cogent
social
sciences

Big data business models: Challenges and opportunities

Ralph Schroeder[1]*

Abstract: This paper, based on 28 interviews from a range of business leaders and practitioners, examines the current state of big data use in business, as well as the main opportunities and challenges presented by big data. It begins with an account of the current landscape and what is meant by big data. Next, it draws distinctions between the ways organisations use data and provides a taxonomy of big data business models. We observe a variety of different business models, depending not only on sector, but also on whether the main advantages derive from analytics capabilities or from having ready access to valuable data sources. Some major challenges emerge from this account, including data quality and protectiveness about sharing data. The conclusion discusses these challenges, and points to the tensions and differing perceptions about how data should be governed as between business practitioners, the promoters of open data, and the wider public.

Subjects: e-Business; Economics, Finance, Business & Industry; Sociology of Media

Keywords: big data; business; digital media

1. Introduction

Big data is increasingly seen as an essential element of a well-functioning economy. A number of reports and academic publications have pointed to the growing use of big data across economic sectors (Brynjolfsson, Hitt, & Heekyung, 2011; Bulger, Taylor, & Schroeder, 2014; George, Haas, & Pentland, 2014; Manyika et al., 2011; Schroeck, Shockley, Smart, Romero-Morales, & Tufano, 2012; Taylor & Schroeder, 2014; Taylor, Schroeder, & Meyer, 2014; Thomas & McSharry, 2015) and its potential to bolster productivity, efficiency, and growth. The realisation that data use is likely to become increasingly important and widespread in the coming years has led to discussions of how

ABOUT THE AUTHOR

Ralph Schroeder is a professor and director of the Master's degree in Social Science of the Internet at the Oxford Internet Institute. Before coming to Oxford University, he was a professor in the School of Technology Management and Economics at Chalmers University in Gothenburg (Sweden). His recent books are Rethinking Science, Technology and Social Change (Stanford University Press, 2007) and, co-authored with Eric T. Meyer, Knowledge Machines: Digital Transformations of the Sciences and Humanities (MIT Press 2015). He is the author of 6 books, editor and co-editor of 4 volumes, and has published more than 125 papers on virtual environments, Max Weber, sociology of science and technology, e-Research, and other topics.

PUBLIC INTEREST STATEMENT

The opportunities of big data in business have been much discussed in recent years, but this is also still an emerging area, with many uncertainties about what business models will succeed. This paper develops a typology of these models, centred on the sources of data that are used and the challenges of these sources. The paper is based on 28 interviews with leading figures and with different types of expertise. Among the findings is that much depends on the quality of data sources and how they can be deployed. Another notable finding is how often openly available data-sets are used in conjunction with proprietary data-sets. The paper concludes with reflections on the future of big data in business, and how insufficient thought is given to how big data is defined, and the possibilities and limits of various data sources.

cogent ·oa

best to promote big data approaches by means of policy and regulation (Brown & Marsden, 2013; Pasquale, 2015). Any such policies must be predicated upon a thorough understanding of the prevailing landscape of how big data is being used by firms, and especially how different data sources are being used.

This article aims to furnish such an understanding by addressing three questions relevant to the use of big data. Firstly, what are the practitioners' views of big data: how do they define this concept and how is its influence being felt in their industry? This helps to inform a general picture of the current data landscape and sheds light on the way businesses and other stakeholders are adopting technologies and techniques to increase their use of data. Secondly, we can turn to business models with a view to learn, in broad terms, what opportunities have so far been identified and the various pathways to their implementation. Thirdly, the article investigates the main challenges faced by organisations working in the data economy and seeks to understand what, if any, steps can be taken to mitigate these challenges. In addressing these questions, the article draws on a series of 28 interviews with industry experts from a range of sectors.

The main findings can be summarised as follows: First, industry leaders acknowledge a wide variety of exciting opportunities connected to an ever increasing capacity to collect, store and analyse data. However, there has not, so far, been a big bang moment at which entire sectors simultaneously and completely transform, thanks to the increased use of data. Rather, business continues to undergo a significant *but gradual* transition towards a more data-driven landscape. In particular, many businesses started the wholesale incorporation of data into their business model decades ago, and many industries are still exploring the space of possible applications for (and sources of) data. There continue to be, therefore, opportunities for additional firms to realise the benefits of increased data utilisation—especially when doing so in new or innovative ways. Secondly, we can identify three classes of big data business models: data users, data suppliers and data facilitators. These three classes are mutually dependent but analytically separate, and a well-functioning, data-oriented economy will simultaneously cultivate the growth of all three. Thirdly, there are a number of significant challenges facing big data firms. These challenges are regarded mostly as internal and reflect procedural problems in collecting, archiving and handling data. As such, it appears that an important objective for policy-makers seeking to encourage the efficient use of data in the economy should be to promote good practice in these areas. Yet there are also wider challenges about the use of big data within society-at-large. These have been widely discussed in the literature (Boyd & Crawford, 2012; Ekbia et al., 2015), but these discussions are not based on definitions of data and big data and so they do not distinguish adequately between commercial and government uses as opposed to research uses. In the conclusion, such an analytical definition (Cowls & Schroeder, 2015; Schroeder, 2014) will be presented which allows making such a distinction and thus provides a perspective on the findings derived from the interviews but also on the social implications of big data more generally. The aim here in the first instance is to provide a picture of how practitioners see big data affecting business and policy in practice, rather than to impose a definition from the start.

2. Related literature
There is a growing literature about the use of big data in business, but business models as such have not received sustained analysis. Mayer-Schönberger and Cukier (2013) described a range of early uses of big data and made suggestions for tackling some of the emerging legal and regulatory issues. A more recent analysis of these issues can be found in Pasquale (2015) (see also Lane & Stodden, 2014). While an overview of the broader issues is beyond the scope of this paper, we will see that the main concern among our interviewees is that government should create an appropriate environment such that privacy and other legal issues do not impede taking advantage of new sources of data. The advantages of these new sources for economics have been discussed by Einav and Levin (2014) and for the developing world, including economic development by (Taylor & Schroeder, 2014). Brynjolfsson et al. (2011) showed that firms using big data had better performance than those that did not, although this study would require updating in view of the rapid development of this field. Apart from this, a number of reports showcase individual uses of big data (for government,

see Clarke & Margetts, 2014; for research, see Borgman, 2014; Eagle & Greene, 2014; Pentland, 2014; the current state of the art in relation to policy uses, see http://www.data4policy.eu/#!sota/cbiv).

