1

CONCEPTUALISING DATA

Data are commonly understood to be the raw material produced by abstracting the world into categories, measures and other representational forms – numbers, characters, symbols, images, sounds, electromagnetic waves, bits – that constitute the building blocks from which information and knowledge are created. Data are usually representative in nature (e.g., measurements of a phenomena, such as a person's age, height, weight, colour, blood pressure, opinion, habits, location, etc.), but can also be implied (e.g., through an absence rather than presence) or derived (e.g., data that are produced from other data, such as percentage change over time calculated by comparing data from two time periods), and can be either recorded and stored in analogue form or encoded in digital form as bits (binary digits). Good-quality data are discrete and intelligible (each datum is individual, separate and separable, and clearly defined), aggregative (can be built into sets), have associated metadata (data about data), and can be linked to other datasets to provide insights not available from a single dataset (Rosenberg 2013). Data have strong utility and high value because they provide the key inputs to the various modes of analysis that individuals, institutions, businesses and science employ in order to understand and explain the world we live in, which in turn are used to create innovations, products, policies and knowledge that shape how people live their lives.

Data then are a key resource in the modern world. Yet, given their utility and value, and the amount of effort and resources devoted to producing and analysing them, it is remarkable how little conceptual attention has been paid to data in and of themselves. In contrast, there are thousands of articles and books devoted to the philosophy of information and knowledge. Just as we tend to focus on buildings and neighbourhoods when considering cities, rather than the bricks and mortar used to build them, so it is the case with data. Moreover, just as we think of bricks and mortar as simple building blocks rather than elements that are made within factories by companies bound within logistical, financial, legal and market concerns, and are distributed, stored and traded, so we largely do with data. Consequently, when data are the focus of enquiry it is usually to consider, in a largely technical sense, how they should be generated and analysed, or how they can be leveraged

into insights and value, rather than to consider the nature of data from a more conceptual and philosophical perspective.

With this observation in mind, the principal aim of this book is threefold: to provide a detailed reflection on the nature of data and their wider assemblages; to chart how these assemblages are shifting and mutating with the development of new data infrastructures, open data and big data; and to think through the implications of these new data assemblages with respect to how we make sense of and act in the world. To supply an initial conceptual platform, in this chapter the forms, nature and philosophical bases of data are examined in detail. Far from being simple building blocks, the discussion will reveal that data are a lot more complex. While many analysts may accept data at face value, and treat them as if they are neutral, objective, and pre-analytic in nature, data are in fact framed technically, economically, ethically, temporally, spatially and philosophically. Data do not exist independently of the ideas, instruments, practices, contexts and knowledges used to generate, process and analyse them (Bowker 2005; Gitelman and Jackson 2013). Thus, the argument developed is that understanding data and the unfolding data revolution requires a more nuanced analysis than much of the open and big data literature presently demonstrates.

WHAT ARE DATA?

Etymologically the word data is derived from the Latin *dare*, meaning 'to give'. In this sense, data are raw elements that *can be* abstracted from (given by) phenomena – measured and recorded in various ways. However, in general use, data refer to those elements that *are* taken; extracted through observations, computations, experiments, and record keeping (Borgman 2007). Technically, then, what we understand as data are actually capta (derived from the Latin *capere*, meaning 'to take'); those units of data that have been selected and harvested from the sum of all potential data (Kitchin and Dodge 2011). As Jensen (1950: ix, cited in Becker 1952: 278) states:

> it is an unfortunate accident of history that the term *datum* ... rather than *captum* ... should have come to symbolize the unit-phenomenon in science. For science deals, not with 'that which has been given' by nature to the scientist, but with 'that which has been taken' or selected from nature by the scientist in accordance with his purpose.

Strictly speaking, then, this book should be entitled *The Capta Revolution*. However, since the term data has become so thoroughly ingrained in the language of the academy and business to mean capta, rather than confuse the matter further it makes sense to continue to use the term data where capta would be more appropriate.

Beyond highlighting the etymological roots of the term, what this brief discussion starts to highlight is that data harvested through measurement are always a selection from the total sum of all possible data available – what we have chosen to take from all that could potentially be given. As such, data are inherently partial, selective and representative, and the distinguishing criteria used in their capture has consequence.

Other scholars have noted that what has been understood as data has changed over time with the development of science. Rosenberg (2013) details that the term 'data' was first used in the English language in the seventeenth century. As a concept then it is very much tied to that of modernity and the growth and evolution of science and new modes of producing, presenting and debating knowledge in the seventeenth and eighteenth century that shifted information and argument away from theology, exhortation and sentiment to facts, evidence and the testing of theory through experiment (Poovey 1998; Garvey 2013; Rosenberg 2013). Over time, data came to be understood as being pre-analytical and pre-factual, different in nature to facts, evidence, information and knowledge, but a key element in the constitution of these elements (though often the terms and definitions of data, facts, evidence, information and knowledge are conflated). As Rosenberg (2013: 18) notes,

> facts are ontological, evidence is epistemological, data is rhetorical. A datum may also be a fact, just as a fact may be evidence ... [T]he existence of a datum has been independent of any consideration of corresponding ontological truth. When a fact is proven false, it ceases to be a fact. False data is data nonetheless.

In rhetorical terms, data are that which exists prior to argument or interpretation that converts them to facts, evidence and information (Rosenberg 2013). From this perspective, data hold certain precepts: they are abstract, discrete, aggregative (they can be added together) (Rosenberg 2013), and are meaningful independent of format, medium, language, producer and context (i.e., data hold their meaning whether stored as analogue or digital, viewed on paper or screen or expressed in any language, and 'adhere to certain non-varying patterns, such as the number of tree rings always being equal to the age of the tree') (Floridi 2010). Floridi (2008) contends that the support-independence of data is reliant on three types of neutrality: taxonomic (data are relational entities defined with respect to other specific data); typological (data can take a number of different non-mutually exclusive forms, e.g., primary, secondary, metadata, operational, derived); and genetic (data can have a semantics independent of their comprehension; e.g., the Rosetta Stone hieroglyphics constitute data regardless of the fact that when they were discovered nobody could interpret them).

Not everyone who thinks about or works with data holds such a narrow rhetorical view. How data are understood has not just evolved over time, it varies with

respect to perspective. For example, Floridi (2008) explains that from an epistemic position data are collections of facts, from an informational position data are information, from a computational position data are collections of binary elements that can be processed and transmitted electronically, and from a diaphoric position data are abstract elements that are distinct and intelligible from other data. In the first case, data provide the basis for further reasoning or constitute empirical evidence. In the second, data constitute representative information that can be stored, processed and analysed, but do not necessarily constitute facts. In the third, data constitute the inputs and outputs of computation but have to be processed to be turned into facts and information (for example, a DVD contains gigabytes of data but no facts or information per se) (Floridi 2005). In the fourth, data are meaningful because they capture and denote variability (e.g., patterns of dots, alphabet letters and numbers, wavelengths) that provides a signal that can be interpreted. As discussed below, other positions include understanding data as being socially constructed, as having materiality, as being ideologically loaded, as a commodity to be traded, as constituting a public good, and so on. The point is, data are never simply just data; how data are conceived and used varies between those who capture, analyse and draw conclusions from them.

KINDS OF DATA

Whether data are pre-factual and rhetorical in nature or not, it is clear that data are diverse in their characteristics, which shape in explicit terms how they are handled and what can be done with them. In broad terms, data vary by form (qualitative or quantitative), structure (structured, semi-structured or unstructured), source (captured, derived, exhaust, transient), producer (primary, secondary, tertiary), and type (indexical, attribute, metadata).

