4

BIG DATA

Francis Diebold (2012) traces the etymology of the term 'big data' to the mid-1990s, first used by John Mashey, retired former Chief Scientist at Silicon Graphics, to refer to handling and analysis of massive datasets. Initially, the term had little traction. In 2008 very few people were using the term 'big data', either in the academy or industry. Five years later it had become a buzzword, commonly used in business circles and the popular media, with regular commentaries in broadsheet newspapers, such as the *New York Times* and *Financial Times*, and feature pieces and sections in popular and science magazines such as *The Economist*, *Time*, *Nature*, and *Science*. Such was its prevalence and associated boosterism that Gartner's had already declared by January 2013 that it had moved along the hype cycle from 'peak of inflated expectation' to 'trough of disillusionment' (Sicular 2013), with some evangelists already declaring 'big data' dead as a meaningful term, having become too wide-ranging and woolly in definition (e.g., de Goes 2013), some early adopters struggling to convert investment into return, and others voicing scepticism as to its potential benefits. Nonetheless, business, government and research funders have largely remained firm in their conviction that big data is set to rise back up the hype cycle's 'slope of enlightenment' to the 'plateau of productivity', and, what's more, it is set to alter fundamentally how science and business are conducted (Sicular 2013; see Chapters 7 and 8).

As discussed in Chapter 2, across government, industry and academia there have long been very large datasets from which information has been extracted in order to provide insights and knowledge. Governmental departments and agencies routinely generate huge quantities of data. For example, in 2013 the National Archives and Records Administration (NARA) in the US was storing some 4.5 million cubic feet of physical documents from US executive branch agencies, courts, Congress and presidents (just 5 per cent of the federal government's records) to which it adds 30,000 linear feet of new records annually (Ellis 2013), as well as holding more than 500 terabytes of digital data. Likewise, businesses have collated data about their operations, markets and customers, and vast databases of scientific data have been assembled and curated since the start of modern science. So, what is meant by the term 'big data', given these data volumes in previous eras?

Like many terms used to refer to the rapidly evolving use of technologies and practices, there is no agreed academic or industry definition of big data. The most common makes reference to the 3Vs: volume, velocity and variety (Laney 2001; Zikopoulos et al. 2012). Big data are:

- huge in *volume*, consisting of terabytes or petabytes of data;
- high in *velocity*, being created in or near real-time;
- diverse in *variety* in type, being structured and unstructured in nature, and often temporally and spatially referenced.

Prior to big data, databases were constrained across these three attributes: it was only possible to have two at any one time (large and fast; varied and fast; large and varied) (Croll 2012). With enhanced computational power, new database design and distributed storage (see Chapter 5), all three have become simultaneously achievable enabling new forms of analysis and providing very detailed views of large systems in flux. Beyond the 3Vs, the emerging literature denotes a number of other key characteristics, with big data being:

- *exhaustive* in scope, striving to capture entire populations or systems (n = all), or at least much larger sample sizes than would be employed in traditional, small data studies;
- fine-grained in *resolution*, aiming to be as detailed as possible, and uniquely *indexical* in identification;
- *relational* in nature, containing common fields that enable the conjoining of different datasets;
- *flexible*, holding the traits of extensionality (can add new fields easily) and *scalable* (can expand in size rapidly)

(boyd and Crawford 2012; Dodge and Kitchin 2005; Marz and Warren 2012; Mayer-Schonberger and Cukier 2013).

Given the drive to digitise and scale traditional small data into digital infrastructures that are voluminous and varied (such as national archives, censuses and collections of cultural and social heritage; see Chapter 2) it is velocity and these additional characteristics that set big data apart and make them a disruptive innovation (Christensen's 1997) one that radically changes the nature of data and what can be done with them (see Table 2.1). For example, a national household survey has large volume, strong resolution and relationality, but lacks velocity (once a year), variety (usually *c*.30 structured questions), exhaustivity (a sample of perhaps one in twenty households), and flexibility (the fields are fixed, typically across surveys, to enable time-seried analysis). In this chapter, the seven characteristics of big data are elaborated and the next chapter discusses the enablers and sources of big data.

