, from https://googleblog.blogspot.com.au/2008/11/tracking-flu-trends.html

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), 1012–1014. https://doi.org/10.1038/nature07634.

Gordon, M. (2014). *What is Strava Metro?* Retrieved March 2, 2018, from https://support.strava.com/hc/en-us/articles/216918877-What-is-Strava-Metro-

Gunn, J. F., & Lester, D. (2013). Using Google searches on the Internet to monitor suicidal behavior. Journal of Affective Disorders, 148(2–3), 411–412. https://doi.org/10.1016/j.jad.2012.11.004.

Gunnell, D., Bennewith, O., Kapur, N., Simkin, S., Cooper, J., & Hawton, K. (2012). The use of the Internet by people who die by suicide in England: A cross sectional study. Journal of Affective Disorders, 141(2–3), 480–483. https://doi.org/10.1016/j.jad.2012.04.015.

Hachman, M. (2015). *The price of free: How Apple, Facebook, Microsoft and Google sell you to advertisers*. Retrieved January 26, 2018, from https://www.pcworld.com/article/2986988/privacy/the-price-of-free-how-apple-facebook-microsoft-and-google-sell-you-to-advertisers.html

Hagihara, A., Miyazaki, S., & Abe, T. (2012). Internet suicide searches and the incidence of suicide in young people in Japan. *European Archives of Psychiatry and Clinical Neuroscience, 262*(1), 39–46. https://doi.org/10.1007/s00406-011-0212-8.

Harlow, L. L., & Oswald, F. L. (2016). Big Data in psychology: Introduction to the special issue. *Psychological Methods, 21*(4), 447–457. https://doi.org/10.1037/met0000120.

Henderson, V., Storeygard, A., & Weil, D. (2008). Measuring economic growth from outer space. *American Economic Review, 102*(2), 994–1028. https://doi.org/10.1257/aer.102.2.994.

Hern, A. (2018). *Far more than 87m Facebook users had data compromised, MPs told*. Retrieved June 4, 2018, from https://www.theguardian.com/uk-news/2018/apr/17/facebook-users-data-compromised-far-more-than-87m-mps-told-cambridge-analytica

Hilbert, M. (2012a). How much information is there in the "information society"? *Significance, 9*(4), 8–12. https://doi.org/10.1111/j.1740-9713.2012.00584.x.

Hilbert, M. (2012b). How to measure "how much information"? Theoretical, methodological, and statistical challenges for the social sciences. *International Journal of Communication, 6*(1), 1042–1055. http://ijoc.org/index.php/ijoc/article/view/1318/746

Hilbert, M. (2014). What is the content of the world's technologically mediated information and communication capacity: How much text, image, audio, and video? *Information Society, 30*(2), 127–143. https://doi.org/10.1080/01972243.2013.873748.

Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science*, 332(April), 60–65. https://doi.org/10.1126/science.1200970.

Hilbert, M., & López, P. (2012a). How to measure the world's technological capacity to communicate, store, and compute information, Part I: Results and scope. *International Journal of Communication, 6*(1), 956–979. http://ijoc.org/index.php/ijoc/article/view/1562/742

Hilbert, M., & López, P. (2012b). How to measure the world's technological capacity to communicate, store, and compute information, Part II: Measurement unit and conclusions. *International Journal of Communication, 6*, 936–955. http://ijoc.org/index.php/ijoc/article/view/1563/741

Hitt, C. (2018). Woman accused of murdering husband searched for "how to kill someone and not get caught" online. Retrieved January 26, 2018, from http://www.nydailynews.com/news/crime/woman-web-searched-pin-murder-article-1.3753079

Holland, T. (2017). *How Facebook and Google changed the advertising game*. Retrieved January 26, 2018, from https://theconversation.com/how-facebook-and-google-changed-the-advertising-game-70050

Hopke, J. E., & Simis, M. (2017a). Discourse over a contested technology on Twitter: A case study of hydraulic fracturing. *Public Understanding of Science, 26*(1), 105–120. https://doi.org/10.1177/0963662515607725.

