Journal of Strategic Information Systems 28 (2019) 3–16

Contents lists available at ScienceDirect

Journal of Strategic Information Systems

journal homepage: www.elsevier.com/locate/jsis

What we talk about when we talk about (big) data

Matthew Jones

Judge Business School, University of Cambridge, Trumpington Street, Cambridge CB2 1AG, UK

A B S T R A C T

In common with much contemporary discourse around big data, recent discussion of datafication in the *Journal of Strategic Information Systems* has focused on its effects on individuals, organisations and society. Generally missing from such analysis, however, is any consideration of data themselves. What is it that is having these effects? In this Viewpoint article I therefore present a critical analysis of a number of widely-held assumptions about data in general and big data in particular. Rather than being a referential, natural, foundational, objective and equal representation of the world, it will be argued, data are partial and contingent and are brought into being through situated practices of conceptualization, recording and use. Big data are also not as revolutionary voluminous, universal or exhaustive as they are often presented. Some initial implications of this reconceptualization of data are explored. A distinction is made between "data in principle" as they are recorded, and the "data in practice" as they are used. It is only the latter, typically a small and not necessarily representative subset of the former, that will contribute directly to the effects of datafication.

1. Introduction

Data and their effects on individuals, organisations, business models and society have, rightly, attracted growing attention In the *Journal of Strategic Information Systems* (Newell and Marabelli, 2015; Loebbecke and Picot, 2015; Gunther et al., 2017, Markus, 2017). The immediate prompt for this attention has been the "widespread diffusion of digital devices that have the ability to monitor our everyday lives" (Newell and Marabelli, 2015: 3), a process that is referred to as "datafication". Discussions of this phenomenon, however, have largely taken the data themselves for granted and have focused on how data "are being used, and by whom and with what consequences" (Newell and Marabelli, 2015:3). While, as Galliers et al. (2017) argue, the uses of data raise important questions that deserve the attention of scholars in the IS field (and more widely) in this Viewpoint paper I would like to switch the focus around and consider what constitutes these data, the effects of the accumulation of which we have begun to explore. What actually is it that is having these effects?

This enquiry will critically examine a number of commonly held, and often implicit, assumptions about the nature of data. In doing so I hope to extend the discussion of the datafication phenomenon beyond "its issues, impacts and implications" (Galliers et al., 2017: 188) to include an awareness of the particular character of the 'material' on which this phenomenon is based. A better appreciation of this character, it will be argued, may inform a richer understanding of the effects of datafication and open them to more rigorous scrutiny.

What has given questions about the nature of data a particular relevance, of course, is not just the increasing datafication of contemporary life. Rather it is the accumulation of these data in repositories, the analysis of which, often by "pre-determined algorithms that lead to decisions that follow on directly without further human intervention" (Galliers et al., 2017: 185), is seen as transforming work, organisations and society, a development commonly referred to as "big data".

Although, as will be discussed, "big data" are not necessarily a product of datafication and the meaning of the term itself is highly contested, there is little doubt both that the volume of data being created has greatly expanded in recent years and that techniques to analyse data at this scale have significantly advanced. The paper will therefore also examine assumptions that relate specifically to

E-mail address: m.jones@jbs.cam.ac.uk.

https://doi.org/10.1016/j.jsis.2018.10.005
Received 10 August 2018; Received in revised form 15 October 2018; Accepted 23 October 2018
Available online 30 November 2018

this accumulated data, not just data themselves.

2. Assumptions about data

Discussions of data are bedevilled by inconsistencies in the way in with which the term is defined in the literature. A Delphi study of Information Scientists by Zins (2007), for example yielded more than 40 different definitions, while Checkland and Holwell (1998) list seven different definitions from IS textbooks. Although it is certainly beyond the scope of this paper to propose a definitive analysis of these definitions, it would nevertheless seem important to clarify some of the main characteristics that are seen to be associated with data to draw out the assumptions on which they are based. Many of these definitions, however, themselves employ contested terms that carry with them their own implications, so it is hard to find a starting point for this analysis that would be recognised by all parties. The discussion below therefore proceeds from the definition of Davenport and Prusak (1997) on the grounds that it is simple and widely-cited.

For Davenport and Prusak (1997: 9) data are "simple observations about states of the world". A similar view is offered by Ackoff (1989:3) who describes data as "symbols that represent properties of objects and events and their environments", and the Royal Society (2012: 12) who define data as "numbers, characters or images that designate an attribute of a phenomenon". Common to these definitions would seem to be an assumption that data are referential, they stand for something else (states of the world, objects, entities) that pre-date and exist independently of them.

The referentiality of data does not necessarily imply, however, that they are (statistically) representative of the particular states of the world they claim to describe. The observations, symbols or collected attributes are a particular subset rather than the universe of all possible observations, symbols or attributes. This continues to be the case with datafication, even if the range of potential phenomena about which it may be possible to produce data is expanded. Nor does it mean that all data are about actual states of the world. Synthetic data, that have similar properties to sensitive or regulated data may be artificially manufactured (rather than generated by real-world events), for example, to validate or train machine learning models. The utility of such data, however, normally depends on the assumption that they correspond, to a sufficient degree, to the state of the world they are considered as representing, so they are still referential.

A second assumption about data does not necessarily follow from referentiality, but is often associated with it. This is that data themselves are natural, in the sense, as the *Oxford English Dictionary* puts it, of having a real or physical existence. Checkland and Holwell (1998), for example, point out that the term data comes from the Latin *dare*, to give (a sense that is retained in the French term for data, *données*). This may be seen to imply that data have some sort of external existence "out-there" in the world, that precedes any use that may be made of them. Such a view may also be inferred from terms used to describe activities associated with data such as "mining" or "curation".

If data exist independently of their use, however, then only that proportion of data to which attention is paid will ever be in a position to influence human action. Checkland and Holwell (1998) propose the term capta (from the Latin *capere*, to take) for this subset. Capta are therefore the data that at least have the potential to be used and we should be wary of assuming that they are necessarily representative of the data as a whole. Rarely, however, is this distinction observed and capta are often treated as if they were the totality of data (and indeed are often referred to as "the data").

Both the referential view of data and the view that data are natural, moreover, share the assumption that what we take to be data have some direct relationship with state of the world. This is often associated with a third assumption that sees data as foundational. Kitchin (2014a:9), for example, describes data as the "base or bedrock of a knowledge pyramid" and presents the following figure by way of illustration (see Fig. 1).

The pyramid here is a reference to the Data/Information/Knowledge/Wisdom (DIKW) hierarchy which (Rowley, 2007: 163) describes as "one of the [most] fundamental, widely recognized and 'taken-for-granted' models in the information and knowledge literatures". The hierarchy is usually attributed to Ackoff (1989:3) who wrote "Wisdom is located at the top of a hierarchy …

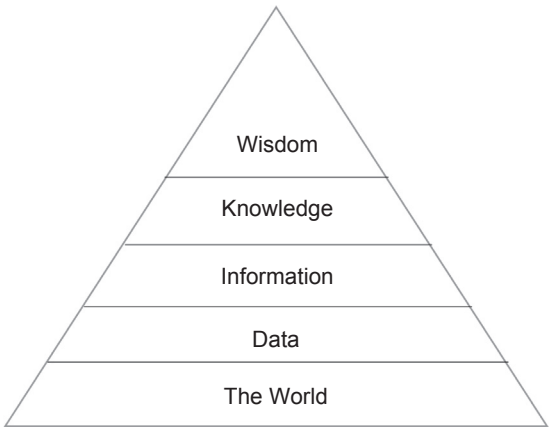

Fig. 1. The Knowledge Pyramid (after Kitchin, 2014a).

Descending from wisdom there are understanding, knowledge information and, at the bottom, data. Each of these includes the categories that fall below it". Note too that it is "the World" on which data directly rests.