VOLUME

The last decade has witnessed an explosion in the amount of data that are being generated and processed on a daily basis. As *Wired* magazine put it in the title of their 2008 special issue: we are entering 'The Petabyte Age' (in fact, we have already entered the zettabyte age; 2^{70} bytes). Several studies have sought to estimate and track the volumes involved (e.g., Hilbert and López 2009; Gantz and Reinsel 2011; Short et al. 2011). They employ different methodologies and definitions, but all are unanimous that the rate of growth has been staggering in scale. Moreover, it is set to grow exponentially for the foreseeable future. The simplest way to illustrate this growth is to give some examples of the global estimates of data volumes and some estimates relating to specific entities. To provide a frame of reference, Table 4.1 details a summary of how data volume is measured.

Table 4.1 Measurements of digital data

Unit	Size	What it means
Bit (b)	1 or 0	Short for 'binary digit', after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code
Kilobyte (KB)	1,000, or 2^{10} bytes	From 'thousand' in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From 'large' in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From 'giant' in Greek. A two-hour film can be compressed into 1–2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From 'monster' in Greek. All of the catalogued books in America's Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All the letters delivered by America's postal service in 2010 amounted to around 5PB of data
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence in 2010 was forecast to be around 1.2ZB

(Continued)

Table 4.1 (Continued)

Unit	Size	What it means
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine
	The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.	

Source: *The Economist* (2010).

Reproduced with the permission of *The Economist*. © The Economist Newspaper Limited, London, issued 11 March 2014.

Zikopoulos et al. (2012) detail that in 2000, *c*.800,000 petabytes of data were stored in the world. According to Short et al. (2011: 7), in their annual report – *How Much Information?* – by '2008, the world's servers processed 9.57 zettabytes of information ... This was 12 gigabytes of information daily for the average worker, or about 3 terabytes of information per worker per year. The world's companies on average processed 63 terabytes of information annually' excluding non-computer sources. By 2010, MGI (cited in Manyika et al. 2011: 3) 'estimated that enterprises globally stored more than 7 exabytes of new data on disk drives ... while consumers stored more than 6 exabytes of new data on devices such as PCs and notebooks'. They further estimated that in '2009, nearly all sectors in the US economy had at least an average of 200 terabytes of stored data ... per company with more than 1,000 employees. Many sectors had more than 1 petabyte in mean stored data per company.' In 2013, EU commissioner for Digital Agenda, Neelie Kroes, stated that 1.7 million billion bytes of data per minute were being generated globally (Rial 2013).

Based on their review of data volume growth, Manyika et al. (2011) projected a 40 per cent rise in data generated globally per year. Gantz and Reinsel (2011) estimated that the 'amount of information created and replicated on the Internet will surpass 1.8 zettabytes (1.8 trillion gigabytes)' in 2011 stored in '500 quadrillion files'. This they reported represented a growth by 'a factor of 9 in just five years', with growth at that time projected to 'more than doubl[e] every two years'. As a result, they predicted that in the decade following their report,

> the number of servers (virtual and physical) worldwide will grow by a factor of 10, the amount of information managed by enterprise datacenters will grow by a factor of 50, and the number of files the datacenter will have to deal with will grow by a factor of 75, at least.

Such is the phenomenal growth in data production, IBM (2012) contended that '90% of the data in the world today has been created in the last two years alone' and Zikopoulos et al. (2012) expect data volumes to reach 35 zettabytes by 2020.