Hopke, J. E., & Simis, M. (2017b). Response to "word choice as political speech": Hydraulic fracturing is a partisan issue. *Public Understanding of Science, 26*(1), 124–126. https://doi.org/10.1177/0963662516643621.

Hopkins, D., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science, 54*(1), 229–247. https://doi.org/10.1111/j.1540-5907.2009.00428.x.

Human Rights Council of the United Nations. (2016). *The promotion, protection and enjoyment of human rights on the Internet*, Pub. L. No. A/HRC/32/L.20 (2016). https://www.article19.org/data/files/Internet_Statement_Adopted.pdf.

IBM. (n.d.-a). *Extracting business value from the 4 V's of big data*. Retrieved January 25, 2018, from http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data.

IBM. (n.d.-b). *Infographic: The four V's of Big Data*. Retrieved January 25, 2018, from http://www.ibmbigdatahub.com/infographic/four-vs-big-data.

Ingram, D. (2018). *Facebook fuels broad privacy debate by tracking non-users*. Retrieved April 16, 2018, from https://www.reuters.com/article/us-facebook-privacy-tracking/facebook-fuels-broad-privacy-debate-by-tracking-non-users-idUSKBN1HM0DR

Jacobs, A. (2009). The pathologies of Big Data. *Queue, 7*(6), 36–44. https://doi.org/10.1145/1563821.1563874.

Jain, A. (2016). *The 5 Vs of Big Data – Watson health perspectives*. Retrieved January 25, 2018, from https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/%0A

Jinha, A. E. (2010). Article 50 million: An estimate of the number of scholarly articles in existence. *Learned Publishing, 23*(3), 258–263. https://doi.org/10.1087/20100308.

Johnson, H. A., Wagner, M. M., Hogan, W. R., Chapman, W., Olszewski, R. T., Dowling, J., & Barnas, G. (2004). Analysis of web access logs for surveillance of influenza. *Studies in Health Technology and Informations, 107*(2), 1202–1206. https://doi.org/10.3233/978-1-60750-949-3-1202.

Kamenetz, A. (2014). *"Mischievous responders" confound research on teens*. Retrieved August 12, 2017, from https://www.npr.org/sections/ed/2014/05/22/313166161/mischievous-responders-confound-research-on-teens

Kelman, C. W., Kortt, M. A., Becker, N. G., Li, Z., Mathews, J. D., Guest, C. S., & Holman, C. D. J. (2003). Deep vein thrombosis and air travel: Record linkage study. *BMJ (Clinical Research Ed.), 327*(7423), 1072. https://doi.org/10.1136/bmj.327.7423.1072.

Kim, J., Brossard, D., Scheufele, D. A., & Xenos, M. (2016). "Shared" information in the age of Big Data. *Journalism & Mass Communication Quarterly, 93*(2), 430–445. https://doi.org/10.1177/1077699016640715.

King, G. (2011). Ensuring the data-rich future of the social sciences. *Science, 331*(6018), 719–721. https://doi.org/10.1126/science.1197872.

King, G. (2014). Restructuring the social sciences: Reflections from Harvard's Institute for Quantitative Social Science. *PS: Political Science & Politics, 47*(01), 165–172. https://doi.org/10.1017/S1049096513001534.

King, G., & Persily, N. (2018). *A new model for industry-academic partnerships*. http://j.mp/2q1IQpH

Kramer, A. D. I. (2014). *Facebook post by A Kramer*. Retrieved July 1, 2014, from https://www.facebook.com/akramer/posts/10152987150867796

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences USA, 111*(24), 8788–8790. https://doi.org/10.1073/pnas.1412469111.

Kristoufek, L., Moat, H. S., & Preis, T. (2016). Estimating suicide occurrence statistics using Google trends. *EPJ Data Science, 5*(1), 32. https://doi.org/10.1140/epjds/s13688-016-0094-0.

La Rue, F. (2011). *Report of the special rapporteur on the promotion and protection of the right to freedom of opinion and expression*. Frank La Rue. http://www2.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf

Laney, D. (2001). 3D data management: Controlling data volume, velocity, and variety. *Application Delivery Strategies, 949*(February 2001), 4. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Larsen, P. O., & von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics, 84*(3), 575–603. https://doi.org/10.1007/s11192-010-0202-z.