A fourth, and generally more contentious, assumption about data relates to their objectivity. This is sometimes presented as a corollary of the distinctions involved in the DIKW hierarchy in which information is viewed as inferred from data. Drucker (1998), for example, defines information as "data endowed with relevance and purpose", while Davenport and Prusak (1997: 4) argue that "data describes only a part of what happened; it provides no judgment or interpretation". Data themselves therefore simply report reality. This view is supported in many IS textbooks that define data as "raw facts that describe a particular phenomenon" (Haag and Cummings, 2013: 508), "a series of facts that have been obtained by observation or research and recorded" (Bocij et al., 2006: 794), or "raw facts that can be processed into accurate and relevant information" (Turban et al., 2006: G-3).

Because data are just these objective facts, neutral atoms that represent the world, a fifth assumption is that all data are equal. In principle, there is nothing to differentiate between them. They may be of different forms (numbers, words, images, video), but, given that from an Information Technology (IT) perspective these are all just bitstrings, they are equal in the eyes of the processor. Indeed part of the appeal of datafication is that it renders more and more of the world accessible to analysis using common techniques. Whether the data represent stock price movements, Tweets or x-rays, they can be analysed in the same way.

In summary, therefore, data are widely assumed to be: referential, they describe a world independent of themselves; foundational, they are the base on which our understanding of the world is built and themselves stand directly on that world; natural, in that they exist independent of their use; objective, in that they represent the world, without interpretation; and equal, as all data can be analysed with the same techniques. While, as will be discussed, these assumptions are not universally shared, they are found, often implicitly, in much discussion of data. With the advent of big data a number of further assumptions have been added.

3. Assumptions about big data

The starting point for much discussion of big data is typically a reference to the increasing volume and velocity of data "flowing into every area of the global economy" (Manyika et al., 2011). This is often illustrated by the quoting of very large numbers with exa, peta and tera prefixes describing how many Facebook posts, or Google searches are undertaken every minute, or comparing the volume of data produced between the dawn of civilisation and the early 2000s and the amount of data now created every day. Big data, these figures suggest, mark a new era, where old rules and understandings cannot be expected to apply. As the subtitle of Anderson's influential Wired article put it "More isn't just more, more is different" (Anderson, 2008a). Big data are revolutionary, they transform how we know the world.

Beyond the shock and awe invoked by these numbers, which has contributed perhaps to the lack of theoretical reflection on big data, they indicate a common fixation on the quantity, rather than the quality, of data. This is somewhat ameliorated by the "third V" of the big data mantra (Laney, 2001) – variety, although, as Kitchin and McArdle (2016) argue, this is not necessarily a defining characteristic of all datasets that are considered as "big". Just from the choice of epithet to refer to the phenomenon, however, it would seem that size is seen as a significant distinguishing feature of big data and some authors even enshrine this in their definition of the term. Manyika et al. (2011), for example, state that big data "refers to datasets that are beyond the ability of typical database software tools to capture, store manage and analyse". Of course what counts as typical is open to debate and the ability of database software to handle such data may not be due solely to their size, but in technical domains, such as genomics or astrophysics, in reports by IT firms such as SAS or IBM, and in online definitions, volume is generally the first, and sometimes the only, descriptor associated with the term.

Another characteristic associated with big data that is highlighted by commentators is its incipient universality. As Cukier and Mayer-Schoenberger (2013: 29) argue "it is tempting to understand big data solely in terms of size. But that would be misleading. Big data is also characterized by the ability to render into data many aspects of the world that have never been quantified before." This may be seen as a specifically quantitative twist on the more general 'informating' capacity of IT (Zuboff, 1988) that textualises previously ephemeral social and organizational practices, rendering them visible to control (and accessible to analytics). This datafication may be deliberate, for example when a person chooses to wear an activity tracker to record their movement throughout the day, or may be an unrecognized byproduct of their web browsing. Nor is datafication restricted to human activity. With the Internet of Things our environment is being datafied too. Consequently, it is suggested, everything is, or potentially will become, data and data will be everywhere.

This prospect of the datafication of everything leads to a final assumption, that big data are exhaustive. "Once the world has been datafied" Mayer-Schoenberger and Cukier (2013: 96) argue, we have a situation where N = all. Data and the world will be coextensive and wholly available for analysis. This, they argue, will bring about a transformation in the way we understand and explore the world: a shift from the traditional reliance on small samples to a situation where we can analyse the entirety of data about a topic; a shift from the pursuit of exactitude to a toleration of the messiness of real-world data; and a shift from the search for causal explanations to a greater reliance on correlation. Whether or not this marks the "end of theory" (Anderson, 2008b) we are reaching the point, it is argued, where we have enough data that, as Anderson (2008b) puts it, "the numbers speak for themselves".

Big data therefore adds to the assumptions about data in general, presenting them as: revolutionary, an unprecedented phenomenon that will transform our understanding of the world; voluminous, accumulating at large scale and at a high rate; universal, data are everywhere and represent everything; and exhaustive, all these data form a vast, comprehensive and accessible resource that provides a complete picture of the world. As was stressed earlier, this is not intended to be a definitive characterisation of (big) data, rather it gives us a set of traits that are commonly associated with these concepts and that may therefore provide a starting point for a critical examination of the nature of data. These assumptions and some of the questions (discussed below) that may be raised about

Table 1
Assumptions about data and big data and potential challenges to them.

Assumption	Description	Potential challenges to the assumption
Data are …		
Referential	Data report a reality that exists independent of themselves	Data may not always report reality accurately or completely
Natural	Data exist independent of their use	Data are constructed
Foundational	Date are the base on which our understanding of the world is built and themselves stand directly on that world	What is taken to be data reflects a particular way of understanding the world
Objective	Data represent the world, without interpretation	What is taken to be data involves subjective judgements
Equal	All data are equivalent	Data vary in their characteristics, provenance and quality
Big data are …		
Revolutionary	Big data are unprecedented and transformational	Big data are not necessarily new and their effects are not determined by their characteristics
Voluminous	Data are accumulating at large scale and at a high rate	The significance of data does not depend on their volume and velocity
Universal	Everything is, or will become, data	Not everything is effectively represented by data. Datafication is selective
Exhaustive	N = all	Data are a sample of the reality they describe. Data that are usable, and used, are a subset of data that are recorded

them are summarised in Table 1.

It should be noted that there is considerable debate about whether big data mark the emergence of a wholly new paradigm (see Kitchin (2014b) and Frické (2015) for example on the "end of theory" argument of Anderson (2008b) and Mayer-Schoenberger and Cukier (2013)) and therefore whether we might expect the assumptions about data in general to apply also to big data. It is not the aim of this paper to settle this question and for present purposes the two sets of assumptions will be treated separately. In keeping with the critical perspective adopted, however, the starting point will be that, in the absence of evidence (as opposed to claims), assumptions about data in general may be expected to apply to big data. That is, whatever may be different about big data, at base they are data (and therefore subject to the same assumptions). This does not preclude the possibility that intensification may be associated with new effects, but these add to, rather than replace the earlier ones. Referentiality still applies to big data, for example, in that they are taken as representing states of the world that exist prior to, and independent of, them. A retailer would not be very interested in customer loyalty data if they did not believe they reported actual activity in their business. The argument of the paper does not depend on this continuity, though, so if future work were to suggest that big data should be treated differently this should not invalidate the assumptions at either end of the scale.

4. Questioning data

To provide a common reference point for the examination of data, the discussion will draw on examples from ongoing research on the implementation and use of electronic medical records in acute hospitals, and particularly in critical care. Electronic medical records are also widely considered to be prime targets for "big data" initiatives (Groves et al., 2013; Raghupathi and Raghupathi, 2014). This is not to suggest that these examples will be representative of all data, but that they highlight issues that may apply in other settings too.

4.1. What do data refer to?

The assumption that data are referential, reporting a reality that exists independently of them, depends on the accuracy and completeness of the recording process and on the security and reliability of the data storage environment. There may be various reasons, however, why what is recorded as data may not always correspond to the state of the world that has been observed. These include, errors or failures in the observation process, communication breakdowns between what is observed and what is recorded; deliberate or accidental changes to the record; or failure or corruption of the record. The potential for such errors may be more evident when the observation or recording process relies on human intervention, but is not eliminated when processes are automated.