Quantitative and qualitative data

Data can take many material forms including numbers, text, symbols, images, sound, electromagnetic waves, or even a blankness or silence (an empty space is itself data). These are typically divided into two broad categories. *Quantitative data* consist of numeric records. Generally, such data are extensive and relate to the physical properties of phenomena (such as length, height, distance, weight, area, volume), or are representative and relate to non-physical characteristics of phenomena (such as social class, educational attainment, social deprivation, quality of life rankings). Quantitative data have four different levels of measurement which delimit how they can be processed and analysed (Kitchin and Tate 1999, see also Table 1.1). Such data can be analysed using visualisations, a variety of descriptive and inferential statistics, and be used as the inputs to predictive and simulation models.

Table 1.1 Levels of data measurement

Levels of measurement	Definition	Example
Nominal data	Categorical in nature, with observations recorded into discrete units.	Unmarried, married, divorced, widowed
Ordinal data	Observations that are placed in a rank order, where certain observations are greater than others.	Low, medium, high
Interval data	Measurements along a scale which possesses a fixed but arbitrary interval and an arbitrary origin. Addition or multiplication by a constant will not alter the interval nature of the observations. Data can either be continuous (e.g., time or length) or discrete (e.g., counts of a phenomenon) in nature.	Temperature along the Celsius scale
Ratio data	Similar to interval data except the scale possesses a true zero origin, and multiplication by a constant will not alter the ratio nature of the observations.	Exam marks on a scale of 0–100

In contrast, *qualitative data* are non-numeric, such as texts, pictures, art, video, sounds, and music. While qualitative data can be converted into quantitative data, the translation involves significant reduction and abstraction and much of the richness of the original data is lost by such a process. Consequently, qualitative data analysis is generally practised on the original materials, seeking to tease out and build up meaning and understanding rather than subjecting the data to rote, computational techniques. However, significant progress is being made with respect to processing and analysing qualitative data computationally through techniques such as machine learning and data mining (see Chapter 6).

Structured, semi-structured and unstructured data

Structured data are those that can be easily organised, stored and transferred in a defined data model, such as numbers/text set out in a table or relational database that have a consistent format (e.g., name, date of birth, address, gender, etc). Such data can be processed, searched, queried, combined, and analysed relatively straightforwardly using calculus and algorithms, and can be visualised using various forms of graphs and maps, and easily processed by computers. *Semi-structured data* are loosely structured data that have no predefined data model/schema and thus cannot be held in a relational database. Their structure are irregular, implicit, flexible and often nested hierarchically, but they have a reasonably consistent set of fields and the data are tagged thus, separating content semantically and providing

loose, self-defining content metadata and a means to sort, order and structure the data. An example of such data are XML-tagged web pages (pages made using Extensible Markup Language [XML] which encode documents in a format that is both human- and machine-readable; Franks 2012; see linked data in Chapter 3).

In contrast, *unstructured data* do not have a defined data model or common identifiable structure. Each individual element, such as narrative text or photo, may have a specific structure or format, but not all data within a dataset share the same structure. As such, while they can often be searched and queried, they are not easily combined or computationally analysed. Such unstructured data are usually qualitative in nature, but can often be converted into structured data through classification and categorisation. Until relatively recently, very large datasets were typically structured in form because they were generally much easier to process, analyse and store. In the age of big data, many massive datasets consist of semi- or unstructured data, such as Facebook posts, tweets, uploaded pictures and videos, and blogs, and some estimates suggest that such data are growing at 15 times the rate of structured data (Zikopoulos et al. 2012), with advances in database design (such as NoSQL databases that do not use the tabular models of relational databases, see Chapter 5) and machine learning techniques (see Chapter 6) aiding storage and analysis.

Captured, exhaust, transient and derived data

There are two primary ways in which data can be generated. The first is that data can be *captured* directly through some form of measurement such as observation, surveys, lab and field experiments, record keeping (e.g., filling out forms or writing a diary), cameras, scanners and sensors. In these cases, data are usually the deliberate product of measurement; that is, the intention was to generate useful data. In contrast, *exhaust data* are inherently produced by a device or system, but are a by-product of the main function rather than the primary output (Manyika et al. 2011). For example, an electronic checkout till is designed to total the goods being purchased and to process payment, but it also produces data that can be used to monitor stock, worker performance and customer purchasing. Many software-enabled systems produce such exhaust data, much of which have become valuable sources of information. In other cases, exhaust data are *transient* in nature; that is, they are never examined or processed and are simply discarded, either because they are too voluminous or unstructured in nature, or costly to process and store, or there is a lack of techniques to derive value from them, or they are of little strategic or tactical use (Zikopoulos et al. 2012; Franks 2012). For example, Manyika et al. (2011: 3) report that 'health care providers ... discard 90 percent of the data that they generate (e.g., almost all real-time video feeds created during surgery)'.

Captured and exhaust data are considered 'raw' in the sense that they have not been converted or combined with other data. In contrast, *derived data* are produced through additional processing or analysis of captured data. For example,

captured data might be individual traffic counts through an intersection and derived data the total number of counts or counts per hour. The latter have been derived from the former. Captured data are often the input into a model, with derived data the output. For example, traffic count data might be an input into a transportation model with the output being predicted or simulated data (such as projected traffic counts at different times or under different conditions). In the case of a model, the traffic count data are likely to have been combined with other captured or derived data (such as type of vehicle, number of passengers, etc.) to create new derived data for input into the model. Derived data are generated for a number of reasons, including to reduce the volume of data to a manageable amount and to produce more useful and meaningful measures. Sometimes the original captured data might be processed to varying levels of derivation depending on its intended use. For example, the NASA Earth Observing System organises its data into six levels that run from unprocessed captured data, through increasing degrees of processing and analysis, to model outputs based on analyses of lower-level data (Borgman 2007; see Table 1.2).

Table 1.2 The six levels of data of NASA's Earth Observing System

Data level	Description
Level 0	Reconstructed, unprocessed instrument and payload data at full resolution, with any and all communications artefacts (e.g., synchronisation frames, communications headers, duplicate data) removed.
Level 1A	Reconstructed, unprocessed instrument data at full resolution, time-referenced, and annotated with ancillary information, including radiometric and geometric calibration coefficients and georeferencing parameters computed and appended but not applied to Level 0 data.
Level 1B	Level 1A data that have been processed to sensor units.
Level 2	Derived geophysical variables at the same resolution and location as Level 1 source data.
Level 3	Variables mapped on uniform space-time grid scales, usually with some completeness and consistency.
Level 4	Model output or results from analyses of lower-level data (e.g., variables derived from multiple measurements).

Source: Adapted from https://earthdata.nasa.gov/data/standards-and-references/processing-levels

Primary, secondary and tertiary data

Primary data are generated by a researcher and their instruments within a research design of their making. *Secondary data* are data made available to others to reuse and analyse that are generated by someone else. So one person's primary data can be

another person's secondary data. *Tertiary data* are a form of derived data, such as counts, categories, and statistical results. Tertiary data are often released by statistical agencies rather than secondary data to ensure confidentiality with respect to whom the data refer. For example, the primary data of the Irish census are precluded from being released as secondary data for 100 years after generation; instead the data are released as summary counts and categorical tertiary data. Many researchers and institutions seek to generate primary data because they are tailored to their specific needs and foci, whereas these design choices are not available to those analysing secondary or tertiary data. Moreover, those using secondary and tertiary data as inputs for their own studies have to trust that the original research is valid.

In many cases researchers will combine primary data with secondary and tertiary data to produce more valuable derived data. For example, a retailer might seek to create a derived dataset that merges their primary sales data with tertiary geodemographics data (data about what kind of people live in different areas, which are derived from census and other public and commercial data) in order to determine which places to target with marketing material. Secondary and tertiary data are valuable because they enable replication studies and the building of larger, richer and more sophisticated datasets. They later produce what Crampton et al. (2012) term 'data amplification'; that is, data when combined enables far greater insights by revealing associations, relationships and patterns which remain hidden if the data remain isolated. As a consequence, the secondary and tertiary data market is a multi-billion dollar industry (see Chapter 2).