Lazer, D., & Kennedy, R. (2015). *What we can learn from the epic failure of Google Flu Trends*. Retrieved January 25, 2018, from https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014a). The parable of Google Flu: Traps in Big Data analysis. *Science, 343*(6176), 1203–1205. https://doi.org/10.1126/science.1248506.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2014b). Computational social science. *Science*, 323, 721–723. https://doi.org/10.1126/science.1169410.

Leathern, R. (2018). *Shining a light on ads with political content*. Retrieved June 2, 2018, from https://newsroom.fb.com/news/2018/05/ads-with-political-content/

Lesk, M. (1997). *How much information is there in the world?* Retrieved December 8, 2017, from http://www.lesk.com/mlesk/ksg97/ksg.html

Lester, D. (2009). The Nordic prescription databases as a resource for pharmacoepidemiological research—A literature review. *Clinical Neuropsychiatry, 6*(5), 188–191. https://doi.org/10.1002/pds.

Lewis, P. (2018). *"Utterly horrifying": Ex-Facebook insider says covert data harvesting was routine*. Retrieved March 21, 2018, from https://www.theguardian.com/news/2018/mar/20/facebook-data-cambridge-analytica-sandy-parakilas

Lewis, P., & Carrie Wong, J. (2018). *Facebook employs psychologist whose firm sold data to Cambridge Analytica*. Retrieved June 4, 2018, from https://www.theguardian.com/news/2018/mar/18/facebook-cambridge-analytica-joseph-chancellor-gsr

Li, N., Akin, H., Yi-Fan, L. S., Brossard, D., Xenos, M., & Scheufele, D. A. (2016). Tweeting disaster: An analysis of online discourse about nuclear power in the wake of the Fukushima Daiichi nuclear accident. *Journal of Science Communication, 15*(5), 1–20. https://jcom.sissa.it/archive/15/05/JCOM_1505_2016_A02

Library of Congress. (2017). *General information*. Retrieved January 17, 2018, from https://www.loc.gov/about/general-information/

Lohr, S. (2013, February 1). The origins of "Big Data": An etymological detective story. *The New York Times*. https://bits.blogs.nytimes.com/2013/02/01/the-origins-of-big-data-an-etymological-detective-story/

London's Tube and Big Data: Underground movement. (2013). Retrieved January 26, 2018, from https://www.newscientist.com/article/in415-londons-tube-and-big-data-underground-movement/

Ma-Kellams, C., Or, F., Baek, J. H., & Kawachi, I. (2016). Rethinking suicide surveillance: Google search data and self-reported suicidality differentially estimate completed suicide risk. *Clinical Psychological Science, 4*(3), 480–484. https://doi.org/10.1177/2167702615593475.

Mabe, M., & Amin, M. (2001). Growth dynamics of scholarly and scientific journals. *Scientometrics, 51*(1), 147–162. https://link.springer.com/article/10.1023/A:1010520913124

Madrigal, A. C. (2014). *In defense of Google Flu Trends*. Retrieved January 25, 2018, from https://www.theatlantic.com/technology/archive/2014/03/in-defense-of-google-flu-trends/359688/

Malik, M. T., Gumel, A., Thompson, L. H., Strome, T., & Mahmud, S. M. (2011). "Google Flu Trends" and emergency department triage data predicted the 2009 pandemic H1N1 waves in Manitoba. *Canadian Journal of Public Health, 102*(4), 294–297. http://www.jstor.org/stable/41995614

Marr, B. (2014). *Big Data – The 5 Vs everyone must know*. Retrieved January 25, 2018, from https://www.slideshare.net/BernardMarr/140228-big-data-volume-velocity-variety-varacity-value%0A

Mars, B., Heron, J., Biddle, L., Donovan, J. L., Holley, R., Piper, M., Potokar, J., Wyllie, C., & Gunnell, D. (2015). Exposure to, and searching for, information about suicide and self-harm on the Internet: Prevalence and predictors in a population based cohort of young adults. *Journal of Affective Disorders*, 185, 239–245. https://doi.org/10.1016/j.jad.2015.06.001.

