Chapter 1

The data explosion

What is data?

In 431 BCE, Sparta declared war on Athens. Thucydides, in his account of the war, describes how besieged Plataean forces loyal to Athens planned to escape by scaling the wall surrounding Plataea built by Spartan-led Peloponnesian forces. To do this they needed to know how high the wall was so that they could make ladders of suitable length. Much of the Peloponnesian wall had been covered with rough pebbledash, but a section was found where the bricks were still clearly visible and a large number of soldiers were each given the task of counting the layers of these exposed bricks. Working at a distance safe from enemy attack inevitably introduced mistakes, but as Thucydides explains, given that many counts were taken, the result that appeared most often would be correct. This most frequently occurring count, which we would now refer to as *the mode*, was then used to calculate the height of the wall, the Plataeans knowing the size of the local bricks used, and ladders of the length required to scale the wall were constructed. This enabled a force of several hundred men to escape, and the episode may well be considered the most impressive example of historic data collection and analysis. But the collection, storage, and analysis of data pre-dates even Thucydides by many centuries, as we will see.

Notches have been found on sticks, stones, and bones dating back to as early as the Upper Paleolithic era. These notches are thought to represent data stored as tally marks, though this is still open to academic debate. Perhaps the most famous example is the Ishango Bone, found in the Democratic Republic of Congo in 1950, and which is estimated to be around 20,000 years old. This notched bone has been variously interpreted as a calculator or a calendar, although others prefer to explain the notches as being there just to provide a better grip. The Lebombo Bone, discovered in the 1970s in Swaziland, is even older, dating from around 35,000 BCE. With twenty-nine lines scored across it, this fragment of a baboon's fibula bears a striking resemblance to the calendar sticks still used by bushmen in distant Namibia, suggesting that this may indeed be a method that was used to keep track of data important to their civilization.

While the interpretation of these notched bones is still open to speculation, we know that one of the first well-documented uses of data is the census undertaken by the Babylonians in 3800 BCE. This census systematically documented population numbers and commodities, such as milk and honey, in order to provide the information necessary to calculate taxes. The early Egyptians also used data, in the form of hieroglyphs written on wood or papyrus, to record the delivery of goods and to keep track of taxes. But early examples of data usage are by no means confined to Europe and Africa. The Incas and their South American predecessors, keen to record statistics for tax and commercial purposes, used a sophisticated and complex system of coloured knotted strings, called *quipu*, as a decimal-based accounting system. These knotted strings, made from brightly dyed cotton or camelid wool, date back to the third millennium BCE, and although fewer than a thousand are known to have survived the Spanish invasion and subsequent attempt to eradicate them, they are among the first known examples of a massive data storage system. Computer algorithms are now being developed to try to decode the full meaning of the *quipu* and enhance our understanding of how they were used.

Although we can think of and describe these early systems as using data, the word 'data' is actually a plural word of Latin origin, with 'datum' being the singular. 'Datum' is rarely used today and 'data' is used for both singular and plural. The *Oxford English Dictionary* attributes the first

known use of the term to the 17th-century English cleric Henry Hammond in a controversial religious tract published in 1648. In it Hammond used the phrase ‘heap of data’, in a theological sense, to refer to incontrovertible religious truths. But although this publication stands out as representing the first use of the term ‘data’ in English, it does not capture its use in the modern sense of denoting facts and figures about a population of interest. ‘Data’, as we now understand the term, owes its origins to the scientific revolution in the 18th century led by intellectual giants such as Priestley, Newton, and Lavoisier; and, by 1809, following the work of earlier mathematicians, Gauss and Laplace were laying the highly mathematical foundations for modern statistical methodology.