Indexical and attribute data and metadata

Data also vary in kind. *Indexical data* are those that enable identification and linking, and include unique identifiers, such as passport and social security numbers, credit card numbers, manufacturer serial numbers, digital object identifiers, IP and MAC addresses, order and shipping numbers, as well as names, addresses, and zip codes. Indexical data are important because they enable large amounts of non-indexical data to be bound together and tracked through shared identifiers, and enable discrimination, combination, disaggregation and re-aggregation, searching and other forms of processing and analysis. As discussed in Chapter 4, indexical data are becoming increasingly common and granular, escalating the relationality of datasets. *Attribute data* are data that represent aspects of a phenomenon, but are not indexical in nature. For example, with respect to a person the indexical data might be a fingerprint or DNA sequence, with associated attribute data being age, sex, height, weight, eye colour, blood group, and so on. The vast bulk of data that are generated and stored within systems are attribute data.

Metadata are data about data. Metadata can either refer to the data content or the whole dataset. Metadata about the content includes the names and descriptions of

specific fields (e.g., the column headers in a spreadsheet) and data definitions. These metadata help a user of a dataset to understand its composition and how it should be used and interpreted, and facilitates the conjoining of datasets, interoperability and discoverability, and to judge their provenance and lineage. Metadata that refers to a dataset as a whole has three different forms (NISO 2004). Descriptive metadata concerns identification and discovery and includes elements such as title, author, publisher, subject, and description. Structural metadata refers to the organisation and coverage of the dataset. Administrative metadata concerns when and how the dataset was created, details of the technical aspects of the data, such as file format, and who owns and can use the data. A common metadata standard for datasets that combines these three types of metadata is the Dublin Core (http://dublincore.org/). This standard requires datasets to have 15 accompanying metadata fields: title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, and rights. Metadata are essential components of all datasets, though they are often a neglected element of data curation, especially amongst researchers who are compiling primary data for their own use rather than sharing.

DATA, INFORMATION, KNOWLEDGE, WISDOM

What unites these various kinds of data is that they form the base or bedrock of a knowledge pyramid: data precedes information, which precedes knowledge, which precedes understanding and wisdom (Adler 1986; Weinberger 2011). Each layer of the pyramid is distinguished by a process of distillation (reducing, abstracting, processing, organising, analysing, interpreting, applying) that adds organisation, meaning and value by revealing relationships and truths about the world (see Figure 1.1).

While the order of the concepts within the pyramid is generally uncontested, the nature and difference between concepts often varies between schools of thought. Information, for example, is a concept that is variously understood across scholars. For some, information is an accumulation of associated data, for others it is data plus meaning, or the signal in the noise of data, or a multifaceted construct, or tertiary data wherein primary data has been reworked into analytical form. To a physicist, data are simply zeros and ones, raw bits; they are noise. Information is when these zeros and ones are organised into distinct patterns; it is the signal (von Baeyer 2003). Airwaves and communication cables then are full of flowing information – radio and television signals, telephone conversations, internet packets – meaningful patterns of data within the wider spectrum of noise. For others, information is a broader concept. Floridi (2010: 74), for example, identifies three types of information:

Factual: information *as* reality (e.g., patterns, fingerprints, tree rings)

Instructional: information *for* reality (e.g., commands, algorithms, recipes)

Semantic: information *about* reality (e.g., train timetables, maps, biographies).

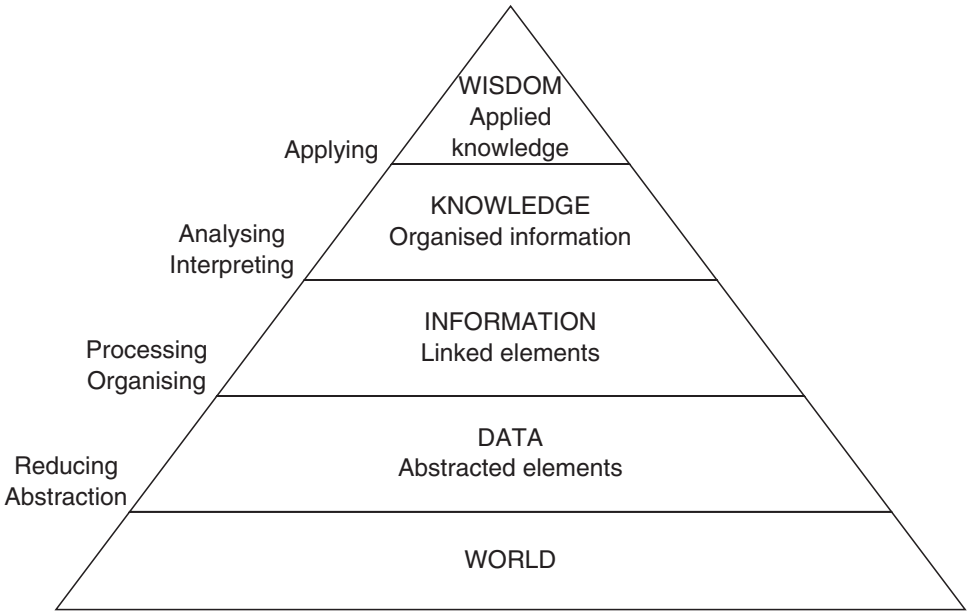

Figure 1.1 Knowledge pyramid (adapted from Adler 1986 and McCandless 2010)

The first is essentially meaningful data, what are usually termed facts. These are data that are organised and framed within a system of measurement or an external referent that inherently provides a basis to establish an initial meaning that holds some truth. Information also extends beyond data and facts through adding value that aids interpretation. Weinberger (2011: 2) thus declares: 'Information is to data what wine is to the vineyard: the delicious extract and distillate.' Such value could be gained through sorting, classifying, linking, or adding semantic content through some form of text or visualisation that informs about something and/or instructs what to do (for example, a warning light on a car's dashboard indicating that the battery is flat and needs recharging, Floridi, 2010). Case (2002; summarised in Borgman 2007: 40) argues that differences in the definition of information hinge on five issues:

> uncertainty, or whether something has to reduce uncertainty to qualify as information; physicality, or whether something has to take on a physical form such as a book, an object, or the sound waves of speech to qualify as information; structure/process, or whether some set of order or relationships is required; intentionality, or whether someone must intend that something be communicated to qualify as information; and truth, or whether something must be true to qualify as information.

Regardless of how it is conceived, Floridi (2010) notes that given that information adds meaning to data, it gains currency as a commodity. It is, however, a particular kind of commodity, possessing three main properties (which data also share):

Non-rivalrous: more than one entity can possess the same information (unlike material goods)

Non-excludable: it is easily shared and it takes effort to seek to limit such sharing (such as enforcing intellectual property rights agreements or inserting pay walls)

Zero marginal cost: once information is available, the cost of reproduction is often negligible.

While holding the properties of being non-rivalrous and non-excludable, because information is valuable many entities seek to limit and control its circulation, thus increasing its value. Much of this value is added through the processes enacted in the information life cycle (Floridi 2010):

Occurrence: discovering, designing, authoring

Transmission: networking, distributing, accessing, retrieving, transmitting

Processing and management: collecting, validating, modifying, organising, indexing, classifying, filtering, updating, sorting, storing

Usage: monitoring, modelling, analysing, explaining, planning, forecasting, decision-making, instructing, educating, learning.

It is through processing, management and usage that information is converted into the even more valuable knowledge.