We can get a sense of why this digital data explosion is occurring by considering some specific examples. TechAmerica estimates that each day 114 billion e-mails and 24 billion text messages are sent, and 12 billion phone calls made globally (Strohm and Homan 2013). According to CISCO, in 2013 there were estimated to be 10 billion objects (devices and sensors) making up the Internet of things, each of which is producing data in varying quantities, with this figure set to rise to 50 billion by 2020 (Farber 2013). With respect to online activity, in 2012 Google was processing 3 billion search queries daily, each one of which it stored (Mayer-Schonberger and Cukier 2013) and about 24 petabytes of data every day (Davenport et al. 2012). In 2011, Facebook's active users spent more than 9.3 billion hours a month on the site (Manyika et al. 2011), and by 2012 Facebook reported that it was processing 2.5 billion pieces of content (links, stores, photos, news, etc.), 2.7 billion 'Like' actions and 300 million photo uploads per day (Constine 2012). In 2012, over 400 millions tweets a day were produced, growing at a rate of 200 per cent a year, each tweet having 33 discrete items of metadata (Mayer-Schonberger and Cukier 2013). Much of these data are unstructured in nature. A similar explosion in structured data has taken place. For example, with respect to retail data concerning stock and sales, collected through logistics chains and checkouts, Walmart was generating more than 2.5 petabytes of data relating to more than 1 million customer transactions every hour in 2012 ('equivalent to 167 times the information contained in all the books in the Library of Congress'; Open Data Center Alliance 2012: 6), and the UK supermarket Tesco was generating more than 1.5 billion new items of data every month in 2011 (Manyika et al. 2011).

Likewise, governments and public bodies are generating vast quantities of data about their own citizens and other nations. For example, transit bodies have started to monitor the constant flow of people through transport systems, for example, collating the time and location of the use of pre-paid travel cards such as the Oyster Card in London. Many forms of tax payment, or applications for government services, are now conducted online. In 2009, the US Government produced 848 petabytes of data (TechAmerica Foundation 2012). The 16 intelligence agencies that make up US security, along with the branches of the US military, screen, store and analyse massive amounts of data hourly, with thousands of analysts employed to sift and interpret the results. To get a sense of the scale of some military intelligence projects, the ARGUS-IS project, unveiled by DARPA and the US Army in 2013, is 'a 1.8-gigapixel video surveillance platform that can resolve details as small as six inches from an altitude of 20,000 feet (6km)' (Anthony 2013). It collects '1.8 billion pixels, at 12 fps [frames per second], generat[ing] on the order of 600 gigabits per second. This equates to around 6 petabytes ... of video data per day.' Using a supercomputer, analysis is undertaken in near real-time and the system can simultaneously track up to 65 moving objects within its field of vision. This is just one project in an arsenal of similar and related intelligence projects.

Similarly, with respect to scientific projects, a personal human genome sequence consists of about 100 gigabytes of data (Vanacek 2012): multiply that across thousands of individuals and the database soon scales into terabytes and petabytes of data. When the Sloan Digital Sky Survey began operation in 2000, its telescope in New Mexico generated more observational data in the first couple of months than had previously been collected in the history of astronomy up to that point (Cukier 2010). In 2010, its archive was 140 TB of data, an amount soon to be collected every five days by the Large Synoptic Survey Telescope due to become operational in Chile in 2016 (Cukier 2010). Even more voluminous, the Large Hadron Collider at CERN, Europe's particle-physics laboratory, generates 40 terabytes every second (*The Economist* 2010). In this, and other cases, the data generated are so vast that they neither get analysed nor stored, consisting instead of transient data. Indeed, the capacity to store all these data does not exist because, although storage is expanding rapidly, it is not keeping pace with data generation (Gantz et al. 2007; Manyika et al. 2011).

EXHAUSTIVITY

With small data studies a process of sampling is used in order to produce a representative set of data from the total population of all potential data at a particular time and place. Such sampling is employed because the total population might be very large and it is unfeasible in terms of time and resources to harvest all data. In contrast, big data projects strive towards capturing entire populations (n = all), or at least much larger sample sizes than would traditionally be employed in small data studies (Mayer-Schonberger and Cukier 2013). On the one hand, this is a by-product of the technologies that are employed to generate data, along with the huge growth in the ability to store data (see Chapter 5), and on the other a conviction that 'more is better' and will provide greater representativeness and validity in the analysis.