It can mentioned that the literature bearing on this topic comes from a variety of academic disciplines, including law, economics and business and management studies, sociology and development, and computer science and information science (to name only the main ones). As we shall see, the same is true of our interviewees, whose backgrounds span a wide range. Within such an emerging area, it is no surprise to find such a range of perspectives, though in the conclusion it will be argued that a focus on data sources can anchor the discussion and provide a purchase on the challenges ahead. This paper seeks to go beyond the existing literature by concentrating, even if still at an early stage, on the different kinds of business models that are deployed with big data uses, and to identify opportunities and challenges that arise with different approaches.

3. Data
In addition to desk research, this study relies on interviews with 28 industry experts (the roles and institutional affiliations of interviewees cited are provided in the text below) from a range of sectors using semi-structured interviews (10 items were structured to provide consistency across answers). These interviews included open-ended questions to allow exploring emerging issues in depth, as common in research in novel areas. The interviews focused on types of business models, challenges in collecting and using data, characteristics of businesses poised to take advantage of big data, skills needed and how government can promote productive use.

4. Methods
Experts were identified through their prominence in industry publications and conferences and via the snowball method. Participants were invited via email. Interviews were conducted face-to-face, via Skype, or telephone between September 2013 and January 2014. Each interview lasted between 30 and 90 min, and interviews were transcribed for analysis. Eight women and 20 men participated in the study representing public, private and civil society organisations from a range of sectors including finance, advertising, data management, software development, analysis platforms, risk analysis, human resources, research/education, retail, health, public service and the quantified self. Experts from the public sector included representatives from data.gov.uk, data.gov, the UK's Open Data User Group and Administrative Data Liaison Service. Experts from the civil society sector represented the World Bank (a data scientist making data available for use especially by civil society) and the World Wide Web Foundation. The topic was also discussed informally with analysts in Seattle, Silicon Valley and London who requested anonymity. The study was governed by the ethical procedures of the University of Oxford (https://www.admin.ox.ac.uk/curec/).

5. Additional materials
In addition to interviews, data were collected on the organisations represented by our experts, covering dimensions such as industry, and organisation age and size. Companies included in the study represent a range of sizes, from Mappiness, managed by one founder, to IBM with over 400,000 employees on five continents. A blend of newer and older companies is represented, including the 180-year old Willis Group and more recently founded companies Gild, Drawbridge and Datacratic. Of the 14 companies represented in the study, 9 are headquartered in the US (4 of which have offices in London), 4 are headquartered in the UK and 1 is based in Canada.

6. The evolving big data landscape
Attempts by scholars and industry leaders to study and conceptualise the (potential) benefits of a big data economy have been stymied by difficulties in defining big data. Indeed, if this concept is to be defined by its "bigness" then many would say that the size of a data-set usually exists on a continuum with no obvious threshold for qualification. Dr. Boris Mouzykantskii, Founder and CEO of IPONWEB, remarked "I don't think anybody really talks about small data any more...anything which is data is now big data."

More concretely, proposed definitions might broadly be classified into absolute and relative approaches. Absolute definitions lay down a set of criteria that any data collection or analysis activity must satisfy in order to be classified as big data. This approach is typified by a pair of studies by Gartner, the consultancy firm, and IT firm IBM (Schroeck et al., 2012). These studies jointly define big data in terms of four V's: volume, velocity, variety and veracity. These four dimensions, respectively, account for the amount of data generated or processed, the speed or frequency with which it is recorded and analysed, the range of sources and data types (e.g. demographic, textual, geographic, image, etc.) that are brought together and the reliability with which measurements are conducted and data are captured. This approach is well-known within the business community and a number of our interviewees referred to one or more of these dimensions in describing the nature of their work with data. The absolute approach, though, has a number of drawbacks. Firstly, the threshold issue remains. Although there may be consensus that big data frequently involve large volumes of high-velocity data, it remains unclear what *exactly* the minimum volume and frequency are. Selecting a cut-off for variety and veracity is even more problematic given that these dimensions are not readily quantifiable. Secondly, it remains unclear how data that partially satisfies this criteria are to be categorised. For example, should large volumes of data that exhibit little variety, velocity or veracity be thought of as big data? What about reliable (high veracity) data that exists only in small volumes?

A third problem with the absolute approach is that it ignores the dynamic nature of the technological environment within which big data is used. Little more than a decade ago, a five gigabyte data-set would have exceeded the computational and storage capacities of most desktop computers and might therefore have been viewed as satisfying the volume criterion. In 2014, however, the storage and processing of a data-set of this size is essentially trivial. Indeed, in response to the IBM (2012) survey, over half of the IT and business professionals described big data as between one terabyte and one petabyte in volume. This hints at an alternative definitional paradigm in which big data is defined not in absolute terms, but rather *relative* to the prevailing technological and analytical capacity of the day. The meaning of big data, then, changes in lockstep with our ability to handle it.