As with all the concepts in the pyramid, knowledge is similarly a diversely understood concept. For some, knowledge is the 'know-how that transforms information into instructions' (Weinberger 2011: 3). For example, semantic information can be linked into recipes (first do this, then do that ...) or a conditional form of inferential procedures (if such and such is the case do this, otherwise do this) (Floridi 2010). In this framing, information is structured data and knowledge is actionable information (Weinberger 2011). In other words, 'knowledge is like the recipe that turns information into bread, while data are like the atoms that make up the flour and the yeast' (Zelany 1987, cited in Weinberger 2011). For others, knowledge is more than a set of instructions; it can be a practical skill, a way of knowing how to undertake or achieve a task, or a system of thought that coherently links together information to reveal a wider picture about a phenomenon. Creating knowledge involves applying complex cognitive processes such as perception, synthesis, extraction, association, reasoning and communication to information. Knowledge has more value than information because it provides the basis for understanding, explaining and drawing insights about the world, which can be used to formulate policy and actions. Wisdom, the pinnacle of the knowledge pyramid, is being able to sagely apply knowledge.

While not all forms of knowledge are firmly rooted in data – for example, conjecture, opinions, beliefs – data are clearly a key base material for how we make sense of the world. Data provide the basic inputs into processes such as collating, sorting, categorising, matching, profiling, and modelling that seek to create information and knowledge in order to understand, predict, regulate and control phenomena. And generating data over time and in different locales enables us to track, evaluate and compare phenomena across time, space and scale. Thus, although information and knowledge are rightly viewed as being higher order and more valuable concepts, data are nonetheless a key ingredient with significant latent value that is realised when converted to information and knowledge. Whoever then has access to high-quality and extensive data has a competitive advantage over those excluded in being able to generate understanding and wisdom. A key rationale for the open data movement, examined in Chapter 3, is gaining access to the latent value of administrative and public sector datasets.

FRAMING DATA

So far in this chapter, data have already started to be framed conceptually in terms of discussing the ontology of data (what data are), their different forms, and how they fit within the pyramid of knowledge. There is a myriad of other ways in which data can be thought about and understood, for example from a technical perspective concerning data quality, validity, reliability, authenticity and usability, and how they can be processed, structured, shared and analysed; or an ethical perspective concerning the reasons why data are generated and the uses to which data are put; or a political or economic perspective that considers how data are normatively conceived and contested as public goods, political capital, intellectual property or a commodity, and how they are regulated and traded; or a spatial and/or temporal perspective that considers how technical, ethical, political and economic regimes concerning data production and their uses develop and mutate across space and time; or a philosophical perspective that considers the varied and contested ontologies and epistemologies of data. Many of the issues, discussed in brief in this section, are returned to throughout the book.

Technically

Across all disciplines, data are considered from a normative, technical viewpoint. What is at stake is the extent to which methods of capture and measurement generate certain, clean, and accurate data, and how such data can and should be processed, structured, shared and analysed in ways that maintain their integrity, thus ensuring that reliable and valid conclusions can be drawn from them. There

are always doubts about the veracity of data because they are inherently abstracted, generalised and approximated through their production (Goodchild 2009). Considerable attention is thus directed at issues such as data representativeness, uncertainty, reliability, error, bias, and calibration within research design and implementation, with this information recorded as metadata.

Given that data are a surrogate for some aspect of a phenomenon – light representing a star, physical characteristics representing a plant, words representing a person's thoughts – *representativeness* relates to how well data capture the phenomenon they seek to represent, and how well the sample of data generated represents the overall population. With respect to the former, the key question is the extent to which we can be confident that scientific techniques accurately capture the phenomenon in question. This has been a particular problem in the social sciences and humanities and has proven difficult to resolve. For example, it is well noted that what people say they will do and what they do are often quite different, and what people do is often not what they intended. There is therefore a question over how well interview data represent human behaviour, or how well behaviour represents conscious thought. Similarly, there are concerns over the extent to which key indicators adequately capture and represent how a domain is performing. For example, to what extent do indicators such as citation counts, an h-index, and patents registered denote high-quality performance by university staff (with respect to humanities faculty they are considered very poor indicators)? The solution has been to try and develop more and more sophisticated research designs that counteract the shortcomings of different methods, or to largely ignore the shortcomings.

With respect to how well a sample represents a population, we might decide to generate detailed, longitudinal, astronomical data with respect to 50 stars in order to better understand their nature, but to what extent can we be confident that these 50 stars represent the qualities of the billions of stars that exist? Even in the age of big data, which seeks to be exhaustive rather than selective in data generation (see Chapter 4), the data are inherently a sample (not all people use social media, or shop with credit cards, and indeed many people across the world do not have access to phones or computers), meaning the data are representative of a set of people, even if that set is very large. Again, the solution has been to devise a range of sampling techniques that seek to ensure representativeness under different conditions (often dependent on the sample being random), and statistical methods that calculate the extent to which we can be confident that the sample represents the population (Kitchin and Tate 1999).

Reliability concerns the repeatability or consistency in obtaining the same finding from the administering of a research instrument. Golledge and Stimson (1997) describe three kinds of reliability: (1) quixotic reliability, where a single method of observation continually yields an unvarying measurement; (2) diachronic reliability, the stability of an observation through

time; and (3) synchronic reliability, the similarity of observations within the same time period. Reliability is important because it is generally accepted that the more consistent a measure in producing data, the more confidence can be attributed to it.

Error is the difference between a measured and a real value, and can include absences (missing data), mistakes (such as miscoding or misclassification or the misapplication of a technique), and misunderstandings. *Bias* is a particular kind of error, where the data are skewed due to a consistent pattern of error. Bias is usually caused by the method, instrument or sampling technique used to generated the data having an undue influence on what data are produced, or can be introduced due to the ideological position or aspirations of the researcher often in a non-deliberate manner (Kitchin 1996). Processing techniques such as aggregation can introduce bias by reducing variance in a dataset leading to ecological fallacy errors – that is, assuming that an aggregate value accurately represents the individuals aggregated (for example, if we had two people weighing 50 kilograms and two weighing 150 kilograms their mean aggregate weight would be 100 kilograms, yet nobody in the set weighs that amount) (Kitchin and Fotheringham 1997). *Uncertainty* concerns the extent to which a researcher can be confident concerning the accuracy of the data and any analysis based on them. With respect to quantitative data it relates to the certainty of a statistical test given the data inputted, and is usually calculated as probabilities and expressed as confidence levels (Goodchild 2009). Uncertainty with respect to qualitative data is more likely to be assessed by expert judgement based on prior experience.

Underpinning the drive to tackle these concerns is a belief that such issues arise due to human frailties in research design or deficiencies in the instruments or methods used and that they can be fixed through technical solutions. That is, they can be addressed by improving the quality of procedures and equipment used, implementing regimes of standardisation that provide known benchmarks of data quality (such as those endorsed by the ISO), and finding ways to compensate for uncertainty, error and bias in the modes of analysis employed.

Ethically

Ethics is concerned with thought and practice related to value concepts such as justice, equality, fairness, honesty, respect, rights, entitlements and care. Every society operates with respect to a mix of commonsensical, informal and taken-for-granted moral values, and highly codified ethical positions enshrined in rules, principles, policies, licences and laws, subject to enforcement by state and other agencies. These ethical positions are often contested, with different groups taking contrasting views on values themselves and the extent to which ethical stances should be legislated for, and their debate is an exercise in moral philosophy. Such contestation also exists with respect to data, especially concerning what data are

generated and the means of production, how data are shared, traded and protected, and to what ends data are employed.

While some data are considered relatively benign, for example measurements relating to the weather, other data are considered to be highly sensitive, for example those related to individuals which can be used to produce a detailed picture of the lives they lead and to regulate those lives. In some cases, generating data might do harm, for example interviewing the survivors of war crimes might cause psychological distress. Here, there are questions about the extent to which data generation and various forms of dataveillance (surveillance enacted through the processing and analysing of data records) and data analysis infringe on privacy and other human rights, and can be used to actively socially sort individuals (provide differential service based on their characteristics) (Graham 2005). These concerns are exacerbated given that digital data can be easily combined, shared and traded, and we live in an age of widespread invasive data generation and surveillance. It is perhaps no surprise then that agencies funding academic research and higher education institutions now routinely evaluate the ethical dimensions of research projects as to their potential wider implications, and nations have enacted legislation, such as data protection acts and privacy laws, to try and prevent the misuses and abuses of data. These and related issues are discussed more fully in Chapter 10.