In a medical setting data may be recorded by clinical staff, for example prescribing a drug or recording that a dose has been administered, or recorded automatically from the output of sensors placed on, or sometimes in, the patient's body, or from the output of devices supporting a patient's vital functions, such as a respirator. Although clinical information systems are designed to avoid drug prescription and administration errors, busy clinical staff can still make mistakes. Similarly while medical devices are designed for high reliability and accuracy, they can sometimes malfunction or sensors can become misplaced, and cables can sometimes break or become disconnected. As a result medical records can contain significant levels of erroneous and incomplete data (Rahman and Reddy, 2015). Even if these may not be consequential for clinical care, they mean that the data recorded do not accurately report the patient's condition.

This is not to say that most data do not report an independent reality, but that we cannot necessarily know from the data themselves how accurately they represent the state of the world they are taken as describing. This question is magnified the further we get from the original source of the data. Errors may be evident to a person who has observed the conditions that the data are intended to represent, but once these data are bundled up into datasets that are analysed by people who have no other point of

reference, then they become hard to detect. Data cleaning may catch some of the more obvious errors, but may miss more subtle ones and may potentially also reduce the fidelity of the data to the phenomena they are considered as reporting, for example if data are genuinely anomalous.

4.2. Are data natural?

It might be considered self-evident that data are givens that precede their use. After all, we have vast and ever-growing quantities of data that are available for analysis, the existence of which is not dependent on the uses to which they may be put. This is to assume, however, that all phenomena naturally manifest themselves as data and that this manifestation is independent of the use to be made of them. What is it though, that we take to be data?

Let us consider data on a patient's blood pressure in a medical setting. The pressure itself is not what is held in a database, but rather some particular representation of it. In the case of blood pressure this is the values measured by some device at particular instants in time. As Gitelman and Jackson (2013: 3) write, "like events imagined and enunciated against the continuity of time, data are imagined and enunciated against the seamlessness of phenomena."

At least in Western medicine, these values are expressed in units (millimetres of mercury) that correspond to the readings of a mercury sphygmomanometer, even though the readings are likely to have been taken with a digital or aneroid device. What is recorded as blood pressure data are therefore sampled values, created in conformance with a particular convention. This does not make them any less effective in patient care, but does highlight their constructed character. "Data does not just exist – it has to be generated", as Manovich (2012: 224) puts it.

The choice of what values are recorded as a patient's blood pressure and how often these are recorded, moreover, is likely to be shaped by the intended use of the data. Thus intensive care patients may have several blood pressure values recorded in different locations in the venous system and pulmonary vessels. Values may also be recorded every minute in intensive care, whereas a single measurement may be sufficient for an otherwise healthy patient. Observing or representing states of the world therefore involves decisions about what and how they should be recorded as data.

4.3. Are data foundational?

The DIKW hierarchy presents data as a bedrock on which information, knowledge and wisdom are based. Tuomi (1999), however, makes the case that the hierarchy should in fact be reversed. That is, "Data emerge last—only after knowledge and information are available. There are no "isolated pieces of simple facts" unless someone has created them using his or her knowledge" (Tuomi, 1999: 107). Drawing on Fleck's analysis of historical evolution of the definition of syphilis (Fleck, 2012), Tuomi proposes that, rather than being the raw material of information, data are created from information and that this information is itself shaped by a body of knowledge held by a particular thought community.

Using the example of temperature measurement, Tuomi argues that it is the knowledge informing the design of thermometers that defines what temperature is. Furthermore, to use data we draw on knowledge of what those values should be and what variation from them is significant. In a medical setting there are various different ways in which a patient's temperature may be measured that vary in their accuracy and their equivalence to core body temperature (the temperature of blood at the hypothalamus, the organ that controls the body's response to temperature). The patient's condition will also influence how frequently their temperature should be recorded. What is taken to be data on a patient's temperature is therefore not a universal formula, but varies depending on a range of factors that are informed by a body of knowledge about the performance of different methods of measurement and patients' sensitivity and susceptibility to temperature change. As Gitelman and Jackson (2013: 2) argue, therefore, "data are always already "cooked" and never entirely "raw"".

4.4. Are data objective?

If data are never raw, then our conception of their objectivity is already in trouble, but the common description, of data as facts (raw or otherwise) also deserves critical scrutiny. It is the givenness of data, Frické (2009) argues, that leads to the perception of them as facts, based on two features: truth and certainty. Fallibilist arguments, of writers such as Peirce and Popper, that all knowledge is conjectural and certain knowledge is unattainable, may have placed the latter feature in doubt, but the truth of data remains a necessary assumption of mainstream views (and of the DIKW hierarchy), as the earlier definitions from the IS literature demonstrate. Nor is this view restricted to positivist authors. Checkland and Holwell (1998) also treat data and facts as interchangeable.

At least three objections may be raised to the equation of data with facts. The first is pragmatic. Thus we may commonly talk of data as "wrong", "incorrect" or "unreliable". This being so, it is evidently not a necessary characteristic of data that they are true. Rather as Rosenberg (2013: 18) argues, "when a fact is proven false, it ceases to be a fact. False data is data nonetheless." The second objection is that data are not limited to what is observable (whether by humans or machines). Data may be non-empirical (Frické, 2009), so cannot be judged on their truth. Nor can we assume that even where data accurately report an empirical phenomenon there will be consensus on the truth of what they describe, as recent controversies over "fake news" illustrate.

All three issues may be present with medical data. What is recorded as data may not correspond to the state of the world that was observed, for example where a clinician makes a mistake in entering data, or where a faulty sensor reports erroneous values. There may be subjective components of the record, for example describing a patient's experience of pain. While it may be hoped that medical records do not contain fake data, certain data may be artificial. For example a patient's heart rate may be set by a pacemaker,

or their blood pressure may be controlled by medication. The values recorded may be accurate, but they are mediated by other factors, the influence of which may not be precisely known.

4.5. Are data equal?

That data are not logical atoms that truthfully and directly report an external reality raises the possibility that some data may be better than others. While it may be the case computationally that all data are equal, in practise there are a range of dimensions on which data may be differentiated, as Kitchin (2014a) discusses. Thus data may be: quantitative or qualitative; structured, semi-structured or unstructured; captured, derived, exhaust or transient; primary, secondary or tertiary; or indexical, attribute or meta-data. Variation on these dimensions may be expected to affect the perceived value of the data, although this evaluation may vary across different fields. For example captured, primary quantitative data may be considered the gold standard in some settings, but unstructured qualitative data may be highly valued in others.

The provenance of data may also affect their perceived value. In evidence-based medicine, for example, various hierarchies assessing the strength of different methods as sources of evidence have been proposed. These generally place randomised controlled trials at the top, with expert opinion and anecdotal experience at the bottom and observational studies, cross-sectional surveys and cohort studies somewhere in between. More informally, the interpretation of health record data may be influenced by knowledge of who recorded them.

Data also vary in their quality. Reviewing literature on data quality in Electronic Health Records, for example, Weiskopf and Weng (2013) identify five dimensions of quality: completeness, are values recorded for all data elements; correctness, are the values recorded true; concordance, are the values consistent with each other; plausibility, are values in agreement with general medical knowledge; and currency, are values recorded in a reasonable period of time following measurement. While different application domains may have other quality criteria, these dimensions indicate that the recording of data per se does not make them equivalent. As boyd and Crawford (2012: 671) write “Data are not generic. There is value to analyzing data abstractions, yet retaining context remains critical, particularly for certain lines of inquiry.”

Many of the assumptions prevalent in discussions of data, in general, would therefore seem debatable. Many of these assumptions too, would appear to carry over to discussions of big data. Indeed, with all the hype around the concept they are possibly even less likely to receive critical scrutiny. Big data also bring additional assumptions that merit attention.

4.6. Are big data revolutionary?

The argument that big data will have a transformative effect on individuals, organisations and society amounts to a claim that the existence of data at scale per se necessitates a particular response. This is to neglect, however, the reasons for the accumulation of data and the socioeconomic conditions of their creation and use. Why is it that organisations or governments choose to develop the systems that create and analyse large volumes of data? What are the circumstances under which individuals contribute to these datasets? This is not to suggest any necessarily malign intention of either corporations or governments in accumulating data, or that individuals are necessarily dupes in this process, but rather to point out that, from an IS perspective, we might be expected to be cautious of mono-causal accounts of organisational and social change that exclude any consideration of human agency. The use of large datasets may well be implicated in significant social and organisational change, but they are neither an autonomous creation nor are they the independent cause of these changes.