Although many of our interviewees talked in terms of their data's volume, velocity, variety or veracity, their broader view of big data and its role within business appears to be of a more relative nature. Dr. Phil Mui, Chief Product and Engineering Officer at Acxiom, for example, defined big data with reference to "the access methods and manipulation technologies to make sense of the data", whilst Chris Nott, Chief Technology Officer of Big Data and Analytics at IBM UK, describes big data as an "evolution of capability". Indeed, there appears to be a fairly widespread view that businesses are undergoing a process of change that, while rapid, represents an evolution rather than a revolution. Many of our interviewees thought that the main opportunities created by big data have come from increases in the scale, speed or accuracy of existing processes rather than fundamentally new activities. As Basem Nayfeh, Chief Technology Officer at Audience Science describes: "The real shift is the processing. It's the ability to capture and store and then wrangle the data and to come back with some response in a reasonable amount of time". This sentiment was also reflected in the emphasis our respondents placed on the cost reduction dimension of technological progress in this area. Boris Mouzykantskii described the process as: "The amount of data which could be accessed [...] at a reasonable price point gets bigger and bigger." Where genuinely novel opportunities have arisen, firms appear to be involved in an ongoing process of experimentation. Nigel Davis of Willis Group, for example, remarked "We're [still] learning about the potential value of some of the newer data sources like social media and other feeds of data to different areas of our business."

This reserved optimism notwithstanding, our experts identified a number of ways in which big data and the attendant processing technologies have had a clear qualitative impact on what they are able to achieve. Many of these advances are related to the ability to link different datasets in a single analysis. In the past, data were often in silos—collected and analysed for single purposes due to the costs of storage and testing models. This has now changed. Mark Elliot of the Centre for Census and Survey Research at the University of Manchester says that "the thing that changes

everything for me is the fact that there are linkages. There are linkages between data and there are linkages from data to people. So the tie between ourselves and our data is much tighter and becoming increasingly so". Bringing together large data-sets allows for matching and connections that were not previously possible. Linkages between, for example, weather data and satellite images for the catastrophic modelling performed by the Willis Group, or online and offline purchasing behaviours performed by Tesco, potentially enable businesses to make better informed decisions.

A second important practice enabled by new technological capabilities is more powerful prediction. Paul Malyon of Experian says that "the main difference between big data and the standard data analytics that we've always done in the past is that big allows us to predict behaviour. Also, predict events based upon lots of sources of data that we can now combine in ways that we weren't able to before". Prediction is not a new phenomenon for commerce, but the difference with big data is the method of prediction. Traditional methods emphasised the "why" of behaviours or phenomena— e.g. why are more units of coffee sold in one region over another. The answer was then used to predict what would happen next. This represents a "theory first" approach to data analysis. Relatively small amounts of data were used to construct a theoretical understanding of the underlying process or behaviour which could serve as the basis for prediction.

Big data has brought a shift since the volumes and varieties of data often mean that prediction can be decoupled from understanding the underlying conceptual processes (Mayer-Schönberger & Cukier, 2013). Instead, esoteric patterns in the data can be used for forecasting the future without any intuitive or obvious reason for why the prediction should work as it does. For example, Google discovered that it was able to forecast flu epidemics ahead of official indicators purely by looking at traffic for a subset of search keywords (Ginsberg et al., 2009). These keywords were chosen for their correlation with the variable of interest rather than their semantic content. This kind of prediction comes with risks: Without an understanding of the underlying mechanism, predictions are vulnerable to changes in the underlying structure of behaviour or the environment that render the implicit model invalid. Recent failures in Google's flu prediction (Lazer, Kennedy, King, & Vespignani, 2014) are a case in point, and also highlight the general lack of transparency about how the data were arrived at, which presents a barrier to replicability. (It can be added that this obstacle is not insuperable: a subsequent study using Wikipedia to predict flu and other diseases was both more powerful and is open to replication since the data source is open; see Generous, Fairchild, Deshpande, Del Valle, & Priedhorsky, 2014).

In sum, the general sentiment among our experts was that the increased use of data is having a positive impact on their sector. Nevertheless, many business leaders do not see "big data" as a new phenomenon. Rather, it is perceived as being a continuation of a process by which companies seek competitive advantage or efficiency through the application of (data) science and technology. What's new is the scope of opportunity offered by big data, along with the cost-effectiveness for businesses of all sizes.

It should be emphasised that a consistent finding in our interviews was the crucial role that open data, often provided by governments or civil society groups, plays in facilitating these new opportunities. According to Tariq Khokhar, data scientist at the World Bank, "One of the biggest providers and creators of administrative data is government". Paul Maylon explained that "We often use open data as a basic building block or bedrock on top of which we layer other sources of data". More generally, it became quickly apparent that open data, especially from public sources, is absolutely foundational for many of the new economic opportunities being created through the intensive use of data in business. There is also evidence that this important role is being recognised by open data providers: Jeanne Holm, Evangelist at Data.gov, the US government's open data portal, observes that "We're seeing a whole sector that is looking at augmenting traditional services with open data to create either new services or smarter services within a company". Similarly, Susan Bateman, Head of Data Science at the UK Cabinet Office reports that a priority for her team is to think of ways to make the data more relevant for businesses.

cogent ·· social sciences

A number of our interviewees expressed the need for caution in embracing the promises of data or encouraging others to do so. Firstly, data, like any resource, has an associated cost and these costs should not be understated. Even though costs for data storage have decreased, they still must be balanced against benefit and value. There is a prevailing belief that more data are always better—allowing for improved predictive analysis. Jeremy Barnes, Co-founder and Chief Technology Officer of Datacratic, challenges this notion, asking "is the value of having that extra bit of information worth the price you're going to pay for it?" More generally, value is a consistent concern among our experts, yet how data are valued and extracted varies for each sector. Business strategy and purpose are considered critical determinants of how truly valuable particular data-sets are to a business.

Secondly, the benefits of big data—so hyped by the popular media—should not be overstated. Boris Mouzykantskii asserts that in truth, analysis is still a far way from predicting behaviours or tailoring advertising to an individual with an ideal degree of accuracy and personalisation. Personalised advertising, for instance, still relies on heuristic techniques such as segmentation—the categorising of people's predicted behaviours based on the aggregated behaviours of others with similar purchasing or viewing behaviours. Mouzykantskii remarks: "The online industry shot themselves in the foot. They basically overhyped internally their ability to learn from the data".