Politically and economically

A consideration of the ethics of data starts to reveal the ways in which data are framed by wider political and economical concerns. What data are generated, and how they are processed, analysed and employed are contextualised with respect to: how they are normatively conceived within a population and employed by states, and notions of how they should be regulated and legislated for; the discourses employed within discursive regimes that support or oppose their generation and application; decision-making about funding and investing in data; the unfolding of capitalism and the ways in which data are used to manage endeavours and leverage value and profit; and are traded as a commodity with the emergence of a multi-billion-dollar data marketplace made up of a diverse set of players (producers, aggregators, sellers, analysts, and consumers; see Chapter 2). Those producing data have to consider public and political opinion, ethical considerations, the regulatory environment, the funding available, and the soundness of their investment vis-à-vis resourcing. And those in charge of the legislative and funding arenas have to ponder and make decisions about how to shape the landscape in which producers and users of data operate, as well as consider their own data regimes and what they reveal about their agendas, priorities, and modes of governance and governmentality (Lauriault 2012).

In both cases, a diverse set of politics and economic rationalities is at play, with competing voices seeking to influence opinion and the wider data terrain. The open data movement, for example, casts data as a public good that should constitute a commons and be freely accessible (with the exception of sensitive, personal data) or be accessible through fair use agreements. In contrast, business views data as a valuable commodity that, on the one hand, needs to be protected through intellectual property regimes (copyright, patents, ownership rights) and, on the other, not be so tied down by ethical concerns that they cannot be exploited for capital gain. For communities and states, data are a means by which political agendas and work can be legitimated, conducted and contested by enabling the construction of evidence-informed narratives and counter-discourses that have greater rhetorical value than anecdote or sentiment (Wilson 2011; Garvey 2013). In other words, data constitute in Foucault's (1981) terms a form of power/knowledge; a means through which people, phenomena and territory can be surveyed and regulated (Lauriault 2012). These alternative interests can often become aligned in paradoxical ways, though they may have quite different agendas, for example the support of big business for the open data movement with respect to public data (see Chapter 3). In other words, data are manifested and situated within complex and contested political economies and, at the same time, they are used to shape such debates and regimes.

Moreover, data constitute an economic resource, one that is a key component of the next phase of the knowledge economy, reshaping the mode of production to one that it is data-driven (see Chapter 7). Since the late 1980s, scholars such as Castells (1988, 1996) have argued that the latest cycle of capitalism is underpinned by the production of knowledge that creates new products and forms of labour, facilitates economic restructuring, and enhances productivity, competitiveness, efficiencies, sustainability and capital accumulation. Big data, in particular, is the latest development in deepening and advancing this cycle, providing a wealth of evidence that is being used by companies to, on the one hand, monitor and evaluate company performance in real time, reduce waste and fraud, and improve corporate strategy, planning and decision-making and, on the other, to design new commodities, identify and target new markets, implement dynamic pricing, realise untapped potential, and gain competitive advantage (Manyika et al. 2011; Zikopoulos et al. 2012). In so doing, the production and analysis of data enables companies to be run more intelligently with respect to how they are organised and operate, promoting flexibility and innovation, reducing risks, costs and operational losses, improving customer experience, and maximising return on investment and profits. By driving capital accumulation, big data facilitates new divisions of labour and the next round of uneven development. Data can thus be understood as an agent of capital interests.

Temporally and spatially

Data have both a temporality and a spatiality. What data are produced and the ways in which they are processed, analysed, stored or discarded varies across time and space; data and the assemblages surrounding them have histories and geographies. How data are processed and analysed mutates over time, affected by organisational change and improvements in enumeration and administration, new laws regarding data handling and protection, new technologies, new methods of data sorting and analysis, varying statistical geographies (such as new local area or county boundaries), and new statistical techniques. Moreover, the data assemblages operating in one jurisdiction can be quite different from another. Even within a jurisdiction, how one entity produces and manages data can vary due to institutional or personal vagaries.

Consider population censuses. A census consists of a comprehensive survey of an area and its population, usually conducted every ten years. The aim is to establish key information about who is living in a locale and their characteristics (e.g., age, gender, marital status, household composition, religion, race, social class, etc.) and aspects about their lives (their work, accommodation, etc.). To enable change to be measured censuses require continuity with respect to the questions asked and how they are administered. At the same time, in order to capture new data of interest that reflect broader changes in society, transformation is required, such as new or modified questions (see Figure 1.2: note, even when questions were maintained across censuses, how they were phrased was often quite different). Further, how the census is subsequently administered is shaped by institutional, political and economic forces and new technical developments: see Linehan (1991) for a history of the Irish census 1821–1991, and Lauriault (2012) for an analysis of the Canadian census 1871–2011. Moreover, the construction of a census is contested and negotiated as vested interests compete to include, alter or remove questions. In some cases, changes can be quite radical, such as the decision in Germany to discontinue their census in the 1980s (see Hannah 2011). As a consequence, a national census is always caught in a tension between continuity and change, but nonetheless evolves over time and has varying geographies. To date, however, there have been few histories and geographies of data assemblages (though see Alder 2002; Desrosières 1988; Hannah 2011; Hewitt 2010; Lauriault 2012; Poovey 1998; Porter 1995).

Philosophically

For some, at the ontological level data are benign. Data are simply data, essential elements that are abstracted from the world in neutral and objective ways subject to technical constraints. They 'do not have any inherent meaning, do not necessarily

PARTICULARS FOR INDIVDUALS ON FORM A AT EACH CENSUS

- = Not collected; Y = precoded: E = entered by Enumerator; y = reply menu given in notes