On a more practical level, an extensive amount of data was collected on the 1854 cholera outbreak in Broad Street, London, allowing physician John Snow to chart the outbreak. By doing so, he was able to lend support to his hypothesis that contaminated water spread the disease and to show that it was not airborne as had been previously believed. Gathering data from local inhabitants he established that those affected were all using the same public water pump; he then persuaded the local parish authorities to shut it down, a task they accomplished by removing the pump handle. Snow subsequently produced a map, now famous, showing that the illness had occurred in clusters around the Broad Street pump. He continued to work in this field, collecting and analysing data, and is renowned as a pioneering epidemiologist.

Following John Snow’s work, epidemiologists and social scientists have increasingly found demographic data invaluable for research purposes, and the census now taken in many countries proves a useful source of such information. For example, data on the birth and death rate, the frequency of various diseases, and statistics on income and crime is all now collected, which was not the case prior to the 19th century. The census, which takes place every ten years in most countries, has been collecting increasing amounts of data, which eventually has resulted in more than could realistically be recorded by hand or the simple tallying machines previously used. The challenge of processing these ever-increasing amounts of census data was in part met by Herman Hollerith while working for the US Census Bureau.

By the 1870 US census, a simple tallying machine was in operation but this had limited success in reducing the work of the Census Bureau. A breakthrough came in time for the 1890 census, when Herman Hollerith's punched cards tabulator for storing and processing data was used. The time taken to process the US census data was usually about eight years, but using this new invention the time was reduced to one year. Hollerith's machine revolutionized the analysis of census data in countries worldwide, including Germany, Russia, Norway, and Cuba.

Hollerith subsequently sold his machine to the company that evolved into IBM, which then developed and produced a widely used series of punch card machines. In 1969, the American National Standards Institute (ANSI) defined the Hollerith Punched Card Code (or Hollerith Card Code), honouring Hollerith for his early punch card innovations.

Data in the digital age

Before the widespread use of computers, data from the census, scientific experiments, or carefully designed sample surveys and questionnaires was recorded on paper—a process that was time-consuming and expensive. Data collection could only take place once researchers had decided which questions they wanted their experiments or surveys to answer, and the resulting highly structured data, transcribed onto paper in ordered rows and columns, was then amenable to traditional methods of statistical analysis. By the first half of the 20th century some data was being stored on computers, helping to alleviate some of this labour-intensive work, but it was through the launch of the World Wide Web (or Web) in 1989, and its rapid development, that it became increasingly feasible to generate, collect, store, and analyse data electronically. The problems inevitably generated by the very large volume of data made accessible by the Web then needed to be addressed, and we first look at how we may make distinctions between different types of data.

The data we derive from the Web can be classified as structured, unstructured, or semi-structured.

Structured data, of the kind written by hand and kept in notebooks or in filing cabinets, is now stored electronically on spreadsheets or databases, and consists of spreadsheet-style tables with rows and columns, each row being a record and each column a well-defined field (e.g. name, address, and age). We are contributing to these structured data stores when, for example, we provide the information necessary to order goods online. Carefully structured and tabulated data is relatively easy to manage and is amenable to statistical analysis, indeed until recently statistical analysis methods could be applied only to structured data.

In contrast, unstructured data is not so easily categorized and includes photos, videos, tweets, and word-processing documents. Once the use of the World Wide Web became widespread, it transpired that many such potential sources of information remained inaccessible because they lacked the structure needed for existing analytic techniques to be applied. However, by identifying key features, data that appears at first sight to be unstructured may not be completely without structure. Emails, for example, contain structured *metadata* in the heading as well as the actual unstructured message in the text and so may be classified as semi-structured data. Metadata tags, which are essentially descriptive references, can be used to add some structure to unstructured data. Adding a word tag to an image on a website makes it identifiable and so easier to search for. Semi-structured data is also found on social networking sites, which use hashtags so that messages (which are unstructured data) on a particular topic can be identified. Dealing with unstructured data is challenging: since it cannot be stored in traditional databases or spreadsheets, special tools have had to be developed to extract useful information. In later chapters we will look at how unstructured data is stored.