Discussions of big data also frequently treat data as a free resource, the effects of which flow directly and unproblematically from their existence. What this neglects, however, are the costs of producing, storing, retrieving and using data which may mean that these effects are significantly attenuated. In the medical context, for example, producing data of sufficient completeness, correctness, concordance, plausibility and currency (Weiskopf and Weng, 2013) that they may be reliably reused often requires extra work, such as validating data entry or collecting additional data that may not be necessary for immediate patient care. Data entry may also be more time consuming as systems enforce completion of all fields and keyboards replace handwriting. It can be challenging to persuade busy clinicians to undertake this work when it offers no immediate benefits to their clinical practice. Nor does enforcing data entry necessarily improve data quality as clinical staff may enter only the minimum data that will enable them to get on with their work, even if this makes the data less suitable for reuse.

While the direct costs of storing electronic medical record data may be low, infrastructure and staff costs to ensure that systems are secure and reliable may be an additional burden. Data protection standards may also restrict data reuse. Even if there is the potential for unrestricted reuse of data, this does not mean that the resources or the skills to extract data from medical records and/or to combine them into a consistent and valid dataset will necessarily be available. Nor will effective use of data necessarily be straightforward, as relevant analytical skills may be in short supply. Furthermore, without evidence of clear benefits, busy clinicians may be reluctant to devote attention to data reuse if this detracts from the immediate care of their patients. Even well-publicised cases of successful data reuse may be insufficient to encourage other sites to change their practices. We therefore cannot assume that we can reliably predict the degree or nature of changes that may be associated with data directly from the scale of their production.

4.7. Are big data voluminous?

The distinctiveness of big data, Kitchin (2014a) argues, cannot rest solely on their size. Large datasets, such as a national census or economic statistics are not a recent invention. Kitchin reports for example that 4.5 million cubic feet of government records were held

by the US National Archives and Records Administration in 2013. Nor, Kitchin and McArdle (2016) argue, are many datasets that are cited as examples of big data necessarily distinguished by their volume (or their variety). Contra Manyika et al. (2011) they can be captured, stored, managed and analysed with conventional database software tools (and, with technical advances, the boundaries of what can be handled by these tools keeps expanding, so what counts as big data changes over time). Rather, they argue, it is velocity and exhaustiveness that are the special features of big data. In particular it is the comprehensiveness of what is recorded (no more need to sample as storage space is cheap and effectively unlimited) and the speed and ease with which data can be retrieved and analysed that is the most striking difference from the past. Paper records were voluminous, but were laborious to compile and hard to extract data from. Now (nearly) everything can be easily stored and can be (more) readily found and processed.

Datafication has undoubtedly played a role in the proliferation of large datasets, as even if the size of individual data records may be small, the frequency of recording, the number of attributes being recorded and the number of contributing data sources can combine to produce a significant accumulation and complexity of data. Nevertheless, the simple volume of data would seem a poor indicator of their significance. Thus Newell and Marabelli (2015) argue that "little data", granular data that captures the minutiae of individual behaviour, can have important social and organisational effects. While Newell and Marabelli (2015: 5) define this "little" data as "big data … simply used in a more targeted way", other authors, such as Lindstrom and Heath (2016) and boyd and Crawford (2012), argue for the continuing value of "small data", often collected by ethnographic methods, the contribution of which risks being lost in the hype around data volume.

While it is certainly the case that massive datasets that can only be handled with specialist software open up new opportunities in the healthcare field, especially related to genomics, restricting attention to such data would be to overlook the possibly more significant changes associated with datasets which, although small in technical terms (and therefore able to be handled with conventional software tools), nevertheless enable an enlarged role for data in healthcare practice. Moving from a situation in which clinicians typically made little use of past data, because they were so unreliable and difficult to find, to one where at least some of these data may be available for reuse may make more difference to clinical work practices than breakthroughs from the analysis of data that formally meets the definition of "big".

4.8. Are big data universal?

If there is little question that the volumes of data currently being produced are of an entirely different order from those in the past, it does not follow, as often appears to be assumed, that we are anywhere near the situation in which it might legitimately be claimed that the "world has been datafied" (Mayer-Schoenberger and Cukier, 2013: 96). There are a number of reasons for this. For many proponents of big data it would seem to be axiomatic that there is nothing that cannot, and will not, be datafied. It may be questioned, however, whether all phenomena are capable of being quantified and whether what is quantified is either an accurate representation of the phenomenon, or an appropriate treatment of the phenomenon. Thus we might consider whether there would be likely to be common agreement on a standard measure of the quality of life between an individual with a particular medical condition or their family and a clinician or a random member of the public. This is not to say that efforts to define such measures are not possible, quality-adjusted life years (QALYs) are already used in the economic evaluation of medical interventions, but they remain controversial (Raftery, 2018).

The converse of this argument provides a second line of questioning of claims of the universality of datafication. This is that what gets datafied, tends to be phenomena that are more readily amenable to quantification (or that in many cases are quantified already). Big data are therefore heavily biased towards particular types of phenomena, the suitability of which in representing "the state of the world" may be questioned.

A third reason to doubt that universal datafication is immanent relates to the highly selective nature of datafication so far. Much is made, for example, of the volumes of data handled by particular companies, such as Google or Facebook, but their user base is not even universal within their primary audiences in Western countries (notwithstanding Facebook's claims to have more users in certain demographics in the USA than are recorded by the census (McCarthy, 2017)). This user base is also skewed towards particular ethnic, socio-economic and age groups. The representativeness of whatever insights might be gained from the data held by these companies may therefore be debatable even for the populations they are targeted towards, let alone for populations where take-up of their services is more limited. This may be of little concern to Google et al. if they can continue to sell advertising to other companies, but it raises questions about how universal "big data" actually are.

Another concern relating to the universality of big data is the effects of their current selectiveness on the algorithms used to analyse them. Many of the machine learning techniques used in big data analysis rely on the use of a portion of the dataset to "train" the algorithms. If the datasets themselves are biased, then the algorithms will also reflect those biases (O'Neil, 2016). As many of these algorithms, moreover, employ opaque models that make it impossible to audit fully the process by which they achieve their results, this may prevent the sources of potential discrimination being identified or addressed. While it may be possible to achieve some post hoc transparency on the relative influence of particular variables on an algorithm's output, as Hosanagar and Jair (2018) discuss, significant, potentially regulatory, incentives would seem likely to be required for this to become common practice. Claims of the universality of data may thus serve to deflect criticisms of algorithmic decision-making, even if claims about data biases may, in specific instances, be well-founded.

4.9. Are big data exhaustive?

The exhaustiveness of data may be seen as a special case of universality. Thus particular phenomena may be considered to be

completely datafied, such that N = all for that phenomenon, without claiming that everything is necessarily datafied. The criticisms of the selectivity of big data, that they only relate to certain categories, therefore apply too in relation to exhaustiveness. Thus, while Twitter may have access to all tweets ever posted, these are exhaustive only of the behaviour, in a particular context, of that subset of the population that use the service (and of some proportion of bots). The extent to which this represents an N of all for any wider population may therefore be questioned.

This is not necessarily a problem for Twitter if they wish to analyse the practices of their users, but for researchers seeking to understand public opinion, say, or popular communication this a significant, if not always acknowledged, limitation. This situation is made all the more difficult when such datasets are proprietary and access is controlled by the owners. In the case of Twitter, for example, much research relies on the APIs that the company makes available, which currently provide what is claimed to be a "small random sample" or, for a fee, a "10% random sample" of the real-time stream (Twitter, 2018). The terms of this access are also frequently revised (some researchers were previously given access to the full "fire hose", albeit up to an undocumented limit (Driscoll and Walker, 2014)), making it difficult to compare results over time.