Mashey, J. R. (1998). *Big Data and the next wave of infraStress*. University of California, Berkeley. http://static.usenix.org/event/usenix99/invited_talks/mashey.pdf

Matsa, K. E., Mitchell, A., & Stocking, G. (2017). *Methodology*. Retrieved February 26, 2018, from http://www.journalism.org/2017/04/27/google-flint-methodology/

McCarthy, M. J. (2010). Internet monitoring of suicide risk in the population. *Journal of Affective Disorders, 122*(3), 277–279. https://doi.org/10.1016/j.jad.2009.08.015.

McNulty, E. (2014). *Understanding Big Data: The seven V's*. Retrieved January 25, 2018, from http://dataconomy.com/2014/05/seven-vs-big-data/

Mellish, L., Karanges, E. A., Litchfield, M. J., Schaffer, A. L., Blanch, B., Daniels, B. J., Segrave, A., & Pearson, S.-A. (2015). The Australian pharmaceutical benefits scheme data collection: A practical guide for researchers. *BMC Research Notes, 8*(1), 634. https://doi.org/10.1186/s13104-015-1616-8.

Mellon, J. (2013). Where and when can we use Google trends to measure issue salience? *PS: Political Science & Politics, 46*(02), 280–290. https://doi.org/10.1017/S1049096513000279.

Moe, W. W., & Schweidel, D. A. (2017). Opportunities for innovation in social media analytics. *Journal of Product Innovation Management, 34*(5), 697–702. https://doi.org/10.1111/jpim.12405.

Monnappa, A. (2017). *How Facebook is using Big Data – The good, the bad, and the ugly*. Retrieved January 26, 2018, from https://www.simplilearn.com/how-facebook-is-using-big-data-article

Moore, G. E. (1965). Cramming more components onto integrated circuits. *Electronics, 38*(8), 114–117. https://doi.org/10.1109/N-SSC.2006.4785860.

Moriarity, M. (2017). *15 Things you can learn from 1 trillion Posts: How 1,000,000,000,000 posts can change the world we live in*. Retrieved January 17, 2018, from https://www.crimsonhexagon.com/blog/15-things-you-can-learn-from-1-trillion-posts/

Mullin, J. (2012). *How much do Google and Facebook profit from your data?* Retrieved January 26, 2018, from https://arstechnica.com/tech-policy/2012/10/how-much-do-google-and-facebook-profit-from-your-data/

Nanji, A. (2017). *The most popular social networks with Millennials, Gen X and Baby Boomers*. Retrieved March 2, 2018, from https://www.marketingprofs.com/charts/2017/31792/the-most-popular-social-networks-with-millennials-gen-x-and-baby-boomers

National Highway Traffic Safety Administration. (2016). *Fatality Analysis Reporting System (FARS): Analytical User's Manual 1975–2015* (No. DOT HS 812 315). ftp://ftp.nhtsa.dot.gov/FARS/FARS-DOC/Analytical User Guide/USERGUIDE-2015.pdf.

Obar, J. A., & Oelof-Hirsch, A. (2016). The biggest lie on the Internet: Ignoring the privacy policies and terms of service policies of social networking services. In *The 44th Research Conference on Communication, Information and Internet Policy 2016*. https://doi.org/10.2139/ssrn.2757465

Olson, D. R., Konty, K. J., Paladini, M., Viboud, C., & Simonsen, L. (2013). Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: A comparative epidemiological study at three geographic scales. *PLoS Computational Biology, 9*(10). https://doi.org/10.1371/journal.pcbi.1003256.

Ong, T. C., Mannino, M. V., Schilling, L. M., & Kahn, M. G. (2014). Improving record linkage performance in the presence of missing linkage data. *Journal of Biomedical Informatics*, 52, 43–54. https://doi.org/10.1016/j.jbi.2014.01.016.