The term ‘data explosion’, which heads this chapter, refers to the increasingly vast amounts of structured, unstructured, and semi-structured data being generated minute by minute; we will look next at some of the many different sources that produce all this data.

Introduction to big data

Just in researching material for this book I have been swamped by the sheer volume of data available on the Web—from websites, scientific journals, and e-textbooks. According to a recent worldwide study conducted by IBM, about 2.5 *exabytes* (Eb) of data are generated every day. One Eb is 10^{18} (1 followed by eighteen 0s) bytes (or a million *terabytes* (Tb); see the Big data byte size chart at the end of this book). A good laptop bought at the time of writing will typically have a hard drive with 1 or 2 Tb of storage space. Originally, the term 'big data' simply referred to the very large amounts of data being produced in the digital age. These huge amounts of data, both structured and unstructured, include all the Web data generated by emails, websites, and social networking sites.

Approximately 80 per cent of the world's data is unstructured in the form of text, photos, and images, and so it is not amenable to the traditional methods of structured data analysis. 'Big data' is now used to refer not just to the total amount of data generated and stored electronically, but also to specific datasets that are large in both size and complexity, with which new algorithmic techniques are required in order to extract useful information from them. These big datasets come from different sources so let's take a more detailed look at some of them and the data they generate.

Search engine data

In 2015, Google was by far the most popular search engine worldwide, with Microsoft's Bing and Yahoo Search coming second and third, respectively. In 2012, the most recent year for which data is publicly available, there were over 3.5 billion searches made per day on Google alone.

Entering a key term into a search engine generates a list of the most relevant websites, but at the same time a considerable amount of data is being collected. Web tracking generates big data. As an exercise, I searched on 'border collies' and clicked on the top website returned. Using some basic tracking software, I found that some sixty-seven third-party site connections were generated just by clicking on this one website. In order to

track the interests of people who access the site, information is being shared in this way between commercial enterprises.

Every time we use a search engine, logs are created recording which of the recommended sites we visited. These logs contain useful information such as the query term itself, the IP address of the device used, the time when the query was submitted, how long we stayed on each site, and in which order we visited them—all without identifying us by name. In addition, *clickstream logs* record the path taken as we visit various websites as well as our navigation within each website. When we surf the Web, every click we make is recorded somewhere for future use. Software is available for businesses allowing them to collect the clickstream data generated by their own website—a valuable marketing tool. For example, by providing data on the use of the system, logs can help detect malicious activity such as identity theft. Logs are also used to gauge the effectiveness of online advertising, essentially by counting the number of times an advertisement is clicked on by a website visitor.

By enabling customer identification, cookies are used to personalize your surfing experience. When you make your first visit to a chosen website, a *cookie*, which is a small text file, usually consisting of a website identifier and a user identifier, will be sent to your computer, unless you have blocked the use of cookies. Each time you visit this website, the cookie sends a message back to the website and in this way keeps track of your visits. As we will see in Chapter 6, cookies are often used to record clickstream data, to keep track of your preferences, or to add your name to targeted advertising.

Social networking sites also generate a vast amount of data, with Facebook and Twitter at the top of the list. By the middle of 2016, Facebook had, on average, 1.71 billion active users per month, all generating data, resulting in about 1.5 *petabytes* (Pb; or 1,000 Tb) of Web log data every day. YouTube, the popular video-sharing website, has had a huge impact since it started in 2005, and a recent YouTube press release claims that there are over a billion users worldwide. The valuable data produced by search engines and social

networking sites can be used in many other areas, for example when dealing with health issues.

Healthcare data

If we look at healthcare we find an area which involves a large and growing percentage of the world population and which is increasingly computerized. Electronic health records are gradually becoming the norm in hospitals and doctors' surgeries, with the primary aim being to make it easier to share patient data with other hospitals and physicians, and so to facilitate the provision of better healthcare. The collection of personal data through wearable or implantable sensors is on the increase, particularly for health monitoring, with many of us using personal fitness trackers of varying complexity which output ever more categories of data. It is now possible to monitor a patient's health remotely in real-time through the collection of data on blood pressure, pulse, and temperature, thus potentially reducing healthcare costs and improving quality of life. These remote monitoring devices are becoming increasingly sophisticated and now go beyond basic measurements to include sleep tracking and arterial oxygen saturation rate.