Thirdly, even the relatively mild form of audience segmentation referred to by Mouzykantskii may already seem invasive to some and society is still in the process of setting the boundaries of how and when personal data can collected and used. Simon Thompson of ESRI illustrates this point with a now infamous example. In 2012, Target, a US retailer of grocery and home goods, sent coupons for baby clothes and cribs to a 15-year-old girl having successfully predicted her pregnancy before her family knew. This incident has been widely used as an example of how invasive data analysis has become. In an ironic twist, as Thompson points out, the Target case also shows how far data analysis must still go: Target's analysis system obviously did not know the girl's age. A more recent case of invasiveness is the Facebook "emotional contagion" experiment, performed on more 700,000 Facebook users, without their knowledge, by changing the words in their newsfeed to see if they would react positively or negatively (Kramer, Guillory, & Hancock, 2014). This study has raised not just issues of research ethics, but also broader questions of whether this kind of research can condition people.

We return to the challenge of maintaining an appropriate contextual frame as data-sets grow in the conclusion. However, it is worth signalling already at this point that this study raised the issue not just of prediction and conditioning people, but also about the relation between academic researchers on the one hand, and on the other researchers and data from users in the private sector (in this case, Facebook): without going into details (the study is discussed in Schroeder, 2014), the relevance for business models is that there was a considerable public outcry (BBC, 2014; Guardian, 2014a, 2014b) and debate among researchers (Grimmelman, 2014; Schneier, 2015) which is still ongoing. It can be mentioned that researchers were in two camps: some argued that restricting this research among academics would only drive it underground to be pursued within private companies without making it open to public scrutiny (Meyer, 2014). Others (Schroeder, 2014) argued that transparency and replicability were important for science and that the larger issues about using big data to shape behaviour deserve a broader debate.

7. Big data business models

7.1. Data business models

It is easy to find an example of a company that might be considered a poster child for big data. Usually, such success stories will involve the innovative use of big data to deliver new products or to achieve large efficiency gains in some particular way. In this section, we step back from particular practices to take a more abstract view of the ways that data can serve as a central component of a business model and the opportunities that these imply. It should be stressed that these business models are not mutually exclusive and many of the organisations in our sample engaged with the data economy through more than one of these channels.

cogent • social sciences

7.1.1. Informing business decisions

Before discussing ways to directly monetise proprietary data, it is important to remark that in many instances data need not be directly monetised at all in order to have an appreciable economic impact. Indeed companies have used data internally to inform strategic decisions and refine business processes since long before data-enabled business became fashionable (Beniger, 1986). In these cases, data are used as an input into the management process. The effect of big data has been to amplify this practice. In the most advanced organisations, first-party data are used to inform internal business decisions on an extremely fine-grained scale. For business-to-business vendors such as Rolls Royce, the primary business model is the sale of equipment to clients such as Boeing or Virgin Atlantic. In the background, however, data collected via remote sensors installed on their equipment alert the company to maintenance issues allowing them to provide better service and informing research and development. Likewise, retailers such as Tesco and Starbucks have been pioneers in the use of reward cards to collect data about their customers and match their online and offline purchasing behaviours. These data also inform decisions about products, pricing, promotions, stock keeping and overall business strategy. While the primary business model for these companies is retail, first-party data inform many major decisions.

7.1.2. Data brokers

One obvious way to monetise proprietary first-party data is to treat it like any product and sell it to other parties. Thus, first-party data is treated as an output in its own right. A relatively pure example of this kind of business model can be observed in Nielsen, a market research company, which provides data and analysis on audience behaviours. Nielsen collects its first-party data using audience panels based either on its own research areas or contracted research. The business model for Nielsen is provision of data related to audience research in diverse formats based on client specifications.

In other cases, firms are discovering that data they generate through the everyday operation of their business can have a market value in its own right. Social media firms such as Twitter, for example, sell access to the data they host to third parties that use it for a variety of purposes such as market insight and sentiment analysis. Likewise, news organisations and online media platforms collect data from visitors to their websites. This first-party data primarily reflects web behaviours (searching, views, clicks, downloads and posts) and location and device information. While this data informs internal decision-making, website owners also act as data brokers and sell this data to third parties. They additionally have the option of collaborating with other businesses such as advertisers to run campaigns based around this data.

Nigel Davis, Analytics IT Director at Willis Group, helps us to understand why data brokers are so important to the evolving data economy: "It's a broadening spectrum of data that we wish to use and analyse, and the range of sources is ever increasing. These range from regular feeds of live data as web services hosted by companies and agencies through to statistics, demographics, and risk *datasets from an increasing number of third parties* [emphasis added]". As it becomes increasingly common to combine data from disparate sources during analysis, it becomes more likely that organisations will turn to third parties to supply that data. It is no more obviously efficient or practical for every organisation to collect its own data than it is for every organisation to drill its own oil or generate its own electricity.

Data brokerage is not new. Customer lists have long been viewed as valuable and marketable proprietary assets, while various companies have provided real-time stock price data from the floor of the New York Stock Exchange since the late nineteenth century. However, two important new trends are emerging. Firstly, as analytical techniques and computational capacity expand to encompass text and other forms of unstructured data, the scope of what constitutes data (and, therefore, of what constitutes a data broker) has grown. Since the complete text of a newspaper archive might now properly be regarded as data, a company holding such an archive can become a data broker simply by making access possible via an API. Secondly, there has been a large growth in the number of business activities, transactions, and interactions that are digitally mediated. This means that

data that would previously have been discarded or never captured in the first place is now stored digitally "from birth". Thus companies that had not been concerned with data find themselves in possession of data that may be of value to others.

7.1.3. Data analytics as a service

Many of our experts emphasised that the value of data lies not in its intrinsic merits, but rather in the actions resulting from analysis. "People want answers; they don't want more data", as Vivienne Ming of Gild put it. However, many organisations that are not data companies per se do not currently have the internal expertise or capacity to perform that analysis. As a result, a common business model for companies in the big data sphere is the provision of analytics as a service. The potential forms for analytics as a service are diverse. Examples in our sample range from large organisations such as Experian, which draws on massive data-sets to provide consumer credit scoring, to start-ups such as Gild, which recruits on behalf of technology firms using web data to profile potential employees. In any event, the defining feature of analytics-as-service businesses is that they take as an input data (its own proprietary data, data supplied by its client, some third party source of data or any combination of these) and produce as an output a data summary, analysis, insight, advice or some other product derived from that data.