Y	Year of Census	1841	1851	1861	1871	1881	1891	1901	1911	1926	1936	1945	1951	1958	1961	1966	1971	1979	1981	1988	1991	Y
M	Month	6	3	4	4	4	4	3	4	4	4	5	4	4	4	4	4	4	4	4	4	M
D	Date	6	30	7	2	3	5	31	2	18	26	12	8	8	9	17	18	1	5	13	21	D
C	Capacity of Form=Number of Persons	20	15	15	12	15	15	15	15	10	10	9	10	11	10	10	10	10	6	8	8	C
1	Name and surname	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	1
2	Sex	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	2
3	Relationship to head of household	X	X	X	X	X	X	X	X	X	X	X	X	-	X	X	X	X	X	X	X	3
4	Age (A); (B) Years, Months; (C) Date of Birth	A	A	A	A	A	A	A	A	B	B	C	C	C	C	C	C	C	C	C	C	4
5	Marital Status (S shows separated)	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	Y	Y	YS	YS	5
6	Place of birth (County, City, Country)	X	X	X	X	X	X	X	X	X	X	X	-	-	X	-	X	-	X	X	X	6
7	Absent family members (separate table)	X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	7
8	Deaths in family members since last Census (separate table)	X	X	X	X	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8
9	Literacy (Read, write); y = notes gave reply menu	y	y	y	y	y	y	y	y	-	-	-	-	-	-	-	-	-	-	-	-	9
10	Occupation: ↑ pre 1926 no separate "industry"	↑X	↑X	↑X	↑X	↑X	↑X	↑X	↑X	X	X	X	X	-	X	X	X	-	X	X	X	10
11	Marriage date "D" or duration "d"; w women only	D	D	D	D	-	-	-	dw	dw	-	Dw	-	-	Dw	-	Dw	-	Dw	-	-	11
12	Irish: *as addendum to Literacy question; y as 9	-	*y	*y	*y	y	y	y	y	y	y	y	-	-	y	-	y	-	y	y	y	12
13	Incapacity (O=deaf, dumb, blind; L = Lunatic)	-	D	D	D	DL	DL	DL	DL	-	-	-	-	-	-	-	-	-	-	-	-	13
14	Religion	-	-	X	X	X	X	X	X	X	X	X	-	-	X	-	X	-	X	-	X	14
15	Live births to present marriage for married women	-	-	-	-	-	-	-	X	X	-	X	-	-	X	-	X	-	X	-	-	15
16	Children of present marriage still living	-	-	-	-	-	-	-	X	-	-	-	-	-	-	-	-	-	-	-	-	16
17	Orphanhood (for children)	-	-	-	-	-	-	-	-	X	X	X	-	-	-	-	-	-	-	-	-	17
18	Dependents = no. <16 for married men and all widowed	-	-	-	-	-	-	-	-	X	-	X	-	-	-	-	-	-	-	-	-	18
19	Industry (Question "Employer and Employers' Business")	-	-	-	-	-	-	-	-	X	X	X	X	-	X	X	X	-	X	X	X	19
20	Area, valn, agric. holdings; H=household; I=individual	-	-	-	-	-	-	-	-	H	H	H	H	-	H	I	H	-	I	-	-	20
21	Period of unemployment previous year; 3 causes	-	-	-	-	-	-	-	-	-	X	X	-	-	-	-	-	-	-	-	-	21
22	Period of residence for immigrants	-	-	-	-	-	-	-	-	-	-	X	-	-	X	-	-	-	-	-	-	22
23	Home address of visitors	-	-	-	-	-	-	-	-	-	-	-	X	X	-	-	-	-	-	-	-	23
24	Employ, status-sep. ques. re employee/own account	-	-	-	-	-	-	-	-	-	-	-	-	-	X	X	X	-	Y	Y	Y	24
25	Subsidiary occupation	-	-	-	-	-	-	-	-	-	-	-	-	-	X	-	-	-	-	-	-	25
26	Age at which fulltime education ceased	-	-	-	-	-	-	-	-	-	-	-	-	-	-	X	X	-	X	X	X	26
27	Types of school etc. attended fulltime: I - duration	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Y	IY	-	IY	-	-	27
28	Scientific or technological qualifications	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	X	-	X	X	X	28
29	Usual residence now	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	X	-	X	X	X	29
30	Usual residence one year ago	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	X	-	X	X	X	30
31	Means of travel to work, school etc.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	X	-	Y	Y	Y	31
32	Distance to work, school etc.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	X	-	X	X	X	32
33	Became resident within past year: Yes/No	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Y	-	-	-	33
34	Address of place of work, school or college	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	X	X	X	34
35	Present status - sep. ques. re labour force status	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Y	Y	Y	35
36	Lived outside > - 1 year; when, whence came	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	X	X	36
37	Highest level of education completed - [for no. 27]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	Y	37
38	Is person farming - principal or subsidiary?	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	y	38

Figure 1.2 Questions concerning individuals on the Irish census 1841–1991

Source: Reproduced from Linehan 1991

present any interpretations or opinions, and do not carry indicative characteristics that may reveal their importance or relevance' (Pérez-Montoro and Díaz Nafría 2010). They are pre-analytical and pre-factual. From this perspective, a sensor has no politics or agenda. It simply measures light or heat or humidity, and so on – producing readings that reflect the truth about the world unless tainted by a technical glitch. In other words, the sensor produces an objective, realist view of the world revealing things as they really are, wherein the reality of thing being measured is independent of the measuring process (Desrosières 1998). Within measurement processes in which people play a central role – in a lab or conducting a survey or interview – a form of mechanical objectivity is employed that adheres to defined rules and rigorous, systematic method to produce distant, detached, impartial and transparent data that is free of researcher bias and preferences, and is thus independent of local customs, culture, knowledge and context (Porter 1995). As such, science when practised properly has no politics or ulterior agenda and data then can be taken at face value. Indeed, the terms commonly used to detail how data are handled suggest benign technical processes: 'collected', 'entered', 'compiled', 'stored', 'processed' and 'mined' (Gitelman and Jackson 2013). It is only the uses of data that are political. In other words, it is people who corrupt data and twist them to their own ends, not science.

For others, such a view is untenable. How we conceive of data, how we measure them, and what we do with them actively frames the nature of data. For them data does not pre-exist their generation; they do not arise from nowhere. Data are produced through measuring, abstracting and generalising techniques that have been conceived to perform a task and are recorded into forms and measures that conform with standards invented by people (e.g., the metric system). They are epistemological units, made to have a representational form that enables epistemological work, and data about the same phenomena can be measured and recorded in numerous ways, each providing a different set of data that can be analysed and interpreted through varying means (Poovey 1998). How data are generated is not inevitable: protocols, organisational processes, measurement scales, categories, and standards are designed, negotiated and debated, and there is a certain messiness to data generation. Take the case of measuring the population of a country: decisions need to be taken as to who is and is not counted (e.g., to include visitors, legal and illegal aliens, those who avoided taking part either deliberately or not, etc.) and where they should be counted (e.g., where they are on census night or where they normally reside); all kinds of rules and procedures are set up, but there is still indeterminacy and variation across enumerators (Porter 1995).

Consequently, how data are ontologically defined and delimited is not a neutral, technical process, but a normative, political, and ethical one that is often contested and has consequences for subsequent analysis, interpretation and action (Bowker and Star 1999). However, once in place, data ontologies work to foreclose variability and define what will be visible and invisible within a dataset, though this process of convergence and stability is always open to resistance and

reworking due to the multiplicity of actors, things and processes at work, and the contrariness of data that do not easily fit within a system (Bowker and Star 1999). Moreover, once data are produced they can sorted, spliced and diced in any number of ways into various categories. In other words, data are not independent of the thought system and the instruments underpinning their production (Bowker and Star 1999). And such thought systems are infused with philosophical assumptions and beliefs and are differentially practised. Indeed, as Borgman (2007: 38) notes, while science seeks to portray itself as universal, 'their practices are local and vary widely'. Thus, data do not simply represent the reality of the world; they are constructions about the world (Desrosières 1998; Poovey 1998).

From such a perspective 'scientific knowledge is produced – rather than innocently "discovered"' (Gitelman and Jackson 2013: 4). As such,

> [d]ata are difficult to separate from the software, equipment, documentation, and knowledge required to use them. For example, if data are produced by an instrument such as a sensor network, interpreting those data requires an understanding of the instrument – for example, what do the sensors detect, under what conditions, at what frequency of observation, and with what type of calibration? (Borgman 2007: 183)

Yet science often tries to shear data free and independent of such a contextual understanding, organising and sharing the data via databases in which the messiness of their creation is ameliorated and users are 'protected' from having to know how the data were produced and organised (Gitelman and Jackson 2013). Ribes and Jackson (2013: 165) thus argue that scientific conceptions of data as neutral and objective are fictions that 'assume or project a world where data floats free of its origins, shedding form, substance, and history, and is thereby rendered free to travel the world as an undifferentiated and universal currency'. In contrast, they assert 'data is stickier than that'.

Consequently, no data are pre-analytic, or objective and independent elements. As Gitelman and Jackson (2013: 2, following Bowker (2005)) put it, 'raw data is an oxymoron'; 'data are always already "cooked" and never entirely "raw"'. 'Data need to be imagined *as* data to exist and function as such' (Gitelman and Jackson 2013: 3). Data are both social, in that they are situated in context, and material, in that they have a form (as bits, as symbols, as numbers, etc.), stored on paper, magnetic tapes, hard disks, etc. (Wilson 2011; Gitelman and Jackson 2013). Both actively shape the constitution of data. For example, big data are reliant on the discursive, political and economic discourses that support their roll-out (see Chapter 7), and on the databases, computers, servers, and networks that enable their production, processing, sharing, analysis and storage (see Chapter 5). Such media facilitate the rotting of data, the misplacing or forgetting or deletion of data, or its erosion through bit-rot (the breakdown of storage media such as the decaying of computer tape and damaged hard drives) (Boellstorff 2013). Indeed, data are never only cooked but are

also open to 'the unplanned, unexpected, and accidental', 'transformed in parahuman, complexly material, and temporally emergent ways that do not always follow a pre-ordained, algorithmic "recipe"' (Boellstorff 2013).