Some companies offer incentives in order to persuade employees to use a wearable fitness device and to meet certain targets such as weight loss or a certain number of steps taken per day. In return for being given the device, the employee agrees to share the data with the employer. This may seem reasonable but there will inevitably be privacy issues to be considered, together with the unwelcome pressure some people may feel under to opt into such a scheme.

Other forms of employee monitoring are becoming more frequent, such as tracking all employee activities on the company-provided computers and smartphones. Using customized software, this tracking can include everything from monitoring which websites are visited to logging individual keystrokes and checking whether the computer is being used for private purposes such as visiting social network sites. In the age of massive data leaks, security is of growing concern and so corporate data must be

protected. Monitoring emails and tracking websites visited are just two ways of reducing the theft of sensitive material.

As we have seen, personal health data may be derived from sensors, such as a fitness tracker or health monitoring device. However, much of the data being collected from sensors is for highly specialized medical purposes. Some of the largest data stores in existence are being generated as researchers study the genes and sequencing genomes of a variety of species. The structure of the deoxyribonucleic acid molecule (DNA), famous for holding the genetic instructions for the functioning of living organisms, was first described as a double-helix by James Watson and Francis Crick in 1953. One of the most highly publicized research projects in recent years has been the international human genome project, which determines the sequence, or exact order, of the three billion base-pairs that comprise human DNA. Ultimately, this data is helping research teams in the study of genetic diseases.

Real-time data

Some data is collected, processed, and used in real-time. The increase in computer processing power has allowed an increase in the ability to process as well as generate such data rapidly. These are systems where response time is crucial and so data must be processed in a timely manner. For example, the Global Positioning System (GPS) uses a system of satellites to scan the Earth and send back huge amounts of real-time data. A GPS receiving device, maybe in your car or smartphone (‘smart’ indicates that an item, in this case a phone, has Internet access and the ability to provide a number of services or applications (apps) that can then be linked together), processes these satellite signals and calculates your position, time, and speed.

This technology is now being used in the development of driverless or autonomous vehicles. These are already in use in confined, specialized areas such as factories and farms, and are being developed by a number of major manufacturers, including Volvo, Tesla, and Nissan. The sensors and computer programs involved have to process data in real-time to reliably

navigate to your destination and control movement of the vehicle in relation to other road users. This involves prior creation of 3D maps of the routes to be used since the sensors cannot cope with non-mapped routes. Radar sensors are used to monitor other traffic, sending back data to an external central executive computer which controls the car. Sensors have to be programmed to detect shapes and distinguish between, for example, a child running into the road and a newspaper blowing across it; or to detect, say, an emergency traffic layout following an accident. However, these cars do not yet have the ability to react appropriately to all the problems posed by an ever-changing environment.

The first fatal crash involving an autonomous vehicle occurred in 2016, when neither the driver nor the autopilot reacted to a vehicle cutting across the car's path, meaning that the brakes were not applied. Tesla, the makers of the autonomous vehicle, in a June 2016 press release referred to the 'extremely rare circumstances of the impact'. The autopilot system warns drivers to keep their hands on the wheel at all times and even checks that they are doing so. Tesla state that this is the first fatality linked to their autopilot in 130 million miles of driving, compared with one fatality per 94 million miles of regular, non-automated driving in the US.

It has been estimated that each autonomous car will generate on average 30 Tb of data daily, much of which will have to be processed almost instantly. A new area of research, called *streaming analytics*, which bypasses traditional statistical and data processing methods, hopes to provide the means for dealing with this particular big data problem.