Big data analytics are also becoming available for personal use with the rise of consumer-facing analytics firms. Mappiness, a mobile application that allows users to report their levels of happiness and receive feedback, collects personal data, analyses it and reports it in a usable form to users, whilst BrightScope offers retirement plan ratings and investment analytics based on a combination of open and personal data. An important part of the business model of firms such as Amazon and Netflix is also the provision of data-driven recommendations to consumers that both enhance the customer's experience and improve customer retention.

7.1.4. Consultancy and advisement

Fully realising the benefits of big data require expertise in technology, data analysis, business and organisational strategy, ethics and a host of other areas. Dozens of questions must be addressed in formulating and implementing a coherent data strategy. Examples include what is the best architecture for the physical data storage infrastructure, how should data workers be situated within a managerial hierarchy, what security protocols should be introduced to protect the integrity of the data and what is the appropriate ethical stance on handling personal data? However, just as some companies are not well-positioned to perform their own data analysis, others lack the in house expertise to tackle all dimensions of this strategic problem. This has given rise to an industry of firms, such as IBM, that provide consultancy and expertise on precisely these matters.

7.1.5. Tools providers

Storage media, servers and workstations, barcode scanners, statistical analysis and visualisation software, database software, remote sensors, encryption technology and networking equipment and many other examples of hardware and software constitute the tools of trade for a data-intensive business. The producers of these tools are therefore an important part of the big data economy. Examples of companies that provide big data tools include IPONWEB, a provider of infrastructure and technology for the online advertising industry, and ESRI, which provides a geospatial software analysis platforms.

7.2. A typology of big data business models

The big data business models described above can be grouped into three categories to yield a novel data business model taxonomy. The first category is what might be termed data users. These are organisations that use data either to inform business decisions, or as an input into other products and services such as credit reports or targeted advertising campaigns. These are organisations engaged in answering the question: how can data be used to create value within our business?

cogent ·· social sciences

Table 1. Big data business model typology		
Type	**Example functions**	**Dependencies**
Data *users*	Using data to inform strategic decisions; building data into products	Depend on *suppliers* for raw data, and on *facilitators* for infrastructure and skills
Data *suppliers*	Gathering primary data; aggregating and packaging data for sale	Depend on *facilitators* for infrastructure and skills, and on *users* both as customers and as sources of data
Data *facilitators*	Supplying infrastructure; consultancy; outsourced analysis	Depend on *users* and on *suppliers* as customers

The second class of business model encompasses data suppliers. These are organisations that either generate data that is of intrinsic value and therefore marketable, or else serve a kind of brokerage role by providing access to an aggregation of first- and third-party data. Such firms need not specialise in the supply of data. Indeed, many organisations are finding that they hold data that are of considerable value when some third party puts it to a use other than that for which is was originally collected. Since, like most information goods, the fixed costs of data production are usually high relative to the variable costs of distribution, there are potentially large efficiency gains from this kind of data reuse.

The third class of business model encompasses the range of activities that support third parties that are lacking in infrastructure or expertise. These data facilitators perform a range of services including advice on how to capitalise on big data, the provision of physical infrastructure and the provision of outsourced analytics services. These organisations are playing an especially important role during the current time of transition when a large number of firms are reorganising to make data more central to their business, but still lack the expertise or capacity to do so entirely internally.

The reason we believe our taxonomy to be useful is that, when viewed at this level of abstraction, it becomes clear that there is a substantial degree of interdependency between different classes of big data business model. Data users depend upon infrastructure and data that is supplied by data facilitators and data suppliers, respectively. Likewise, data facilitators play an important practical role in enabling the collection and aggregation of data by data suppliers. Lastly, both data suppliers and data facilitators are dependent upon an active community of data users to create a market for their products. Since the different types of big data business depend so closely upon each other for success, any policy directed at strengthening the data economy as a whole should take a relatively holistic stance, aiming to foster activities across the entire spectrum of big data business models. The typology is summarised in Table 1.

8. Challenges, obstacles, and policy suggestions

8.1. Data quality
As a practical matter, one significant day-to-day challenge faced by big data users is working with data of a generally low quality. Marwa Mabrouk, Cloud and Big Data Product Manager at ESRI estimates that "typically most data scientists spend between 75% and 80% of their time just cleaning up the data and moving it around and preparing it for analysis". Likewise, Jeremy Barnes, Co-founder and Chief Technology Officer at Datacratic estimates that "90% of the time is spent manipulating and transforming data and 10% is spent doing actual data science". This represents a significant overhead to data work stemming from inconsistencies in the formatting of different data-sets (e.g. if two data-sets store dates in different formats then one must be converted before the data can be merged), or because of generally bad practice in the way that data are collected and stored (e.g. Brian Lorenz, Vice President of Data at BrightScope, reported that historical retirement plan data are made available by the US government only in non-machine readable PDF form). These problems can be partly mitigated if the development of and adherence to consistent standards is encouraged at both the organisation and industry level.

8.2. Context, metadata and data provenance

Valid inferences can only be reliably drawn from data when the analyst has a thorough understanding of the data and the context from which it was drawn, but both are often lost as data-sets are increasingly combined and aggregated. The above-noted example of Target's pregnancy marketing faux pas is a case in point: that particular problem arose because data on purchasing habits were detached from a relevant contextual point (the subject's age). Mouzykantskii neatly summarised the problem thus: "There is no easy or standard way to keep metadata about what the data means together with data in a nice and searchable and consistent way. And that means that the knowledge of what the data actually meant gets separated from the data". The problems created by this lack of standardisation are magnified because they limit the ability of skilled data workers to move seamlessly between industries. Tim Davies, Open Data Research Coordinator for the World Wide Web Foundation, provides the example of working with National Health Service data: those working on health issues, but unfamiliar with NHS codes are likely to face a steep learning curve, while only those familiar with the sector's idiosyncratic practices are likely to "comprehend what the data was".