In open systems like large scientific projects, such as measuring climatic data for weather reporting and meteorological modelling, or collecting astronomical data using a powerful telescope, the drive is towards much larger sets of data, with increased sample sizes across as many variables as possible. For example, in astronomy this means not just collecting light data, but data from across the electromagnetic spectrum, in as high a resolution as possible, for as much of the sky as possible. In the case of closed systems, such as Facebook or buying goods from an online store such as Amazon or sending e-mails, it is possible to record all the interactions and transactions that occur, as well as the level of inaction. And in these cases, that is indeed the case. Every posting, 'like', uploaded photo, link to another website, direct message, game played, periods of absence, etc., is recorded by Facebook for all of its billion or so users. Similarly, Amazon records not only every purchase and purchaser details, but also all the links clicked on and all the goods viewed on its site, as well as items placed in the shopping basket but not purchased. All e-mails are recorded by the servers on which a client e-mail is hosted, storing the whole e-mail and all associated metadata (e.g., who the e-mail was sent to or received from, the time/date, subject, attachments). Even if the e-mail is downloaded locally and deleted it is still retained on the server, with most institutions and companies keeping such data for a number of years.

Like other forms of data, spatial data has grown enormously in recent years, from real-time remote sensing and radar imagery, to large crowdsourced projects such as OpenStreetMap, to digital spatial trails created by GPS receivers being embedded in devices. The first two seek to be spatially exhaustive, capturing the terrain of the entire planet, mapping the infrastructure of whole countries and providing a creative commons licensed mapping dataset. The third provides the ability to track and trace movement across space over time; to construct individual time–space trails that can be aggregated to provide time–space models of behaviour across whole cities and regions. Together they enable detailed modelling of places and mobility, comparison across space, marketing to be targeted at particular communities, new location-based services, and data that share spatial referents to be mashed-up to create new datasets and applications that can be searched spatially (e.g., combining data about an area to create neighbourhood profiles).

Given advances in storage capacity (see Chapter 5), it seems we have reached the stage where in many cases it is easier to record everything, than to sort, sift and sample the data, recording only that which is potentially useful (and who is to know what might prove to be useful in the future?). As Zikopoulos et al. (2012) note: 'it's little wonder we're drowning in data. If we can track and record something, we typically do.' Indeed, Dumbill (2012: 7) suggests that an underlying principle of big data is 'when you can, keep everything'. This is driven by a belief that the more data there are available, the better the chance of making a valid and penetrating insight, and 'the better ... [the] chances of finding the "generators" for a new theory' (John Seely Brown, cited in Bollier 2010: 8). The strategy of seeking exhaustivity, however, contributes enormously to the data deluge, the challenge of seeing the trees from the forest, and raises a host of ethical questions concerning the scope of the data being generated and retained, and the uses to which they are being put or could be put (see Chapter 10). It also raises fundamental epistemological questions (Floridi 2012). For example, given its exhaustivity, Callebaut (2012) asks whether big data analytics is post-reductivist science. Such questions are examined in more detail in Chapter 8.

RESOLUTION AND INDEXICALITY

In addition to data exhaustivity, big data are becoming much more fine-grained in their resolution, together with a move towards strong indexicality (unique labelling and identification) (Dodge and Kitchin 2005). An example of enhanced resolution are remote sensing images. In the late 1980s, the highest resolution images of the Earth's surface available to most non-government researchers were those taken by Landsat satellites, where each pixel relates to a 30 × 30 metre parcel of land. Much of the imagery now available on Google Earth has a resolution of 2.5 × 2.5 metres, enabling much more detail to be viewed and analysed. Similarly, with respect to the output of census data, the resolution of the tertiary data has increased in many jurisdictions. In the Irish case, until recently census data were published for electoral divisions (ED) (3,409 areas with an average population of *c.*1,350, with the population per ED being much higher in cities and towns and lower in rural areas). In 2011, the census data were released for a new statistical geography called Small Areas, of which there were 18,488. These new units typically report the data for 80–150 households (Gleeson et al. 2009). The Small Areas enable analysis of the census to be conducted at neighbourhood or street level, rather than quite large areas, and for areas with roughly equal population numbers to be compared, providing a much more granular understanding of the Irish population and economy. Even more fine-grained in resolution, many data brokers are now collating large volumes of data relating to individuals and households that enables companies to individually target goods and services (see Chapter 2).