Astronomical data

In April 2014 an International Data Corporation report estimated that, by 2020, the digital universe will be 44 trillion *gigabytes* (Gb; or 1,000 *megabytes* (Mb)), which is about ten times its size in 2013. An increasing volume of data is being produced by telescopes. For example, the Very Large Telescope in Chile is an optical telescope, which actually consists of four telescopes, each producing huge amounts of data—15 Tb per night, every night in total. It will spearhead the Large Synoptic Survey, a ten-year

project repeatedly producing maps of the night sky, creating an estimated grand total of 60 Pb (2^{50} bytes).

Even bigger in terms of data generation is the Square Kilometer Array Pathfinder (ASKAP) radio telescope being built in Australia and South Africa, which is projected to begin operation in 2018. It will produce 160 Tb of raw data per second initially, and ever more as further phases are completed. Not all this data will be stored but even so, supercomputers around the world will be needed to analyse the remaining data.

What use is all this data?

It is now almost impossible to take part in everyday activities and avoid having some personal data collected electronically. Supermarket check-outs collect data on what we buy; airlines collect information about our travel arrangements when we purchase a ticket; and banks collect our financial data.

Big data is used extensively in commerce and medicine and has applications in law, sociology, marketing, public health, and all areas of natural science. Data in all its forms has the potential to provide a wealth of useful information if we can develop ways to extract it. New techniques melding traditional statistics and computer science make it increasingly feasible to analyse large sets of data. These techniques and algorithms developed by statisticians and computer scientists search for patterns in data. Determining which patterns are important is key to the success of big data analytics. The changes brought about by the digital age have substantially changed the way data is collected, stored, and analysed. The big data revolution has given us smart cars and home-monitoring.

The ability to gather data electronically resulted in the emergence of the exciting field of data science, bringing together the disciplines of statistics and computer science in order to analyse these large quantities of data to discover new knowledge in interdisciplinary areas of application. The ultimate aim of working with big data is to extract useful information.

Decision-making in business, for example, is increasingly based on the information gleaned from big data, and expectations are high. But there are significant problems, not least with the shortage of trained data scientists capable of effectively developing and managing the systems necessary to extract the desired information.

By using new methods derived from statistics, computer science, and artificial intelligence, algorithms are now being designed that result in new insights and advances in science. For example, although it is not possible to predict exactly when and where an earthquake will occur, an increasing number of organizations are using data collected by satellite and ground sensors to monitor seismic activity. The aim is to determine approximately where big earthquakes are *likely* to occur in the long-term. For example, the US Geological Survey (USGS), a major player in seismic research, estimated in 2016 that ‘there is a 76% probability that a magnitude 7 earthquake will occur within the next 30 years in northern California’. Probabilities such as these help focus resources on measures such as ensuring that buildings are better able to withstand earthquakes and having disaster management programmes in place. Several companies in these and other areas are working with big data to provide improved forecasting methods, which were not available before the advent of big data. We need to take a look at what is special about big data.

Chapter 2

Why is big data special?

Big data didn’t just happen—it was closely linked to the development of computer technology. The rapid rate of growth in computing power and storage led to progressively more data being collected, and, regardless of who first coined the term, ‘big data’ was initially all about size. Yet it is not possible to define big data exclusively in terms of how many Pb, or even Eb, are being generated and stored. However, a useful means for talking about the ‘big data’ resulting from the data explosion is provided by the term ‘small data’—although it is not widely used by statisticians. Big datasets are certainly large and complex, but in order for us to reach a definition, we need first to understand ‘small data’ and its role in statistical analysis.

Big data versus small data

In 1919, Ronald Fisher, now widely recognized as the founder of modern statistics as an academically rigorous discipline, arrived at Rothamsted Agricultural Experimental Station in the UK to work on analysing crop data. Data has been collected from the Classical Field Experiments conducted at Rothamsted since the 1840s, including both their work on winter wheat and spring barley and meteorological data from the field

station. Fisher started the Broadbalk project which examined the effects of different fertilizers on wheat, a project still running today.