Indeed, the problem of context loss is closely related to the broader issue of metadata use and data longevity. In order for data to remain useful in the future (or to be useful to third parties) it is not only necessary that the data be readable, but also that it be documented in a transparent and consistent way so that all users understand what the data represents. However, our respondents reported that metadata is used inconsistently, if at all—raising the spectre of mistakes stemming from future misinterpretations of unlabelled data-sets. This problem is exacerbated in cases where large volumes of data, such as that created on the social web, is curated by users rather than a centralised, institutionalised data management authority because individual users are generally less likely to adhere to any standard labelling practice.

A third point is that, while data collection methods are typically well-understood for those collecting the data, the provenance of third-party data is often much more opaque. In fact, for companies such as Nielsen and ComScore that provide data and analysis based on audience panels, part of the proprietary dimensions of their business may be the formation of these panels. A few of our experts raised concerns about the media industries' shared reliance on these data-sets, asking who exactly the panel members were and what is really known about them. Daryl McNutt, Vice President of Marketing at Drawbridge, observed that even well-respected self-regulatory organisations in the advertising industry, including the Interactive Advertising Bureau (IAB) and the Media Rating Council (MRC) need to be more inclusive about the methods by which they arrive at their ratings. He opined that "There shouldn't be a black box or secret sauce. I think you have to do it in a way that is transparent so that people know there is real science and technology behind it". More generally, many businesses that use third-party data find themselves relying upon data sources without a complete understanding of how they were collected or generated.

8.3. Standards and accessibility

It is not only in the application of metadata that standard practices are lacking. There is a more general lack of standardisation in the way that data is stored and processed. A key theme in our analysis has been the importance of combining and linking data-sets to generate new combinatorial insights. But achieving this often requires that the systems responsible to collecting and processing that data are also linked. Our interviewees have described the nightmare of attempting to introduce an integrated data operation into an organisation with dozens of different computer and software systems, none of which were designed with compatibility in mind. Bret Shroyer of Willis Group describes the challenge thus: "We have no 'go to' tool. We have to think about how do we want to put this together, how are we going to connect it to our database, what sort of model are we going to build and it's a number of manual steps to get there". So long as a common standard that allows the interconnection of systems is absent, this will be a recurring challenge.

cogent •• social sciences

A related issue is that of accessibility. Where tools do exist, they are often designed for implementation and use by specialist data scientists or engineers. Columbia University's Dr. Cathy O'Neil notes "I want to think about the algorithm and not the implementation of the algorithm. I want to press a button, and ignoring costs for a moment, I want it to fire up as many machines on as large a grid as is necessary to do this computation within a given time limit. And I don't want to have to think about that too hard. And things like Hadoop, MapReduce, and other related platforms are a good step toward that. They basically make it possible to do huge calculations, but they don't make it easy yet". As the use of data becomes more pervasive in the economy, it is natural to expect that handling data will become a routine task for an ever larger fraction of the work force. But this can only happen if tools are developed that allow non-expert workers to perform tasks that are currently the preserve of specialists.

8.4. Internal politics

Company politics affect what data are shared internally, both between and within departments, as well as how data are shared with third parties. This very human element can create obstacles that technology alone cannot surmount. Decisions about how data are formatted for sharing and matching across datasets are critical and impact ease of later processing (c.f. the above discussions of data quality and metadata), but these decisions are often shaped as much by organisational structure and hierarchy as by practical or technical considerations.

8.5. The role of government

Our interviewees were concerned that industry expertise is not adequately represented in discussions of regulation, but perspectives varied. Some felt that the potential of big data has been overstated, resulting in uninformed panics. Others worried that decision-makers are not sufficiently informed about the various opportunities presented by big data and the practical reality faced by organisations that wish to take advantage of them. Heather Savory of the UK Open Data User Group remarked "The government should be providing the minimum regulatory infrastructure to allow things to work and allow for economic opportunity and deliver effective public services. It really shouldn't be interfering in businesses. What it should be doing is promoting the opportunities associated with using open data to people who might not have considered them". In short, there is a general desire for a minimum regulatory infrastructure combined with activities targeted at promoting the economic benefits of big data where these may not be well-known.

There was agreement that big data policies should be transparent, clear, fair and consistent. These are hallmarks of any good regulation, but merit special mention because there is a shared sense that the existing regulatory environment fails on a number of these counts. One area of particular friction surrounds the issue of privacy and personal data. The law has lagged behind both the growth in personal data use and developments in technical and statistical anonymisation techniques. There is also a lack of standardisation of privacy practices across jurisdictional boundaries. These failings are reflected in a somewhat piecemeal response to the personal data issues in industry, and there is still no accepted standard for how such issues should be treated—or even what the appropriate definition of personal data should be. Voluntary standards or codes of conduct, according to interviewees, would be a good first step given the likely intractability of a truly global privacy regulation. Germany was cited by Tariq Khokhar as a positive example of a country that provides strong privacy protection, but does so in a fair and transparent manner that also respects the needs of the business community. Yet the regulatory environment is currently in flux, with new European Data legislation on the horizon precisely because of some of the questions raised by new data sources, and other jurisdictions such as in the US similarly in need of updating (Pasquale, 2015).

9. Discussion and conclusion

If we combine our findings about types of business models (data users, facilitators and suppliers) with the challenges we have outlined, it is clear that those pursuing the three models will have quite different bottlenecks in going forward strategically: data quality affect all three, but addressing this issue may be costly for suppliers, may lead users to discount or factor in the reliability of data and

prompt facilitators to seek the best available sources. Similarly if we think of context, metadata and provenance: suppliers will have a strong incentive to provide more well-organised data, whereas users will rely on the best organised data and suppliers will need to bear the substantial costs of this better organisation. The organisational politics are only critical to users if it has a direct bearing on reliability, while suppliers will need to put mechanisms in place whereby they only allow access that keeps their competitive advantage in place, while facilitators need mechanisms so that their services can bridge data sources and skills with those for whom they are useful. This complex web of dependencies is bound to crystallise in the coming years, but it behoves businesses to ask themselves how they are placed to overcome these challenges, given that the type of business model they pursue is bound to be already largely determined by its capacities and resources.