Given the social and material nature of data we actively remake 'our material, technological, geographical, organizational, and social worlds into the kind of environments in which data can flourish ... enter[ing] into a symbiotic relationship with data' (Ribes and Jackson 2013: 152). Thus, while '[d]ata is seen as something that is out there – something that is *about* the real' (Shah 2013, original emphasis), they are more productively understood as both a component of the real and a producer of the real. Data are not merely an abstraction and representative, they are constitutive, and their generation, analysis and interpretation has consequences. As Gitelman and Jackson (2013: 2) state: 'if data are somehow subject to us, we are also subject to data'. Data are captured from the world, but in turn do work in the world. Data are not, and can never be, benign (Shah 2013). Instead, '[d]ata ... need to be understood as framed and framing' (Gitelman and Jackson 2013: 5). In other words, there is much more to conceptualising data than science and business generally acknowledge.

THINKING CRITICALLY ABOUT DATABASES AND DATA INFRASTRUCTURES

In order to make sense of data they are usually pooled into datasets, which are often organised and held in databases (a means of structuring and storing data that can be queried in multiple ways) and data infrastructures such as archives and repositories (see Chapters 2, 3 and 5). As with conceptualising data more generally, it is then important to think critically about the nature of databases and data infrastructures, their sociotechnical production, and how they reflect rationalities about the world at the same time as they reproduce and reinforce such rationalities. Such critical reflection has been largely absent with respect to big data, open data, and the scaling of small data, with the focus to date being more technical and instrumental in nature.

The thesis adopted and developed throughout this book continues the argument set out in the last section, positing that databases and data infrastructures are not simply neutral, technical means of assembling and sharing data; they are not merely products that store captured data about the world, but are bundles of contingent and relational processes that do work in the world (Star and Ruhleder 1996; Kitchin and Dodge 2011). They are complex sociotechnical systems that are embedded within a larger institutional landscape of researchers, institutions and corporations, constituting essential tools in the production of knowledge, governance and capital.

Databases are designed and built to hold certain kinds of data and enable certain kinds of analysis, and how they are structured has profound consequences as

to what queries and analysis can be performed; their formulation conditions the work that can be done on and through them (Ruppert 2012). For example, it is quite different to store data in databases rather than in a narrative form in terms of what is stored, how it is stored, and how it can be recalled and reworked (Bowker 2005). Databases create silences, adding to the inherent gaps in the data, as much as they reveal relationships between data and enable some questions to be answered; they constrain and facilitate through their ontology, producing various presences and absences of relations (Ruppert 2012; Vis 2013). Drawing on Derrida, Bowker (2005: 12) thus argues that databases and archives are jussive: they 'tell us what we can and cannot say' by defining what is remembered and what is ignored and forgotten. Such remembering/forgetting is determined by rules and practices that are political and philosophical acts. The ontologies within databases are thus neither fixed nor natural, but are created by actors with particular aims working within communities of practice, modes of governance, and technical constraints. Databases then are expressions of power/knowledge and they enact and reproduce such relations (Ruppert 2012), for example determining what someone's insurance rate is or whether they can travel between countries. Moreover, databases are dynamic entities that perform a 'constant process of differentiating' (Ruppert 2012: 129) through interactions with their associated assemblage (creators, users, software, hardware, networks, etc.).

At the same time, databases unmoor data analysis from the data by enabling complex queries and calculations without those conducting such analyses having to peruse and work the data themselves or even understand how the data have been compiled and organised (Gitelman and Jackson 2013). This unmooring is aided by techniques such as standardisation of formats and metadata and works to decontextualise and depoliticise the data contained within (Wilson 2011). Importantly, such unmooring enables the power/knowledge of the database to travel and be deployed by others shorn of its complex inner workings and history and politics of production (in the same way as a driver can utilise a car without knowing how all its complex systems are made or what they do or how they interact to shape the driving experience). Researchers can thus utilise government databases such as a census or business survey or economic indicators without knowing the politics of why and how such databases were constructed, the technical aspects of their generation, or having personal familiarity with the phenomena captured. For example, using the Irish Department of Environment's databases of unfinished estates in Ireland post the 2008 economic crash (available at http://www.housing.ie/Our-Services/Unfinished-Housing-Developments.aspx) one can interrogate, map and draw conclusions about the estates without knowing anything about the history and politics of the survey, how it was undertaken, or visiting any of estates (see Kitchin et al., 2012a, b). Such unmooring then enables databases to act as immutable mobiles (that is, stable and transferable forms of knowledge that are portable across space and time) (Latour 1989).

Data infrastructures host and link databases into a more complex sociotechnical structure. As with databases, there is nothing inherent or given about how such archiving and sharing structures are composed. Indeed, as discussed throughout the book, the design and management of data infrastructures are riddled with technical and political challenges that are tackled through messy and contested negotiations that are contextualised by various agendas and governmentalities. The solutions created in terms of standards, protocols, policies and laws inherently have normalising effects in that they seek common shared ground and to universalise practices amongst developers and users (Lauriault 2012), glossing over and ameliorating the tension between enabling interoperability and limiting customisation and constraining innovation, and denying alternative ways of structuring and ordering data (Star and Ruhleder 1996). Given these tensions, normalising processes have to constantly and recursively be reaffirmed through implementation, management and system governance (Star and Lampland, 2009). Star and Ruhleder (1996: 112) thus contend '[t]here is no absolute center from which control and standards flow; as well, no absolute periphery', with 'infrastructure [being] something that emerges for people in practice, connected to activities and structures'.

This emergence, while never fully centred is, however, not free-form and is shaped by wider structural relations. As Graham and Marvin (2001) argue, infrastructures are constitutive of 'long-term accumulations of finance, technology, know-how, and organizational and geopolitical power' (p. 12) and sustain 'sociotechnical geometries of power' (p. 11) of congealed social interests. Such accumulations include regimes of regulation that seek to delimit legally and through forms of governmentality how data are managed, analysed and shared, for example data protection laws (see Chapter 10). Starr (1987: 8) thus proposes that a data infrastructure has

> two kinds of structures – social and cognitive: Its social organization consists of the social and economic relations of individual respondents, state agencies, private firms, professions, international organizations and others involved in producing flows of data from original sources to points of analysis, distribution and use. Cognitive organization refers to the structuring of the information itself, including the boundaries of inquiry, presupposition about social reality, systems of classification, methods of measurement, and official rules for interpreting and presenting data.

As Dourish and Bell (2007) contend, databases and infrastructures then cannot be considered in purely instrumental terms as they are thoroughly cultural, economic and cognitive in nature and steeped in social significance. They thus suggest two lenses through which to understand data infrastructures. The first is a sociopolitical reading which examines them as 'crystallizations of institutional relations' (p. 416). The second perspective is an experiential reading that examines 'how they shape individual actions and experience' (p. 417). In both cases, data infrastructures are understood as relational entities. This relationality reshapes the world contingently

around it, as it in turn is shaped by the world. So as we come to use and rely on databases and data infrastructures to make sense of and do work in the world, our discursive and material practices adapt and mutate in response to them (Star and Ruhleder 1996). The world is not just reflected in data, it is changed by them; 'the work of producing, preserving, and sharing data reshapes the organizational, technological, and cultural worlds around them' (Ribes and Jackson 2013: 147).