If this is the case in terms of specifically digital phenomena then further selectiveness will be involved with data that represent naturally-occurring phenomena. A hospital patient's oxygen saturation, for example, may be measured with a pulse oximeter that produces a continuous stream of data, but only the values measured every 5 min, say, may actually be recorded and only the values at each hour may be archived. Even if all these values are accurately recorded and saved, which is not always the case in medical records, therefore, they represent only a sample of the patient's oxygen saturation levels. What does it mean therefore to say that N = all for such data?

The assumption of exhaustiveness also generally applies with regard to the data that have been recorded. The existence of data in the record, however, does not mean that they may actually be available for reuse, as there may be considerable "data wrangling, data munging or janitorial work" (Endel and Piringer, 2015) required, to identify, extract, clean and integrate these data to produce a dataset that is useable. This work is often not part of the skill set of those who may wish to use data, though, so, without assistance, they may be unable to access some or all of the data recorded. Thus we may distinguish between the "data in principle" that are present in the record and the "data in practice" that can be accessed. Only the latter, which may be a small subset of the former, can actually be made use of and therefore contribute directly[1] to individual, organisational or social change.

In pointing to a distinction between the universe of data and the proportion of that universe that are used, data in principle and practice bear some relation to the data and capta discussed by Checkland and Holwell. In contrast to Checkland and Holwell (1998: 89), however, data are not assumed to be naturally occurring facts, and capta not simply "the small fraction of the available data which we know about or pay attention to, or create". Rather all data are seen to be created, if not necessarily by those who use them, and data in practice are those that are actually used, "materialized in practice" as Orlikowski and Scott (2015: 204) put it, not just known about or paid attention to. A similar distinction is made by Aaltonen and Tempini (2014), drawing on the Aristotelian dichotomy of potentiality v actuality. They discuss actualisation, however, in terms of information rather than data. That is data are a potentiality from which information is extracted. The argument of this paper, though, precedes any consideration of what information might be gained from data. Rather the focus is on how data come to be available such that information could be extracted from them. Unless data are sought, selected, extracted, and interpreted they cannot inform.

It is, then, through their enactment in specific, situated practices that data contribute to the production of social life. As Feldman and Orlikowski (2011) argue, they are consequential. In doing so they are also performative (Orlikowski and Scott, 2015), shaping the conditions of their own materialisation. Thus the data that are enacted in practice, create the understanding of the phenomena they are taken to describe. If QALYs are used to decide what medical interventions are to be authorised then the conception of quality of life that they enact is the one that defines the treatment options for patients. Other data may exist that might adopt a different conception of quality of life, but if they are not materialised in practices that contribute to treatment decisions (in the UK, this predominantly involves their approval by the National Institute for Clinical Excellence (NICE)) then they are unlikely to be consequential.

Whether data get to be used, moreover, depends on whether the perceived value of data is sufficient to bear the cost of access, including, potentially, the costs of employing data specialists to undertake the work. A lot of data that are recorded are never subsequently reused because the resources (skills and funding) are not available to access them. Contrary to discussions of N = all, therefore, it may often be that N = "as much as we are able, and can afford, to access", which may be considerably less than all. This is a situation that may be familiar to academic researchers from their own practice, but it is not clear that its implications are always recognised in discussions of big data.

If data, therefore, are not necessarily, the referential, natural, foundational, objective and equal representation of the world that they are sometimes taken to be and big data are not necessarily as revolutionary voluminous, universal or exhaustive as they are often presented, then how might we understand data differently? Following the distinction between data in principle and data in practice, perhaps a starting point might be to recognise data as having two different aspects. The first relates to the process whereby what is considered to be data comes to be identified as such in the first place. What is it that gets to be recorded as data about a phenomenon? The second relates to the process whereby some, potentially all, of these data come to be in a position to influence individuals, organisations and society.

[1] It might be argued, following Foucault (1977), that the existence of data may influence individual, organisational and social change whether or not they are actually used, but this would not seem to be the position adopted in relation to the exhaustiveness of data. Rather N = all assumes that all data are directly usable.

5. How data come to be

Looking first at how data are produced, it would seem reasonable to maintain the assumption that data are generally intended to be referential. That is, with maybe a few exceptions, data are collected and used on the basis that they tell us something (although perhaps not everything) about the world. The initial stage in the creation of data, therefore, involves a decision on the phenomenon that they are considered to be a representation of. This decision does not necessarily have a single answer, however. The symptoms that are seen as evidence of a particular health condition, for example, will typically be based on established models of the condition's aetiology. Although the model for many conditions may be relatively standardised, at least within the thought community (Fleck, 2012) of western allopathic medicine, there may be differences of emphasis between specialisms (Mol, 2002). These models may also change over time as understanding of the condition's aetiology evolves. Furthermore, as Payer (1996) discusses, there may also be variation in the range of conditions that are recognised in different countries. Phenomena that might be represented by data are therefore not necessarily fixed and objective, but are socially shaped.

Having identified a state of the world on which data are to be obtained, a second necessary stage will be to decide what should be considered to be data about this phenomenon. This may not necessarily be a straightforward matter either, though, as there may be alternative models of a health condition, that may evolve over time, and a variety of data associated with any particular model. It may therefore be necessary to make a choice, even within the worldview of a particular thought community, on which data to record.

It may be noted, moreover, that this process may sometimes be reversed and that the question posed may be what phenomenon can available data be considered to represent, rather than what data most accurately represent the phenomenon. Such behaviour may not be wholly unfamiliar in some forms of research, but it may also be found in healthcare, where data are recorded on parameters that are assumed to be a reasonable proxy for some phenomenon on which it is not possible to obtain data directly.

A third stage in the process of recording data is often a pragmatic decision on what data it is possible to access and record. Inaccessibility may be a matter of technical or practical feasibility, or of cost, or of the effects of data recording practices on other processes. Invasive temperature monitoring of critically ill patients is more accurate than non-invasive methods, for example, but carries a risk of infection. In healthcare settings advances in technology are continuously expanding the range and accuracy of data that can be recorded, although whether these technologies will be deployed in any particular setting is often a matter of economics.

Although the costs of recording data are low, it does not follow, as is sometimes argued in discussions of big data, that all possible data will necessarily be recorded. There will often be a process of selection of which data to record, immediately and in the longer term. This selection may reflect technical constraints on the capacity or complexity of data recording as well as judgements on the appropriate level of detail with which data should be recorded, perhaps in the light of the expectations regarding their future use. What gets recorded will therefore generally be a subset of all possible data and this may be winnowed further when data come to be archived. Judgements on what to record will be shaped by perceptions of the use to which data may be subsequently put, and also potentially by data protection legislation. In a healthcare setting, for example, some physiological parameters may be monitored continuously, but it may be sufficient for patient treatment that they are recorded every five minutes. For a clinician reviewing notes of a treatment episode, even these data may be too detailed to enable a rapid assessment of the patient's condition, so hourly values may be sufficient for archived data.

The final stage of the process by which data come to be held in the record relates to the reliability of recording. As was discussed in relation to the referentiality of data, that a decision was made and procedures established to record particular data does not always mean that the data will actually be recorded. Errors, equipment malfunction or breakdown or corruption of the record may mean that what is found in the record may not always match what was originally registered. What finally constitutes the data recorded about a particular phenomenon, therefore, is not a direct, objective and exhaustive representation of the phenomenon, but the outcome of both a number of choices (which may not always be evident if they are enacted indirectly, for example in the design of technologies) and of a number of technical, social and economic contingencies. This process is summarised in Fig. 2.

6. How data come to be used

Even the presence of data in the record, however, does not necessarily equate to what actually gets used as data about a phenomenon and there is a further process that mediates between the two. This may also be broken down into a number of stages as shown in Fig. 3.

A necessary starting point for the use of data would seem to be some demand that they are perceived to fulfil. A clinician treating a patient for example seeks data that will help them in their task. The specific data they look for will depend on what condition they believe the patient to be suffering from and their model of the condition's aetiology. Although there may be general agreement on this model, there can be individual and specialism-related differences in what data are seen to be relevant to treatment.

Even if data are sought, however, this does not mean that they will necessarily be found. There may be a number of reasons for this. For example, those wishing to use data may not have an understanding of the data that are held, so may request data that are unavailable; or users may lack the search skills to locate relevant data themselves; or may be unable to formulate effective requests that would enable somebody else to find data for them.