The increase in the resolution of data has been accompanied by the identification of people, products, transactions and territories becoming more indexical in nature (see Chapter 5). For example, most items for sale in a supermarket presently have a barcode. This barcode identifies the product, but not the individual item – all bottles of the same brand and range of shampoo share the same barcode – meaning that they cannot be individually discriminated. In contrast, a bottle of shampoo tagged with a RFID chip is uniquely identifiable because each chip has a unique ID code which can be read at a distance by a radio transponder. Consequently, each bottle can be tracked from the place of manufacture through the supply chain into a store and a customer's basket, creating a detailed audit trail. In other words, it has become possible to minutely trace the circulation of individual things across time and space, including those who handle each thing along its path. Similarly, information, especially that in a digital form, is being identified uniquely through digital rights management codes, for example DOIs (digital object identifiers) which can be assigned to creative works available across the Internet (e.g., reports, journals, photos, audio and video files). A DOI is a permanent ID with associated metadata, such as a URL that links to the location of the file. The use of unique identifiers enhances relationality and the ability to interconnect and join data together and provides the practical means for sorting, collating, monitoring, matching, and profiling entities (Lyon 2003a; Dodge and Kitchin 2005; Graham 2005; see also Chapter 10).

RELATIONALITY

Relationality concerns the extent to which different sets of data can be conjoined and how those conjoins can be used to answer new questions. Relationality is at the heart of relational databases (see Chapters 2 and 5), and it is the ability to create data that are highly relational that drives the vast data marketplace and the profits of data brokers and profiling companies (see Chapter 2). It is the high degree of relationality that makes censuses so useful for understanding a nation's population and how it is changing over time and space. Small data studies vary in the extent of their relationality, with those involving structured data tending to have higher degrees of interconnection than unstructured ones. That said, some form of relationality must exist between data for overarching interpretations and conclusions to be drawn from them.

Although big data often do not use a relational database structure (see Chapter 5), a core feature of their nature is strong relationality. As boyd and Crawford (2011: 2) detail, 'Big Data is fundamentally networked. Its value comes from the patterns that can be derived by making connections between pieces of data, about an individual, about individuals in relation to others, about groups of people, or simply about the structure of information itself.' And unlike relational databases, it is equally proficient at handling non-numeric, unstructured data as structured data, and in binding the two together and leveraging value from intersections. It thus becomes possible to interlink diverse sets of data – personal, transactional, interactional, social, financial, spatial, temporal, and so on – and to analyse them on an individual and collective basis for relationships and patterns.

An example of the power of such relationality is evident in President Obama's election campaigns in 2008 and 2012 that made extensive use of big data. As detailed by Issenberg (2012), Obama's team sought to quantify and track all aspects of their campaigns in 2008 and 2012, devising a whole series of metrics that were continuously recorded and mined for useful information, patterns and trends. This included the rigorous monitoring of their own actions, such as placing ads across different media, undertaking mail shots, ringing up potential voters, knocking on doors and canvassing areas, organising meetings and rallies, tracking who they had spoken to and what they had said or committed to. They supplemented this information with hundreds of randomised, large-scale experiments designed to test the effectiveness of different ways of persuading people to vote for Obama or donate funds. Obama's team combined all the information they generated with respect to voters with registration data, census and other government data, polling surveys, and data bought from a whole range of suppliers, including data brokers, credit ratings agencies, and cable TV companies. The result was a set of massive databases about every voter in the country consisting of a minimum of 80 variables (Crovitz 2012), and often many more, relating to a potential voter's demographic characteristics, their voting history, every instance in which they had been approached by the Obama campaign and their reaction, their social and economic history, their patterns of behaviour and consumption, and expressed views and opinions, with the databases updated daily during the campaign as new data was produced or bought. In cases where Obama's analysts did not know the political affiliation of a voter, and they could not access this through direct contact, they employed a sophisticated algorithm to use what variables they did have to predict a person's likely voting preference (Issenberg 2012). The result was billions of pieces of interconnected data that were used to individually profile voters, assess if they were likely to vote and how, and how they might react to different policies and stories. The interlinking of data in Obama's campaign created what Crampton et al. (2012) term an 'information amplifier effect', wherein the sum of data is more than the parts.