Ortiz, J. R., Zhou, H., Shay, D. K., Neuzil, K. M., Fowlkes, A. L., & Goss, C. H. (2011). Monitoring influenza activity in the United States: A comparison of traditional surveillance systems with Google Flu Trends. *PLoS One, 6*(4), 2–10. https://doi.org/10.1371/journal.pone.0018687.

Ortutay, B. (2018). *Facebook scandal affected more users than thought: Up to 87M*. Retrieved April 5, 2018, from https://www.apnews.com/e0e0df2083fe40c0b0ad10ff1946f041

Ortutay, B., Kirka, D., & Katz, G. (2018). *Facebook's Zuckerberg apologizes for 'major breach of trust.'* Retrieved March 22, 2018, from https://apnews.com/c8f615be9523421998b4fcc16374ff37

Page, A., Chang, S.-S., & Gunnell, D. (2011). Surveillance of Australian suicidal behaviour using the Internet? *Australian and New Zealand Journal of Psychiatry, 45*(12), 1020–1022. https://doi.org/10.3109/00048674.2011.623660.

Paige, E., Kemp-Casey, A., Korda, R., & Banks, E. (2015). Using Australian Pharmaceutical Benefits Scheme data for pharmacoepidemiological research: Challenges and approaches. *Public Health Research & Practice, 25*(4), 1–6. https://doi.org/10.17061/phrp2541546.

Parker, J., Cuthbertson, C., Loveridge, S., Skidmore, M., & Dyar, W. (2017). Forecasting state-level premature deaths from alcohol, drugs, and suicides using Google trends data. *Journal of Affective Disorders, 213*(November 2016), 9–15. https://doi.org/10.1016/j.jad.2016.10.038

Pearson, S.-A., Pesa, N., Langton, J. M., Drew, A., Faedo, M., & Robertson, J. (2015). Studies using Australia's Pharmaceutical Benefits Scheme data for pharmacoepidemiological research: A systematic review of the published literature (1987–2013). *Pharmacoepidemiology and Drug Safety*, 24, 447–455. https://doi.org/10.1002/pds.3756.

Pingdom. (2008). *The history of computer data storage, in pictures*. Retrieved January 19, 2018, from http://royal.pingdom.com/2008/04/08/the-history-of-computer-data-storage-in-pictures/

Plume, A., & van Weijen, D. (2014). *Publish or perish? The rise of the fractional author…. Research Trends*, *38*. https://www.researchtrends.com/issue-38-september-2014/publish-or-perish-the-rise-of-the-fractional-author/

Polgreen, P. M., Chen, Y., Pennock, D. M., & Nelson, F. D. (2008). Using Internet searches for influenza surveillance. *Clinical Infectious Diseases, 47*(11), 1443–1448. https://doi.org/10.1086/593098.

Press, G. (2013). *A very short history of Big Data*. Retrieved December 8, 2017, from https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#1eaca84c65a1

Prodromou, T., & Dunne, T. (2017). Data visualisation and statistics education in the future. In T. Prodromou (Ed.), *Data visualization and statistical literacy for open and Big Data* (pp. 1–28). : IGI Global. https://doi.org/10.4018/978-1-5225-2512-7.ch001.

Przybylski, A. K. (2016). Mischievous responding in internet gaming disorder research. *PeerJ*, 4, e2401. https://doi.org/10.7717/peerj.2401.

Puang-ngern, B., Bilgin, A. A., & Kyng, T. J. (2017). Comparison of graduates' and academics' perceptions of the skills required for Big Data analysis: Statistics education in the age of Big Data. In T. Prodromou (Ed.), *Data visualization and statistical literacy for open and Big Data* (pp. 126–152). : IGI Global. https://doi.org/10.4018/978-1-5225-2512-7.ch006.

Rassen, J. A., Glynn, R. J., Brookhart, M. A., & Schneeweiss, S. (2011). Covariate selection in high-dimensional propensity score analyses of treatment effects in small samples. *American Journal of Epidemiology, 173*(12), 1404–1413. https://doi.org/10.1093/aje/kwr001.