Recognizing the mess the data was in, Fisher famously referred to his initial work there as ‘raking over the muck heap’. However, by meticulously studying the experimental results that had been carefully recorded in leather-bound note books he was able to make sense of the data. Working under the constraints of his time, before today’s computing technology, Fisher was assisted only by a mechanical calculator as he, nonetheless successfully, performed calculations on seventy years of accumulated data. This calculator, known as the Millionaire, which relied for power on a tedious hand-cranking procedure, was innovative in its day, since it was the first commercially available calculator that could be used to perform multiplication. Fisher’s work was computationally intensive and the Millionaire played a crucial role in enabling him to perform the many required calculations that any modern computer would complete within seconds.

Although Fisher collated and analysed a lot of data it would not be considered a large amount today, and it would certainly not be considered ‘big data’. The crux of Fisher’s work was the use of precisely defined and carefully controlled experiments, designed to produce highly structured, unbiased sample data. This was essential since the statistical methods then available could only be applied to structured data. Indeed, these invaluable techniques still provide the cornerstone for the analysis of small, structured sets of data. However, those techniques are not applicable to the very large amounts of data we can now access with so many different digital sources available to us.

Big data defined

In the digital age we are no longer entirely dependent on samples, since we can often collect all the data we need on entire populations. But the size of these increasingly large sets of data cannot alone provide a definition for the term ‘big data’—we must include *complexity* in any definition. Instead of carefully constructed samples of ‘small data’ we are now dealing with huge

amounts of data that has not been collected with any specific questions in mind and is often unstructured. In order to characterize the key features that make data big and move towards a definition of the term, Doug Laney, writing in 2001, proposed using the three ‘v’s: volume, variety, and velocity. By looking at each of these in turn we can get a better idea of what the term ‘big data’ means.

Volume

‘Volume’ refers to the amount of electronic data that is now collected and stored, which is growing at an ever-increasing rate. Big data is big, but how big? It would be easy just to set a specific size as denoting ‘big’ in this context, but what was considered ‘big’ ten years ago is no longer big by today’s standards. Data acquisition is growing at such a rate that any chosen limit would inevitably soon become outdated. In 2012, IBM and the University of Oxford reported the findings of their Big Data Work Survey. In this international survey of 1,144 professionals working in ninety-five different countries, over half judged datasets of between 1 Tb and 1 Pb to be big, while about a third of respondents fell in the ‘don’t know’ category. The survey asked respondents to choose either one or two defining characteristics of big data from a choice of eight; only 10 per cent voted for ‘large volumes of data’ with the top choice being ‘a greater scope of information’, which attracted 18 per cent. Another reason why there can be no definitive limit based solely on size is because other factors, like storage and the type of data being collected, change over time and affect our perception of volume. Of course, some datasets are very big indeed, including, for example, those obtained by the Large Hadron Collider at CERN, the world’s premier particle accelerator, which has been operating since 2008. Even after extracting only 1 per cent of the total data generated, scientists still have 25 Pb to process annually. Generally, we can say the volume criterion is met if the dataset is such that we cannot collect, store, and analyse it using traditional computing and statistical methods. Sensor data, such as that generated by the Large Hadron Collider, is just one variety of big data, so let’s consider some of the others.

Variety

Though you may often see the terms ‘Internet’ and ‘World Wide Web’ used interchangeably, they are actually very different. The Internet is a network of networks, consisting of computers, computer networks, local area networks (LANs), satellites, and cellphones and other electronic devices, all linked together and able to send bundles of data to one another, which they do using an IP (Internet protocol) address. The World Wide Web (www, or Web), described by its inventor, T. J. Berners-Lee, as ‘a global information system’, exploited Internet access so that all those with a computer and a connection could communicate with other users through such media as email, instant messaging, social networking, and texting. Subscribers to an ISP (Internet services provider) can connect to the Internet and so access the Web and many other services.