VELOCITY

A fundamental difference between small and big data is the dynamic nature of data generation. Small data usually consist of studies that are freeze-framed at a particular time and space. Even in longitudinal studies, the data are captured at discrete times (e.g., every few months or years). For example, censuses are generally conducted every five or ten years. In contrast, big data are generated on a much more continuous basis, in many cases in real-time or near to real-time. Rather than a sporadic trickle of data, laboriously harvested or processed, data are flowing at speed. Therefore, there is a move from dealing with batch processing to streaming data (Zikopoulos et al. 2012). On the one hand, this contributes to the issue of data volume by producing data more quickly, on the other it makes the entire data cycle much more dynamic, raising issues of how to manage a data system that is always in flux.

Velocity occurs because repeated observations are continuously made over time and/or space (Jacobs 2009) with many systems operating in perpetual, always-on mode (Dodge and Kitchin 2005). For example, websites continuously record logs that track all visits and the activity undertaken on the site; medical equipment constantly monitors vital signs, recording how a body is responding to treatment and triggering an alarm if a threshold is crossed; mobile phone companies track the location, use and identity of devices accessing their networks every few seconds; weather sensor networks monitor atmospheric indicators every few minutes and transmit their findings to a central database for incorporation into weather forecasts; transponders along a city's road and rail routes record the identity of buses and trains as they pass, enabling the public transit authority to know where all of its vehicles are at any time and to calculate the estimated arrival time at different stops; a retailer monitors the sales of thousands of different products by thousands of customers, using the data to know when to restock shelves and order from suppliers; people communicate with each other through social media sites in a never-ending flow of exchanges and interconnections; a telescope continually monitors the heavens measuring fluctuations in radio waves in order to understand the nature of the universe. In all these cases, there is a persistent stream of data requiring continual management and analysis.

Transferring and managing large volumes of dynamically produced data is a technical challenge as capacity issues can quickly create bottlenecks. For example, just as YouTube videos might freeze because the bandwidth is not sufficient to keep up with the data streaming speed required, the same effect can operate with respect to capturing and processing data, with systems unable to keep up with the flow. Solutions to the problem include increasing bandwidth capacity, data sorting and compression techniques that reduce the volume of data to be processed, and efficiency improvements in processing algorithms and data-management techniques. Analysing such streaming data is also a challenge because at no point does the system rest, and in cases such as the financial markets micro-second analysis of trades can be extremely valuable. Here, sophisticated algorithms, alongside visualisations that display dynamic data in flux, are employed to track and evaluate the system.

VARIETY

Both small and big data can be varied in their nature, being structured, unstructured or semi-structured, consisting of numbers, text, images, video, audio and other kinds of data. In big data these different kinds of data are more likely to be combined and linked, conjoining structured and unstructured data. For example, Facebook posts consist of text that is often linked to photos, or video files, or other websites, and they attract comments by other Facebook users; or a company could combine its financial data concerning sales with customer surveys that express product sentiment. Small data, in contrast, are more discrete and linked, if at all, through key identifiers and common fields. A key advance with regards to big data is how they differ from earlier forms of digital data management, which was extremely proficient at processing and storing numeric data using relational databases, and which enabled various kinds of statistical analysis. It was, however, much weaker at handling non-numeric data formats, other than to store them as flat or compressed files. As the Open Data Center Alliance (2012: 7) notes, '[p]reviously, unstructured data was either ignored or, at best, used inefficiently'. However, advances in distributed computing and database design using NoSQL structures (see Chapter 5), and in data mining and knowledge discovery techniques (see Chapter 6), have hugely increased the capacity to manage, process and extract information from unstructured data. Indeed, it is widely suggested that approximately 80 per cent of all big data is unstructured in nature, though as Grimes (2011) details, this figure has become a truism with little evidential support.

FLEXIBILITY

With small data projects, given the logistics, expense and need for representativeness in a small sample size, the research design and data management can be relatively inflexible once the fieldwork and analysis get underway. For example, it is essential that every person captured by the census fills in exactly the same form to ensure that the data are comparable across the whole population. Once the forms are printed, additional fields cannot be added, meaning that the data that can be extracted across these forms is fixed. Similarly, the relational databases in which the data are held tend to have a fixed form and are limited in scale. Likewise, in scientific experiments and environmental studies to enable comparison and replication, the research design is usually inflexible once initiated. In studies that use interviews or ethnographies, however, it is possible for the researcher to be more flexible in their approach, to have free-form questions and to adapt to unfolding situations. The coding, management and analysis of such data can also be relatively flexible, but this is partly due to the limited size and scope of the dataset.