Raubenheimer, J. E. (2019). Google Trends Extraction Tool. https://doi.org/10.5281/zenodo.2620618

Raubenheimer, J. E. (2021). Google Trends Extraction Tool for Google Trends Extended for Health data. Software Impacts, 8, 100060. https://doi.org/10.1016/j.simpa.2021.100060

Reuters. (2018). Cambridge Analytica and British parent shut down after Facebook scandal. Retrieved May 3, 2018, from https://www.reuters.com/article/us-facebook-privacy/cambridge-analytica-and-british-parent-shut-down-after-facebook-scandal-idUSKBN1I32L7

Robb, D. (2017). *The Global Heatmap, now 6x hotter*. Retrieved January 1, 2018, from https://medium.com/strava-engineering/the-global-heatmap-now-6x-hotter-23fc01d301de

Roberts, M. E., Stewart, B. M., & Nielsen, R. (2015). *Matching methods for high-dimensional data with applications to text*. http://www.margaretroberts.net/wp-content/uploads/2015/07/textmatching.pdf

Robinson-Cimpian, J. P. (2014). Inaccurate estimation of disparities due to mischievous responders: Several suggestions to assess conclusions. *Educational Researcher, 43*(4), 171–185. https://doi.org/10.3102/0013189X14534297.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika, 74*(1), 13–26. http://www.jstor.org/stable/2336017

Rosenbaum, P. R. (1989). Sensitivity analysis for matched observational studies with many ordered treatments. *Scandinavian Journal of Statistics, 16*(3), 227–236. http://www.jstor.org/stable/4616136

Runge, K. K., Yeo, S. K., Cacciatore, M., Scheufele, D. A., Brossard, D., Xenos, M., Anderson, A., Choi, D. H., Kim, J., Li, N., Liang, X., Stubbings, M., & Su, L. Y. F. (2013). Tweeting nano: How public discourses about nanotechnology develop in social media environments. *Journal of Nanoparticle Research, 15*(1). https://doi.org/10.1007/s11051-012-1381-8.

Salsburg, D. S. (2017). *Errors, blunders, and lies: How to tell the difference*. CRC Press.

Salzberg, S. (2014). *Why Google Flu is a failure*. Retrieved January 25, 2018, from https://www.forbes.com/sites/stevensalzberg/2014/03/23/why-google-flu-is-a-failure/#42fed4945535

Sari Aslama, N., Cheshire, B. J., & Cheng, T. (2015). *Big Data analysis of population flow between TfL oyster and bicycle hire networks in London*. University College London. http://leeds.gis-ruk.org/abstracts/GISRUK2015_submission_92.pdf

Schaffer, A. L., Buckley, N. A., Dobbins, T. A., Banks, E., & Pearson, S.-A. (2015). The crux of the matter: Did the ABC's catalyst program change statin use in Australia? *Medical Journal of Australia, 11*(11), 591–595. https://doi.org/10.5694/mja15.0010.

Schneeweiss, S. (2006). Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiology and Drug Safety*, 15, 291–303. https://doi.org/10.1002/pds.1200.

Schrage, E., & Ginsberg, D. (2018). *Facebook launches new initiative to help scholars assess social media's impact on elections*. Retrieved June 2, 2018, from https://newsroom.fb.com/news/2018/04/new-elections-initiative/

Scurr, J. H., Machin, S. J., Bailey-King, S., Mackie, I. J., McDonald, S., & Coleridge Smith, P. D. (2001). Frequency and prevention of symptomless deep vein thrombosis in long-haul flights: A randomised trial. *Lancet, 357*, 1485–1489. https://www.thelancet.com/journals/lancet/article/PIIS0140673600046456/abstract

Smith, G. C. S., & Pell, J. P. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomised controlled trials. *BMJ (Clinical Research Ed.), 327*(7429), 1459–1461. https://doi.org/10.1177/154510970400300401.

Solano, P., Ustulin, M., Pizzorno, E., Vichi, M., Pompili, M., Serafini, G., & Amore, M. (2016). A Google-based approach for monitoring suicide risk. *Psychiatry Research*, 246, 581–586. https://doi.org/10.1016/J.PSYCHRES.2016.10.030.