Nor does finding data necessarily mean they can or will be retrieved. Extracting data and making them available in a form that can be used generally requires scarce technical skills that may be beyond the capabilities of many of those who might wish to use data. It may therefore be necessary to employ data scientists to carry out this work, although this introduces additional costs and may give rise to communication problems between domain and IT specialists. Nor is this data wrangling a trivial task. Lohr (2014) reports that transforming, editing, cleaning and assembling data into a form in which they can be used may take up 50–80% of data scientists'

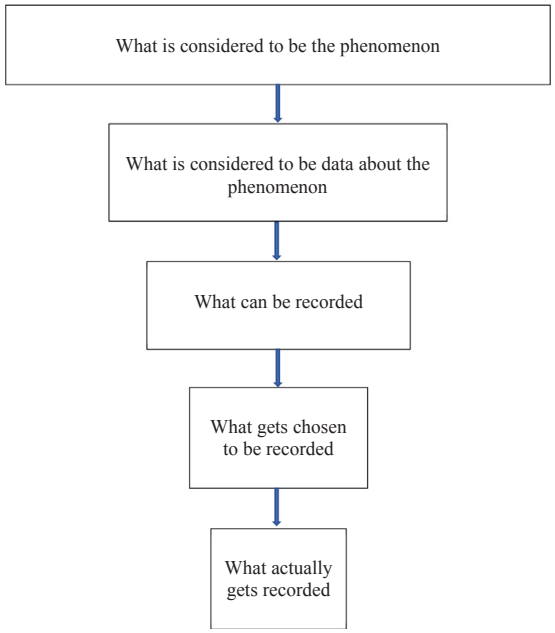

Fig. 2. How data come to be.

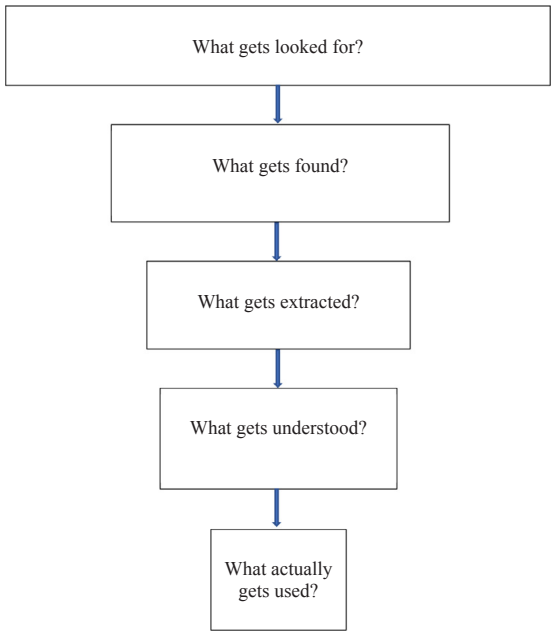

Fig. 3. How data come to be used.

work time. In itself this may be enough to deter data reuse, especially as the associated costs are rarely adequately budgeted for.

Making data available to users may still not be sufficient to ensure they can be used, as there may be many challenges in understanding and interpreting data, especially when dealing with large datasets that are too complex for users to make sense of unaided. Further support, such as visualisations and analytical tools, may therefore be necessary to enable individuals to be in a position to use data. With big data in particular these tools may be the primary, or even only, way in which individuals gain access to data. As Orlikowski and Scott (2015) discuss however, notwithstanding the "aura of objectivity" (Reynolds, 2016) associated with these tools, they are not passive conduits, but enact a selective cut of the data that performs a particular representation of the phenomena they are seen to describe. The tools furthermore add to the costs of data as users need to learn how to interpret their output in relation to the purposes to which they wish to put the data. The individual needs to believe that these costs will be outweighed by the benefits they will gain from data use, although both elements of this calculation may be difficult to assess.

Unless data contribute to some effect on phenomena in the world, possibly indirectly and possibly after some delay, however, they will remain data in principle and never become data in practice. Data in practice are thus the culmination of long series of steps that lead in the first place to the existence of something that we identify as data and then to this data having an effect in the world. This is not an inevitable or necessarily reliable process, however. At each step there is the possibility of breakdown and of what is taken to be data being altered in some way. This would seem a long way from the transparent and complete reporting of states of the world and suggests that data in practice provide a potentially much more selective and partial representation of the phenomena they are seen to describe than is often recognised. It also means that much of what gets referred to as data, are only so in principle.

In healthcare settings, for example, the design of record systems tends to prioritise the immediate care of patients, with less attention to the subsequent reuse of data. While it can be easy to enter data and to view them during an episode of care, therefore, retrieving data on previous patients for comparison or analysis tends to require IT skills that few clinicians possess. Nor, traditionally, have there been sufficient staff with the requisite skills in clinical settings to provide an alternative route to access data. While significant volumes of data may be produced and stored in healthcare settings, as the big data literature describes, often little of this data is actually routinely accessible by clinicians. Without access to these data, moreover, clinicians have no opportunity to make use of them in their practice so there is little clinical incentive to do something about this situation. Data will continue to accumulate in repositories, unused and unconsidered.

Figs. 2 and 3 depict the coming to be of data and the coming to use of data as purely linear processes that operate in just one direction. Not all data, however, may necessarily follow such a trajectory. There is also a growing trend of data reuse (actively promoted through open data initiatives e.g. Open Data Institute (https://theodi.org/)). Although in most cases, such data may be argued to have come into being as a result of the process shown in Fig. 2, shaped and motivated by particular purposes for which they have been recorded, it is also possible that some data may come into being adventitiously, without any particular purpose in mind. Whatever the provenance of these data, their existence creates the potential for experimental or exploratory analyses in search of "interesting patterns" (Aaltonen and Tempini, 2014: 104). Fig. 3 may therefore be short-circuited, with data already extracted, although it would seem important not to overlook how these data will have been shaped by the earlier stages as they were originally recorded and subsequently compiled into accessible datasets.

7. Discussion and conclusion

If we consider data not as givens that are out there in the world, waiting to be gathered, but as contingent representations that are brought into being through situated practices of conceptualization, recording and use, then what might this mean for our understanding of datafication and of big data more generally? While a complete answer to this question is clearly beyond the scope of this initial account of data in practice, four broad domains of implications may be identified.

The first of these proceeds from the problematization of the given character of data. Thus, rather than starting point for our analysis being the existence of data, the issues, impacts and implications of which are to be examined, we need to extend our attention upstream to consider how what are taken to be data come into being. Which particular minutiae of an individual's life are being continuously monitored by digital devices, for example? How accurate and reliable is this monitoring? What representation of the individual's everyday life could this monitoring be considered to provide?

It is not that, to paraphrase Thomas and Thomas (1928), if data define situations as real they are not real in their consequences, but that the reality defined by data may not always be complete or accurate and the consequences therefore not directly a result of the situations data are considered as describing. It is equally the case, as Merton (1995) argues, that the consequences of situations are not dependent on their recognition by data. We therefore need a more critical attitude to data that recognizes that they do not necessarily represent the world transparently, even if, in many situations, we may not have any better way of knowing the actual state of the world. Part of the power of datafication is therefore that it offers a way of knowing the world that it can be hard to challenge. It may not be that individuals' everyday lives are well-represented by available data traces, but that, as the available representations of these lives, held by organisations that wish to know about individuals in certain ways, they become the effective reality. An appreciation of the constructed character of these representations may not undermine their influence, but at least allows us to start to question their objectivity and infallibility.

A second implication of this view of data concerns the necessarily selective character of the representation of phenomena that they offer. The world is always underdetermined by the representations that data provide. The minutiae they record are, pace big data proponents, always a sample of reality. Perhaps this sample is much richer than has been available hitherto, but it is still not congruent with reality. Nor is this selectivity necessarily neutral. While there may sometimes be overt bias in this selection, it may also simply reflect the fact that data tend to report phenomena that are easy to represent as data. Phenomena that are less amenable to datafication are relatively excluded and risk being overlooked. We therefore need to pay attention to the influences that may be acting on this selection.