Data quality also points to an emerging tension that is related to business models but takes us beyond them. Savage and Burrows (2007, 2009) (see also Kitchin, 2014) pointed out some time ago that access to big data provides an advantage for the private sector that is often not available to social scientists (though Wikipedia is an important counterexample). While this is an ongoing debate, it is also important to recognise that big data in academic research is different from big data in the private sector. A definition of big data (Schroeder, 2014; see also Cowls & Schroeder, 2015) can be provided for scientific research and relates to how data are a source for the validity of knowledge. However, big data for business purposes (and indeed purposes outside of scientific knowledge) is a different matter: e.g. the novel predictive capabilities that a number of our interviewees discussed as a key feature of big data approaches does not require scientific validity (if, based on geolocation data, someone is misled to the wrong shop offering discounts or if the prediction that I would like to buy a book on Amazon is wrong—these are errors that may not matter for being able to improve sales, though they would be unacceptable for academic publication). While in some cases there will be legal or consumer rights attendant upon misleading predictions (Pasquale, 2015), these are not the same issues as with big data for scientific validity, which may have few or no legal or rights implications in some cases (e.g. as with the analysis of Wikipedia). We are aware that there are some cases where such issues are involved—again, the Facebook contagion study provides a widely discussed example—but here the issue is when the company might use this knowledge to manipulate its users, while the issue of the scientific validity of the study (again) is separate. This also entails that the data quality issues mentioned by our interviewees only partly overlap with issues of data quality in the case of academic research. Another example here is Twitter, where limited access or access to how the data are collected is an issue for social science research (González-Bailón, Wang, Rivero, Borge-Holthoefer, & Moreno, 2014). However, analysing the same data for marketing purposes, for example, does not require the same standards of validity, even if, of course, marketing companies want their analysis to be as accurate as possible.

In other words, although several of our interviewees raised the issue of data quality and of understanding the context of data, these are practical issues separate from scientific issues, and also only partially overlap with regulatory issues arising from inaccurate data. Hence discussions of the validity of knowledge based on big data (Boyd & Crawford, 2012) and of the need for greater regulation (Pasquale, 2015) can suffer from insufficient differentiation between the types of data sources, their uses and aims (e.g. contributing to scientific knowledge or contributing to increasing sales). The value of data is increasingly being recognised as an asset, whether the data are proprietary or public (in which case they are likely to require cleaning or putting into suitable formats for analysis). This valuableness also raises new issues for business: among the main ones identified so far are the quality of the analyses (see on this point Cowls & Schroeder, 2015) or that the "black boxed" nature the analysis entails that less-than-transparent decisions are made (Pasquale, 2015), which may adversely affect customers and decision-makers.

An additional issue, however, are simply the sources of the data: if public data or data collected from the public are being used, there is a need for regulatory environments in which these uses can take place such that the benefits are regarded as benefitting those who provide the data. This could be states, whose taxpayers provide open data, or companies that provide free services in return for

cogent ∙∙ social sciences

customers providing data, or states and service providers (welfare states or insurance companies) who need data in order to provide services. In all these cases, more actors are engaging in more complex cost–benefit analyses, which require institutional environments which make these analyses predictable and transparent.

Finally, as concerns the role of government, it too, of course is a user of data, but mainly a supplier (see Reimsbach-Kounatze, 2015), and so its role can only benefit other users and facilitators and suppliers. At the same time, government is also in need of private sector data to enable informed policy-making. The coherence and incoherence in the regulatory and legal environment concerning data is therefore emerging as a critical issue, and urgently requires the attention of policy-makers. While this large topic is beyond the scope of this paper, it can be mentioned that there is, of course, a burgeoning literature on privacy laws beyond borders in relation to protecting data (Greenleaf, 2012, 2013; see also Rule, 2007). The importance of law and policy raises a point that was made earlier in this paper (see Section 2), which is that while a typology of big data business models and its implications may be mainly of interest to those in management and business studies, the topic cannot be confined to these disciplines: it must take into account legal and regulatory issues, and also spans questions of the methods and nature of knowledge within academic social science, and further the broader question of how knowledge outside of academia is applied in various contexts.

Hence, finally there is a tension which emerges from our interviews which was not apparent to interviewees themselves, but which has also so far not been apparent in the literature about big data and its social implications: as was discussed, many of the business models discussed here rely at least in part on open data-sets, and the primary concern of interviewees was that governments should enable the use of private and public data-sets by means of providing regulatory or legal frameworks ensuring the maximally productive use of data sources. In society-at-large, on the other hand, there are growing concerns over the use of big data, perhaps triggered by issues which have little to do with big data as discussed here per se (Wikileaks, the Snowden revelations). The main concerns in public discussions are related to uses of big data in social media research which is one of the main new sources of data, especially in academic research (again, the Facebook social contagion study has been the most widely discussed)—though it is a small subset of the data we have discussed here. Thus, we arrive at a tension related to big data business models insofar as these wider concerns, and their potential resolution, dominate public discussion, but are only to limited extent overlapping with the data sources we have discussed here. Big data using social media is only one source among several data sources that have been discussed here. Data apart from social media data also form the bulk of the data in the business models discussed here and the data that governments want to make available. There is moreover a third type of data which is neither social media data nor open government data: proprietary data held within firms, and this is commonly data that the three models we depict rely on (though typically in combination with open data promoted by government. It should be noted that we use open data here to mean publicly available data collected typically by governments. We do not mean data sought by civil society for greater transparency, [see https://okfn.org/opendata/], though the two sometimes overlap).