In other words, databases and data infrastructures do not simply support research, they fundamentally change the practices and organisation of research – the questions asked, how they are asked, how they are answered, how the answers are deployed, who is conducting the research and how they operate as researchers (see Chapter 8). For example, in her study of the evolution of the Canadian Census and the Atlas of Canada, Lauriault (2012) details how each has developed recursively and iteratively based on models of the world which construct ways to imagine and produce Canada. She argues that the data archives and the data themselves constitute an institutional 'extrasomatic memory system that allows for the telling of stories about the nature of Canada ... [through] maps, graphs, models and statistics which rely on sensors, data, interoperability and web mapping standards, portals, metadata and models, science, and open architectures' (p. 27). In turn, these stories modulate the underlying models and thus the data infrastructure mutates, inflecting the means through which the stories are created.

Making sense of databases and data infrastructures then requires carefully unpacking and deconstructing their always emerging, contingent, relational and contextual nature (Star and Ruhleder 1996). This means looking for what Bowker and Star (1999: 34) describe as infrastructural inversion that recognises 'the depths of interdependence of technical networks and standards, on the one hand, and the real work of politics and knowledge production on the other'. As Lauriault (2012) argues, this also requires a historical analysis that documents how databases and data infrastructures develop over time and space.

DATA ASSEMBLAGES AND THE DATA REVOLUTION

The principal argument forwarded in this chapter has been that thinking about data is not straightforward. Data do not exist independently of ideas, techniques, technologies, systems, people and contexts, regardless of them often being presented in this manner (Lauriault 2012; Ribes and Jackson 2013). Data are generated as the product of many minds working within diverse situations, framed and shaped within mileau circumstances and structures.

One way to make sense of data is to think of them as the central concern of a complex sociotechnical assemblage. This data assemblage is composed of many apparatuses and elements that are thoroughly entwined, and develop and mutate over time and space (see Table 1.3). Each apparatus and their elements frame what is possible, desirable and expected of data. Moreover, they interact with and shape each other

Table 1.3 The apparatus and elements of a data assemblage

Apparatus	Elements
Systems of thought	Modes of thinking, philosophies, theories, models, ideologies, rationalities, etc.
Forms of knowledge	Research texts, manuals, magazines, websites, experience, word of mouth, chat forums, etc.
Finance	Business models, investment, venture capital, grants, philanthropy, profit, etc.
Political economy	Policy, tax regimes, public and political opinion, ethical considerations, etc.
Governmentalities and legalities	Data standards, file formats, system requirements, protocols, regulations, laws, licensing, intellectual property regimes, etc.
Materialities and infrastructures	Paper/pens, computers, digital devices, sensors, scanners, databases, networks, servers, etc.
Practices	Techniques, ways of doing, learned behaviours, scientific conventions, etc.
Organisations and institutions	Archives, corporations, consultants, manufacturers, retailers, government agencies, universities, conferences, clubs and societies, committees and boards, communities of practice, etc.
Subjectivities and communities	Of data producers, curators, managers, analysts, scientists, politicians, users, citizens, etc.
Places	Labs, offices, field sites, data centres, server farms, business parks, etc., and their agglomerations.
Marketplace	For data, its derivatives (e.g., text, tables, graphs, maps), analysts, analytic software, interpretations, etc.

through a contingent and complex web of multifaceted relations (see Figure 1.3). And, as Ribes and Jackson (2013) contend, not only do they frame what and how data are produced and to what ends they are employed, but they are themselves organised and managed to produce such data. Data and their assemblage are thus co-determinous and mutually constituted, bound together in a set of contingent, relational and contextual discursive and material practices and relations. Every data assemblage then varies in concert with the arrangement of elements and context, but they share commonalities and echoes of each other due to overarching and overlapping apparatus and conventions that span assemblages. And as new ideas and knowledges emerge, technologies are invented, skill sets develop, and markets open, data assemblages evolve, mutate, coalesce and collapse. As a consequence, there is a huge diversity of data assemblages across domains and jurisdictions.

This book examines the emerging and evolving data assemblages producing open data, data infrastructures and big data. In so doing it advances three key arguments. First, there is a need to develop conceptual and philosophical ways to make sense of data. There has been remarkably little critical reflection and research on data in and of themselves and on the constitution and operation of the

assemblages surrounding them, especially compared to the attention focused on the concepts of information and knowledge. And yet, data are a fundamental element of knowledge production. Second, there is a data revolution underway that constitutes a key moment in evolution and mutation of data assemblages. Due to the confluence of several emerging developments in computing, methodological techniques, and the political and economic realm, the volume, variety, velocity, resolution, and availability of data, and how data are being processed, analysed, stored, and employed to leverage insight and value, is being radically transformed. Third, given the various technical, ethical and scientific challenges that the data revolution raises there is an urgent need to develop a detailed understanding of the new and emerging data assemblages being created. The ten chapters that follow thus aim to provide a broad, synoptic and critical overview of these assemblages and to highlight issues that demand further attention and research.

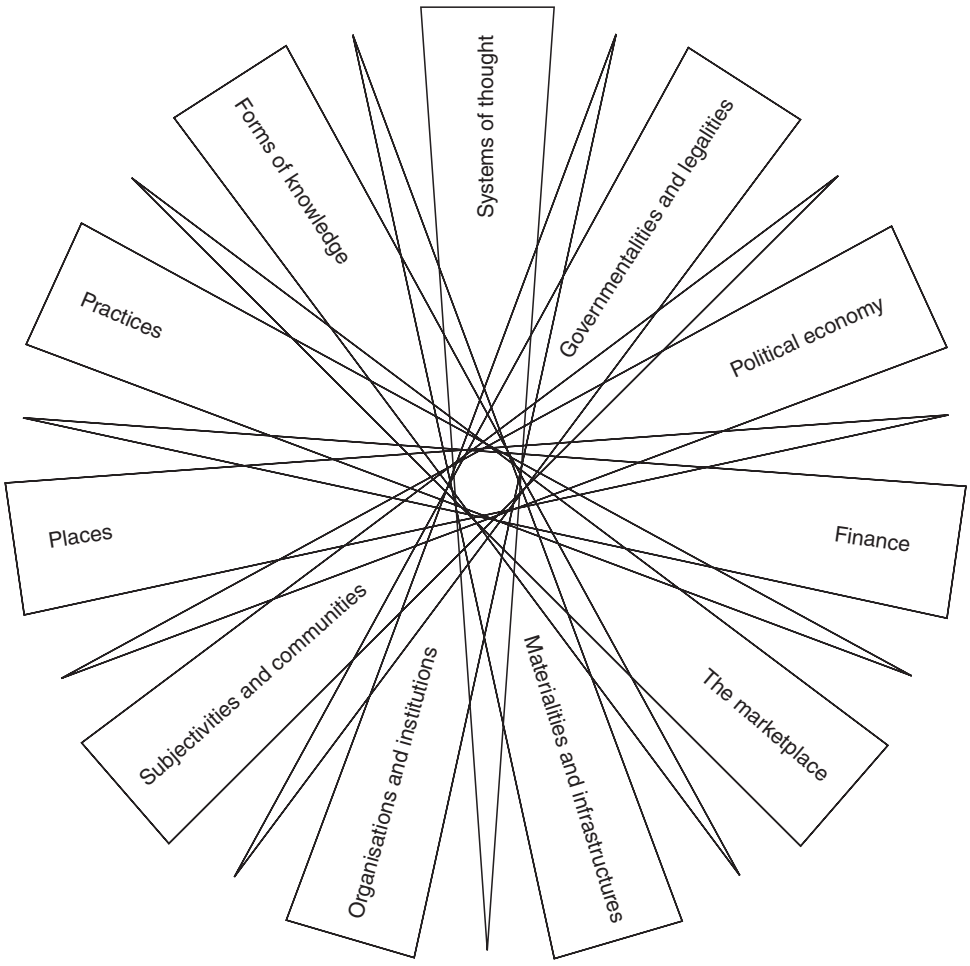

Figure 1.3 The intersecting apparatus of a data assemblage