The processes whereby data come into being and come to influence organizational practices also do not, for the most part, happen by themselves. Work is involved at each stage. This work may in some circumstances be automated, but work will still be needed in the initial design of the requisite algorithms, at least until the singularity occurs. Too often, however, data are discussed as if they simply exist with no need to consider either their provenance or their accessibility. Yet it is the cost of the work in creating and using data that may be a major influence on the effects they have in organisations. Data that are difficult to produce or hard to use are likely to have little impact. Greater attention would seem needed in discussions of data to the cost of making them available and how this affects their use.

The final implication of this view of data follows from the distinction between data in principle and data in practice. What would

seem clear is that much of what is discussed as data, and big data in particular, is data in principle only. Yes, we are accumulating staggering quantities of data, but how much of this is used (or usable)? Perhaps within Google or Facebook all the data they record are completely accessible (notwithstanding their size, geographical range and history of growth by acquisition), but even the creation of a unified "data lake" (Terrizzano et al., 2015) may not mean that data are necessarily ever reused (Stein and Morrison, 2014). Perhaps more typically, data are held in local silos each with their own standards and schemas, making extraction and integration of data difficult, if not impossible in practice. If it is considered desirable that the proportion of data in practice should increase, not least to realise the potential of big data, then solutions need to be found to the proliferation of such "data graveyards". As yet, however, this would seem to have attracted surprisingly little attention in the big data literature.

Custer et al. (2016, 2017) discuss the problem of data graveyards in relation to development data, but their content/channel/choice model of how these may be avoided would seem more widely applicable. The "content" element questions whether data are fit for purpose. In terms of data in principle/practice this may be viewed as tracking back from the envisaged use of data to try to ensure that relevant phenomena are appropriately represented in the data recorded. "Channel" considers whether users can easily find, access and use data, as illustrated in Fig. 3, while "Choice" addresses whether the perceived benefits of using data outweigh their costs. If such a high-level model offers little direct guidance on how more of the ever-growing volume of data observed in society can make a difference in practice, it at least indicates a wider recognition that accumulating data per se will not be enough to bring about individual, organizational or social change and that the barriers to the use of data (often more social and technical) deserve more consideration in discussions of (big) data.

Recognising the complexity, consequentiality and constructed character of data also has implications for the way in which they may be drawn on in strategy. Thus, as Constantiou and Kallinikos (2015) discuss, data play a central role in much strategic planning, providing senior management with an awareness of the organisation's environment and of the efficiency of its internal processes. Indeed proponents of evidence-based management such as Pfeffer and Sutton (2006: 63) argue that managerial judgement and decision-making should be based exclusively on data, or what they describe as "the facts about what works". Although Rousseau (2012: 9) acknowledges that "raw data" may be biased, may omit important information, may be political or may be open to different interpretations, it is argued that, with efforts to identify and transform these data and to reduce errors and unreliability, they can (and need to be) transformed into reliable information. This presupposes, however, that there is a single understanding of a phenomenon that can be fully represented by data and that we have some independent means of establishing whether these data correctly report that understanding, both of which positions would be problematized on the view of data presented here. This is not to say that data cannot inform an understanding of the world, but that we should be aware of the selectivity of that understanding and of the assumptions behind that selection. Evidence-based management may therefore be questioned, not just in terms of who determines what counts as evidence or what counts as a better decision, as its critics suggest (Hornung, 2012), but also in terms of the objectivity and referentiality of the data on which it draws.

The selectivity of big data is already recognized, at least in some circles, as raising ethical issues of bias in the representation and analysis of phenomena, but we may argue that this is an intensification of a more general characteristic of data that deserves greater attention. In the context of big data, these concerns about selectivity may be a reaction to proponents' strong assertions of big data's universality and exhaustiveness, but the novelty of big data also provides an opportunity to question such claims. With data more generally, however, the claims are less prominent and familiarity has perhaps enabled them to fall out of awareness. From a strategy perspective, however, it would seem important not to lose sight of the potential effects of the selectivity of data on the decisions that are informed by them. The willingness of organisations to acknowledge and address these effects may also serve as a means of strategic differentiation.

The argument of Constantiou and Kallinikos (2015) focuses on the implications of big data for strategy and in particular of their dynamic, heterogeneous and unstructured nature for the traditional tools and theories of strategy making. It is not just with big data, however, that assumptions of stability, homogeneity and structure may be questioned. Many of the terms that Constantiou and Kallinikos (2015: 53) use to describe the nature of data in the "standard strategy context" (such as relatively homogeneous, purposeful, alpha-numerical) may be argued to reflect not so much characteristics of the context as of the data that are taken as representing this context. We may therefore view these characteristics as performative effects of the data that are available and valorized in the Fordist mass production paradigm rather than an objective description of a particular state of the world that may, or may not, be receding. This does not mean that the effects are any less powerful or that traditional strategy tools and theories are illegitimate, but it does allow us to question whether they represent a natural order.

The distinction between data in principle and data in practice would also seem to have implications for strategy. Organisations may record large volumes of data, but only those that are materialised in practice will be consequential. This requires not just that data can be accessed, but that they are understood and made use of in practice. How data are understood, however, is a subjective process. There may be well-established conventions for interpreting data, but these are not necessarily infallible or immutable (even if their durability makes them appear so). We therefore need to look not just at the data that are available, but at how the interpretations that are applied to them are arrived at and how these inform situated action. Attention to the doing of strategy has been the hallmark of the practice turn in strategy research (Whittington, 2006). What this paper highlights is a need for a critical focus on the particular ways that data are drawn on in this process of doing.

Furthermore, rather than, as Constantiou and Kallinikos (2015: 51) argue, transcending "the 'nitty-gritty' practices by which situated decision makers make sense of locally embedded phenomena, tweak and fix data, descriptions and objectives", the larger-scale developments they associate with big data, that "frame attention and sample events", always have to be materialized in situated practices. Perhaps the time horizon will be different, the analysis less structured, the process more inductive, but any effects big (or little) data may have in organisations will still be played out through specific, local practices. Addressing the situatedness of data use

does not preclude analysis of larger-scale trends, although it may incline us to greater caution regarding their pervasiveness and the imminence of their supplanting of traditional practices.

In drawing attention to the gulf between the way data are commonly presented and the way that they are produced and used in practice, the aim is not to dismiss their significance or potential to transform organisations and society. Rather it is to seek to encourage a richer awareness of the complexities of data and of the often unrecognized work that is involved in making them available for use. Awareness of this work, moreover, should encourage greater recognition of the social processes that shape data and of the scope for intervention to influence their production and use.

Acknowledgements

The ideas presented in this paper were developed as part of the ReCliC project on the repurposing of clinical data for quality improvement in critical care, a collaboration between the Judge Business School and the Computer Laboratory at the University of Cambridge and Royal Papworth Hospital, funded by the Health Foundation, an independent charity working to improve the quality of healthcare in the UK.