Once we are connected to the Web, we have access to a chaotic collection of data, from sources both reliable and suspect, prone to repetition and error. This is a long way from the clean and precise data demanded by traditional statistics. Although the data collected from the Web can be structured, unstructured, or semi-structured resulting in significant variety (e.g. unstructured word-processed documents or posts found on social networking sites; and semi-structured spreadsheets), most of the big data derived from the Web is unstructured. Twitter users, for example, publish approximately 500 million 140-character messages, or *tweets*, per day worldwide. These short messages are valuable commercially and are often analysed according to whether the sentiment expressed is positive, negative, or neutral. This new area of sentiment analysis requires specially developed techniques and is something we can do effectively only by using big data analytics. Although a great variety of data is collected by hospitals, the military, and many commercial enterprises for a number of purposes, ultimately it can all be classified as structured, unstructured, or semi-structured.

Velocity

Data is now streaming continuously from sources such as the Web, smartphones, and sensors. Velocity is necessarily connected with volume: the faster data is generated, the more there is. For example, the messages on social media that now ‘go viral’ are transmitted in such a way as to have a snowball effect: I post something on social media, my friends look at it, and each shares it with their friends, and so on. Very quickly these messages make their way around the world.

Velocity also refers to the speed at which data is electronically processed. For example, sensor data, such as that being generated by an autonomous car, is necessarily generated in real-time. If the car is to work reliably, the data, sent wirelessly to a central location, must be analysed very quickly so that the necessary instructions can be sent back to the car in a timely fashion.

Variability may be considered as an additional dimension of the velocity concept, referring to the changing rates in flow of data, such as the considerable increase in data flow during peak times. This is significant because computer systems are more prone to failure at these times.

Veracity

As well as the original three ‘v’s suggested by Laney, we may add ‘veracity’ as a fourth. Veracity refers to the quality of the data being collected. Data that is accurate and reliable has been the hallmark of statistical analysis in the past century. Fisher, and others, strived to devise methods encapsulating these two concepts, but the data generated in the digital age is often unstructured, and often collected without experimental design or, indeed, any concept of what questions might be of interest. And yet we seek to gain information from this mish-mash. Take, for example, the data generated by social networks. This data is by its very nature imprecise, uncertain, and often the information posted is simply not true. So how can we trust the data to yield meaningful results? Volume can help in overcoming these problems—as we saw in Chapter 1, when Thucydides described the Plataean forces engaging the greatest possible number of soldiers counting bricks in order to be more likely to get (close to) the

correct height of the wall they wished to scale. However, we need to be more cautious, as we know from statistical theory, greater volume can lead to the opposite result, in that, given sufficient data, we can find any number of spurious correlations.

Visualization and other ‘v’s

‘V’ has become the letter of choice, with competing definitions adding or substituting such terms as ‘vulnerability’ and ‘viability’ to Laney’s original three—the most important perhaps of these additions being ‘value’ and ‘visualization’. Value generally refers to the quality of the results derived from big data analysis. It has also been used to describe the selling by commercial enterprises of data to firms who then process it using their own analytics, and so it is a term often referred to in the data business world.

Visualization is not a characterizing feature of big data, but it is important in the presentation and communication of analytic results. The familiar static pie charts and bar graphs that help us to understand small datasets have been further developed to aid in the visual interpretation of big data, but these are limited in their applicability. Infographics, for example, provide a more complex presentation but are static. Since big data is constantly being added to, the best visualizations are interactive for the user and updated regularly by the originator. For example, when we use GPS for planning a car journey, we are accessing a highly interactive graphic, based on satellite data, to track our position.

Taken together, the four main characteristics of big data—volume, variety, velocity, and veracity—present a considerable challenge in data management. The advantages we expect to gain from meeting this challenge and the questions we hope to answer with big data can be understood through data mining.

Big data mining