In contrast, big data systems are designed to be flexible in nature, holding the traits of extensionality (can add new fields easily) and scalability (can expand rapidly) regardless of volume (Marz and Warren 2012). The use of NoSQL databases means that changeable data can be managed at high velocity, adapting to new fields (see Chapter 5). This means that it is possible to adapt data generation on a rolling basis and to perform adaptive testing. For example, Google, Facebook and other online platforms constantly tweak their design, capturing data about how users respond to these changes (e.g., monitoring click-throughs), analysing the results and using these to make further tweaks designed to encourage certain actions. Because the volumes of people using these sites are vast, their sample sizes are enormous, meaning they can make changes without fear of losing representativeness. For example, to return to Barack Obama's election campaign, his team ran rolling experiments on how effective different tweaks to BarackObama.com were for increasing engagement, volunteering and donations. One test evaluated the effects of changing the 'sign up' button to 'learn more', 'join us now', 'sign up now': over the course of 300,000 visits it became clear that 'join us now' led to a 20 per cent increase in people registering with the site (Issenberg 2012).

Such large-volume sites also have to be scalable, able to cope with surges in demand and data generation, where the amount of traffic would usually collapse a traditional relational database held on a single server. For example, the amount of tweets that Twitter has to deal with can fluctuate markedly, with tens of thousands being posted every few seconds during large events, such as the opening ceremony of the Olympic games or during the Superbowl final. The solution to this has been to configure a hardware system composed of distributed parts where data can be stored in databases split across many servers, enabling storage to scale as needed. Moreover, in some systems, such as Twitter, flexibility can be set by users deciding whether to include data or not. For example, in many mobile and social media apps users decide whether to include their location, and also other key metadata such those relating to identity (Gorman 2013).

CONCLUSION

Big data is a recent phenomena, and given its rapid implementation and deployment there are ongoing debates as to what constitutes big data and its associated characteristics. Some definitions, such as that big data are any dataset too large to fit in an Excel spreadsheet or be stored on a single machine (Strom 2012), are quite trite and unhelpful, reducing big data to merely volume. It is becoming clear that big data have a number of inherent characteristics that make them qualitatively different to previous forms of data. In this chapter it has been argued that big data have seven essential characteristics: volume, velocity, variety, exhaustivity, resolution/indexicality, relationality, and flexibility/scalability that distinguish them from small data (see Table 2.1).

This is an initial first-level pass at providing an ontological assessment of the nature of big data. More work is needed to assess big data generated from multiple sources to establish if there are varieties in the nature of big data. For example, it may be the case that some data hold five or six of these characteristics, but do not fulfil or are weaker in one or two. For example, a dataset may lack variety (be very structured) or volume (small, but exhaustive with n = all) or are weaker in velocity (the data are generated regularly but every month rather than continuously) or lack indexicality (it is anonymised or aggregated), yet hold the other properties. Such data are clearly not small data as discussed in Chapter 2, but are not big data as understood in a narrow sense of holding all seven characteristics. They nevertheless can be considered a form of big data. In other words, there is a need to produce a taxonomy of big data based on strong empirical evidence with case examples that would help us think through more fully the nature of such data. This needs to be accompanied by an examination of other characteristics, such as data quality, veracity, fidelity, and provenance (see Chapter 9).

The seven characteristics of big data also raise questions as to the implications of a deluge of such data. What does it mean for society, government and business to gain access to very large, exhaustive, dynamic, fine-grained, indexical, varied, relational, flexible and scalable data? To what extent can such data provide penetrating insights into the human condition or help address some of the most pressing social, political, economic and environmental issues facing the planet? Or, rather than serve the public good, will such data be used predominately to further private interests? Or serve the interests of the state? How will such data change the epistemology of science across all domains (arts and humanities, social sciences, physical and life sciences, engineering)? Chapters 7, 8 and 10 discuss these issues in detail, providing a critical reflection on the implications and consequences of big data.