References

Aaltonen, A., Tempini, N., 2014. Everything counts in large amounts: a critical realist case study on data-based production. J. Inf. Technol. 29, 97–110.
Ackoff, R., 1989. From data to wisdom. J. Appl. Syst. Anal. 16, 3–9.
Anderson, C., 2008a. The Petabyte age: because more isn't just more – more is different. Wired.
Anderson, C., 2008b. The end of theory: the data deluge makes the scientific method obsolete. Wired.
Bocij, P., Chaffey, D., Hickie, S., Greasley, A., 2006. Business Information Systems: Technology, Development and Management for the e-Business, third ed. Financial Times/Prentice Hall, Harlow.
Boyd, D., Crawford, K., 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Inf. Commun. Soc. 15 (5), 662–679.
Checkland, P., Holwell, S., 1998. Information, Systems and Information Systems: Making Sense of the Field. John Wiley and Sons, Chichester.
Constantiou, I.D., Kallinikos, J., 2015. New games, new rules: big data and the changing context of strategy. J. Inf. Technol. 30 (1), 44–57.
Cukier, K., Mayer-Schoenberger, V., 2013. The rise of big data: how it's changing the way we think about the world. Foreign Aff. 92, 28.
Custer, S., Sethi, T., Custer, J., 2016. Pork to Performance: Open Government and Program Performance Tracking in the Philippines – Phase two. AUS18311. The World Bank, pp. 1–117.
Custer, S., Sethi, T., Custer, J., 2017. Avoiding data graveyards: how can we overcome barriers to data use? Available at: https://www.aiddata.org/blog/avoiding-data-graveyards-how-can-we-overcome-barriers-to-data-use.
Davenport, T.H., Prusak, L., 1997. Information Ecology. Oxford University Press, Oxford.
Driscoll, K., Walker, S., 2014. Working within a black box: transparency in the collection and production of big twitter data. Int. J. Commun. 8, 20.
Drucker, P.F., 1998. The coming of the new organisation. Harvard Bus. Rev.
Endel, F., Piringer, H., 2015. Data Wrangling: Making data useful again. IFAC-PapersOnLine 48 (1), 111–112.
Feldman, M.S., Orlikowski, W.J., 2011. Theorizing practice and practicing theory. Org. Sci. 22 (5), 1240–1253.
Fleck, L., 2012. Genesis and Development of a Scientific Fact. University of Chicago Press, Chicago.
Foucault, M., 1977. Discipline and Punish: The Birth of the Prison. Allen Lane, London.
Frické, M., 2009. The knowledge pyramid: a critique of the DIKW hierarchy. J. Inf. Sci. 35 (2), 131–142.
Frické, M., 2015. Big data and its epistemology. J. Assoc. Inf. Sci. Technol. 66 (4), 651–661.
Galliers, R.D., Newell, S., Shanks, G., Topi, H., 2017. Datification and its human, organizational and societal effects: the strategic opportunities and challenges of algorithmic decision-making. J. Strateg. Inf. Syst. 26 (3), 185–190.
Gitelman, L., Jackson, V., 2013. Introduction. In: Gitelman, L. (Ed.), "Raw Data" is an Oxymoron. MIT Press, Cambridge, MA, pp. 1–14.
Groves, P., Kayyali, B., Knott, D., Van Kuiken, D., 2013. The 'big data' revolution in healthcare. McKinsey Quart. 2 (3).
Günther, W.A., Rezazade, M., Huysman, M., Feldberg, F., 2017. Debating big data: a literature review on realizing value from big data. J. Strateg. Inf. Syst. 26 (3), 191–209.
Haag, S., Cummings, M., 2013. Management Information Systems for the Information Age, ninth ed. McGraw-Hill Irwin, New York, NY.
Hornung, S., 2012. Beyond "New Scientific Management?" Critical reflections on the epistemology of Evidence-based Management. In: Rousseau, D.M. (Ed.), The Oxford Handbook of Evidence-based Management. Oxford University Press, Oxford, pp. 389–403.
Hosanagar, K., Jair, V., 2018. We need transparency in algorithms, but too much can backfire. Harvard Bus. Rev. Digital Art. 2–5.
Kitchin, R., 2014a. The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. SAGE, London.
Kitchin, R., 2014b. Big data, new epistemologies and paradigm shifts. Big Data Soc. 1 (1) 2053951714528481.
Kitchin, R., McArdle, G., 2016. What makes big data, big data? Exploring the ontological characteristics of 26 datasets. Big Data Soc. 3 (1) 2053951716631130.
Laney, D., 2001. 3D Data management: controlling data volume, velocity and variety. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.
Lindstrom, M., Heath, C., 2016. Small Data: The Tiny Clues That Uncover Huge Trends. St. Martin's Press, New York City.
Loebbecke, C., Picot, A., 2015. Reflections on societal and business model transformation arising from digitization and big data analytics: a research agenda. J. Strateg. Inf. Syst. 24 (3), 149–157.
Lohr, S., 2014. For big-data scientists, "janitor work" is key hurdle to insights. New York Times 17 August.
Manovich, L., 2012. Trending: the promises and the challenges of big social data. In: Gold, M.K. (Ed.), Debates in the Digital Humanities. University of Minnesota Press.
Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Hung Byers, A., 2011. Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute.
Markus, M.L., 2017. Datification, organizational strategy, and IS research: what's the score? J. Strateg. Inf. Syst. 26 (3), 233–241.
Mayer-Schönberger, V., Cukier, K., 2013. Big Data: A Revolution that will Transform How We Live, Work and Think. John Murray, London, pp. 2013.
McCarthy, K., 2017. Facebook claims a third more users in the US than people who exist. https://www.theregister.co.uk/2017/09/06/facebook_claims_more_users_than_exist/.
Merton, R.K., 1995. The Thomas theorem and the Matthew effect. Soc. Forces 74 (2), 379–424.
Mol, A., 2002. The Body Multiple: Ontology in Medical Practice. Duke University Press.
Newell, S., Marabelli, M., 2015. Strategic opportunities (and challenges) of algorithmic decision-making: a call for action on the long-term societal effects of "datification". J. Strateg. Inf. Syst. 24 (1), 3–14.
O'Neil, C., 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Allen Lane, London.
Orlikowski, W.J., Scott, S.V., 2015. The algorithm and the crowd: considering the materiality of service innovation. MIS Quart. 39 (1), 201–216.
Payer, L., 1996. Medicine and Culture: Varieties of Treatment in the United States, England, West Germany and France, First Owl Book ed. Holt Paperbacks, New York.
Pfeffer, J., Sutton, R.I., 2006. Evidence-based management. Harvard Bus. Rev. 84 (1), 62.

Raftery, J., 2018. A more fundamental review of QALYs is needed. The BMJ.
Raghupathi, W., Raghupathi, V., 2014. Big data analytics in healthcare: promise and potential. Health Inf. Sci. Syst. 2 (1), 3.
Rahman, R., Reddy, C.K., 2015. Electronic health records: a survey. Healthc. Data Anal. 36, 21.
Reynolds, D.A.J., 2016. Algorithm is gonna get you. The Technoskeptic.
Rosenberg, D., Gitelman, L. (Eds.), 2013. "Raw Data" is an Oxymoron. MIT Press, Cambridge, MA, pp. 15–40.
Rousseau, D.M., 2012. Envisioning evidence-based management. In: Rousseau, D.M. (Ed.), The Oxford Handbook of Evidence-based Management. Oxford University Press, Oxford, pp. 3–24.
Rowley, J., 2007. 'The wisdom hierarchy: representations of the DIKW hierarchy. J. Inf. Sci. 33 (2), 163–180.
Royal Society, 2012. Science as an open enterprise. Available at: http://royalsociety.org/policy/projects/science-public-enterprise/report/ 2012.
Stein, B., Morrison, A., 2014. The enterprise data lake: better integration and deeper analytics. PWC Technol. Forecast: Rethink. Integr.(Issue 1). https://www.pwc.com/us/en/technology-forecast/2014/cloud-computing/assets/pdf/pwc-technology-forecast-data-lakes.pdf.
Terrizzano, I.G. et al., 2015. Data wrangling: the challenging journey from the wild to the lake. In: Proceedings of the Conference on Innovative Data Systems Research, Asilomar California, 4–7 January. Available at: https://pdfs.semanticscholar.org/2a24/f587b68a1ef6539b4ed8725dfe76f0ed40e2.pdf.
Thomas, W.I., Thomas, D., 1928. The Child in America; Behavior Problems and Progress. AA Knopf, New York.
Twitter, 2018. Sample realtime tweets. Available at https://developer.twitter.com/en/docs/tweets/sample-realtime/overview/GET_statuse_sample (accessed 05 October 2018).
Tuomi, I., 1999. Data is more than knowledge. J. Manage. Inf. Syst. 16 (3), 107–121.
Turban, E., 2006. Information Technology for Management: Transforming Organizations in the Digital Economy, fifth ed. Wiley, Hoboken.
Weiskopf, N.G., Weng, C., 2013. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J. Am. Med. Inform. Assoc. 20 (1), 144–151.
Whittington, R., 2006. Completing the practice turn in strategy research. Org. Stud. 27 (5), 613–634.
Zins, C., 2007. Conceptual approaches for defining data, information, and knowledge. J. Am. Soc. Inform. Sci. Technol. 58 (4), 479–493.
Zuboff, S., 1988. In the Age of the Smart Machine. The Future of Work and Power Heinemann, Oxford.