Song, T. M., Song, J., An, J. Y., Hayman, L. L., & Woo, J. M. (2014). Psychological and social factors affecting Internet searches on suicide in Korea: A Big Data analysis of Google search trends. *Yonsei Medical Journal, 55*(1), 254–263. https://doi.org/10.3349/ymj.2014.55.1.254.

Spielberg, S. (2002). *Minority Report*. USA: Twentieth Century Fox. http://www.imdb.com/title/tt0181689

Stephens-Davidowitz, S. (2017). *Everybody lies*. HarperCollins.

Sueki, H. (2011). Does the volume of Internet searches using suicide-related search terms influence the suicide death rate: Data from 2004 to 2009 in Japan. *Psychiatry and Clinical Neurosciences, 65*(4), 392–394. https://doi.org/10.1111/j.1440-1819.2011.02216.x.

Taleb, N. (2013). *Beware the big errors of "Big Data."* Retrieved December 8, 2017, from https://www.wired.com/2013/02/big-data-means-big-errors-people/

The Flu Trends Team. (2015). *The next chapter for Flu Trends*. Retrieved January 25, 2018, from https://research.googleblog.com/2015/08/the-next-chapter-for-flu-trends.html

The Statistics Portal. (2018). *Global shipments of hard disk drives (HDD) from 4th quarter 2010 to 3rd quarter 2017 (in millions)*. Retrieved January 22, 2018, from https://www.statista.com/statistics/275336/global-shipment-figures-for-hard-disk-drives-from-4th-quarter-2010/

Thomas, R., & McSharry, P. (2015). *Big Data revolution: What farmers, doctors and insurance agents teach us about discovering Big Data patterns*. John Wiley & Sons.

Tran, U. S., Andel, R., Niederkrotenthaler, T., Till, B., Ajdacic-Gross, V., & Voracek, M. (2017). Low validity of Google trends for behavioral forecasting of national suicide rates. *PLoS One, 12*(8), 1–26. https://doi.org/10.1371/journal.pone.0183149.

Tromp, M., Ravelli, A. C., Bonsel, G. J., Hasman, A., & Reitsma, J. B. (2011). Results from simulated data sets: Probabilistic record linkage outperforms deterministic record linkage. *Journal of Clinical Epidemiology, 64*(5), 565–572. https://doi.org/10.1016/j.jclinepi.2010.05.008.

Turriago-Hoyos, A., Thoene, U., & Arjoon, S. (2016). Knowledge workers and virtues in Peter Drucker's management theory. *SAGE Open, 6*(1). https://doi.org/10.1177/2158244016639631.

Ueda, M., Mori, K., Matsubayashi, T., & Sawada, Y. (2017). Tweeting celebrity suicides: Users' reaction to prominent suicide deaths on Twitter and subsequent increases in actual suicides. *Social Science and Medicine*, 189, 158–166. https://doi.org/10.1016/j.socscimed.2017.06.032.

Ugander, J., Backstrom, L., Marlow, C., & Kleinberg, J. (2012). Structural diversity in social contagion. *Proceedings of the National Academy of Sciences USA, 109*(16), 5962–5966. https://doi.org/10.1073/pnas.1116502109.

UN Global Pulse. (2014). *Mining Indonesian tweets to understand food price crises*. Jakarta. https://www.unglobalpulse.org/projects/social-media-social-protection-indonesia

Valdivia, A., Lopez-Alcalde, J., Vicente, M., Pichiule, M., Ruiz, M., & Ordobas, M. (2010). Monitoring influenza activity in Europe with Google Flu Trends: Comparison with the findings

of sentinel physician networks—Results for 2009-10. *Euro Surveillance, 15*(29), 1–6. http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19621

Vaughan-Nichols, S. (2014). *We're all just lab rats in Facebook's laboratory*. Retrieved January 25, 2018, from http://www.zdnet.com/article/were-all-just-lab-rats-in-facebooks-laboratory/

Verma, I. M. (2014). Editorial expression of concern and correction. *Proceedings of the National Academy of Sciences USA*, *111*(29), 10779. www.pnas.org/cgi/doi/10.1073/pnas.1412469111.

Vespignani, A. (2009). Predicting the behavior of techno-social systems. *Science, 325*(5939), 425–428. https://doi.org/10.1126/science.1171990.

Walsh, B. (2014). *Google's Flu project shows the failings of Big Data*. Retrieved January 25, 2018, from http://time.com/23782/google-flu-trends-big-data-problems/

Ware, M., & Mabe, M. (2009). The STM report: An overview of scientific and scholarly journal publishing. http://www.markwareconsulting.com/institutional-repositories/the-stm-report-an-overview-of-scientific-and-scholarly-journal-publishing/

Ware, M., & Mabe, M. (2012). The STM report: An overview of scientific and scholarly journal publishing (3rd ed). http://www.stm-assoc.org/2012_12_11_STM_Report_2012.pdf

Ware, M., & Mabe, M. (2015). The STM report: An overview of scientific and scholarly journal publishing (4th ed). http://www.stm-assoc.org/2015_02_20_STM_Report_2015.pdf

Wartzman, R. (2014). *What Peter Drucker knew about 2020*. Retrieved January 23, 2018, from https://hbr.org/2014/10/what-peter-drucker-knew-about-2020

Wettermark, B., Zoëga, H., Furu, K., Korhonen, M., Hallas, J., Nørgaard, M., Almarsdottir, A. B., Andersen, M., Andersson Sundell, K., Bergman, U., Helin-Salmivaara, A., Hoffmann, M., Kieler, H., Martikainen, J. E., Mortensen, M., Petzold, M., Wallach-Kildemoes, H., Wallin, C., & Sørensen, H. (2013). The Nordic prescription databases as a resource for pharmacoepidemiological research—A literature review. *Pharmacoepidemiology and Drug Safety, 22*(7), 691–699. https://doi.org/10.1002/pds.3457.

Why the 3V's are not sufficient to describe Big Data. (2015). Retrieved January 25, 2018, from https://datafloq.com/read/3vs-sufficient-describe-big-data/166

Wikipedia. (2018). *Wikipedia: Database download*. Retrieved January 31, 2018, from wikipedia.org/wiki/Wikipedia:Database_download.

Wilson, N., Mason, K., Tobias, M., Peacey, M., Huang, Q. S., & Baker, M. (2009). Interpreting "Google Flu Trends" data for pandemic H1N1 Influenza: The New Zealand experience. *Euro Surveillance, 14*(44), 1–3. http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=19386

Winkler, W. E. (1993). *Matching and record linkage*. *US Census Bureau – Research Reports*. https://www.census.gov/srd/papers/pdf/rr93-8.pdf

Winkler, W. E. (2014). Matching and record linkage. Wiley Interdisciplinary Reviews: Computational Statistics, 6(5), 313–325. https://doi.org/10.1002/wics.1317.

Yang, A. C., Tsai, S. J., Huang, N. E., & Peng, C. K. (2011). Association of Internet search trends with suicide death in Taipei City, Taiwan, 2004-2009. *Journal of Affective Disorders, 132*(1–2), 179–184. https://doi.org/10.1016/j.jad.2011.01.019.

Youtie, J., Porter, A. L., & Huang, Y. (2017). Early social science research about Big Data. *Science and Public Policy, 44*(1), scw021. https://doi.org/10.1093/scipol/scw021.

Jacques Raubenheimer completed a PhD in Research Psychology at the University of the Free State, South Africa, then worked as a research consultant, before beginning a career as a lecturer in biostatistics in 2008. He moved from South Africa to Australia at the start of 2017 to work as a research fellow with the NHMRC-funded Translational Australian Clinical Toxicology programme. His work focuses on analysing one of the largest cohorts of poisoning admissions data in the world (with data from 2001 to the present), and working with national pharmaceutical dispensing datasets from the Australian population. His current aim is to establish methods for using internet search- and social media data for the monitoring and nowcasting of trends in drug interest, particularly novel substances and substances slipping under the radar of conventional regulation, such as Human Enhancement Drugs (HEDs).