The space in the Venn diagram where these three types of data (social media data, open data promoted by government and proprietary data) overlap is small, but unless these different data sources, and especially their uses and implications, are disentangled, neither the business community nor the public is likely to be satisfied, with government caught amidst different discourses and unable to reconcile demands which conflate a number of data sources and policies about how privacy could be ensured in relation to these. These tensions are likely to affect the future of big data business models. How they can be resolved remains an open question.

Author details
Ralph Schroeder[1]
E-mail: ralph.schroeder@oii.ox.ac.uk
[1] Oxford Internet Institute, Oxford, UK.

References
BBC. (2014). *Facebook faces UK probe over emotion study*. Retrieved from online news website: http://www.bbc.co.uk/news/technology-28102550

Beniger, J. (1986). *The control revolution: Technological and economic origins of the information society*. Cambridge, MA: Harvard University Press.

Borgman, C. (2014). *Big data, little data and beyond*. Cambridge, MA: MIT Press.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication and Society, 15*, 662–679. http://dx.doi.org/10.1080/1369118X.2012.678878

Brown, I., & Marsden, C. (2013). *Regulating code: Good governance and better regulation in the information age*. Cambridge, MA: MIT Press.

Brynjolfsson, E., Hitt, L., & Heekyung, K. (2011). *Strength in numbers: How does data-driven decision making affect firm performance?* Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486

Bulger, M., Taylor, G., & Schroeder, R. (2014). *Engaging complexity: Challenges and opportunities of big data*. London: NEMODE. Retrieved from http://www.oii.ox.ac.uk/research/projects/?id=113

Clarke, A., & Margetts, H. (2014). Governments and citizens getting to know each other? Open, closed, and big data in public management reform. *Policy & Internet, 6*, 393–417. http://dx.doi.org/10.1002/poi3.v6.4

Cowls, J., & Schroeder, R. (2015). Causation, correlation, and big data in social science research. *Policy & Internet, 7*, 447–472.
doi:10.1002/poi3.100

Eagle, N., & Greene, K. (2014). *Reality mining: Using big data to engineer a better world*. Cambridge, MA: MIT Press.

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science, 346*, 715.

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., … Sugimoto, C. R. (2015). Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology, 66*, 1523–1545. http://dx.doi.org/10.1002/asi.2015.66.issue-8

Generous, N., Fairchild, G., Deshpande, A., Del Valle, S. Y., & Priedhorsky, R. (2014). Global disease monitoring and forecasting with Wikipedia. *PLoS Computational Biology, 10*, e1003892. http://dx.doi.org/10.1371/journal.pcbi.1003892

George, G., Haas, M. R., & Pentland, A. (2014). Big data and management. *Academy of Management Journal, 57*, 321–326. http://dx.doi.org/10.5465/amj.2014.4002

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

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks, 38*, 16–27. http://dx.doi.org/10.1016/j.socnet.2014.01.004

Greenleaf, G. (2012). *Global data privacy laws: 89 countries, and accelerating* (Research Paper No. 98/2012). London: Queen Mary University of London, School of Law Legal Studies.

Greenleaf, G. (2013). Data protection in a globalised network. In I. Brown (Ed.), *Research handbook on governance of the internet* (pp. 221–259). Cheltenham: Edward Elgar. http://dx.doi.org/10.4337/9781849805049

Grimmelman, J. (2014). *Personal website, with sources for the facebook emotional manipulation study*. Retrieved from *http://laboratorium.net/archive/2014/06/30/the_facebook_emotional_manipulation_study_source*

Guardian. (2014a, July 1). *Facebook's 'experiment' was socially irresponsible* (newspaper). Retrieved from http://www.theguardian.com/technology/2014/jul/01/facebook-socially-irresponsible

Guardian. (2014b, July 2). *Facebook apologises for psychological experiments on users* (newspaper). Retrieved from http://www.theguardian.com/technology/2014/jul/02/facebook-apologises-psychological-experiments-on-users

Kitchin, R. (2014). *The data revolution*. London: Sage.

Kramer, A., Guillory, J., & Hancock, J. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences, 111*, 8788–8790. http://dx.doi.org/10.1073/pnas.1320040111

Lane, J., & Stodden, V. (Eds.). (2014). *Privacy, big data, and the public good: Frameworks for engagement*. Cambridge: Cambridge University Press.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: Traps in big data analysis. *Science, 343*, 1203–1205. http://dx.doi.org/10.1126/science.1248506

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*. Washington, DC: McKinsey Global Institute. Retrieved from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. New York, NY: Houghton Mifflin Harcourt.

Meyer, M. (2014). Misjudgements will drive social trials underground. *Nature, 511*, 265.

Pasquale, F. (2015). *The black box society*. Cambridge, MA: Harvard University Press. http://dx.doi.org/10.4159/harvard.9780674736061

Pentland, S. (2014). *Social physics: How good ideas spread: The lessons from a new science*. London: Penguin.

Reimsbach-Kounatze, C. (2015). *The proliferation of "big data" and implications for official statistics and statistical agencies: A preliminary analysis* (OECD Digital Economy Papers, No. 245). Paris: OECD. http://dx.doi.org/10.1787/5js7t9wqzvg8-en

Rule, J. (2007). *Privacy in peril: How we are sacrificing a fundamental right in exchange for security and convenience* (Part IV). New York, NY: Oxford University Press.

Savage, M., & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology, 41*, 885–899. http://dx.doi.org/10.1177/0038038507080443

Savage, M., & Burrows, R. (2009). Some further reflections on the coming crisis of empirical sociology. *Sociology, 43*, 762–772. http://dx.doi.org/10.1177/0038038509105420

Schneier, B. (2015). *Data and Goliath*. New York, NY: W.W. Norton.

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., & Tufano, P. (2012). *Analytics: The real-world use of big data*. London: IBM Global Business Services Business Analytics and Optimisation in collaboration with Säid Business School, University of Oxford. Retrieved from http://www-935.ibm.com/services/us/gbs/thoughtleadership/ibv-big-data-at-work.html

Schroeder, R. (2014). Big data and the brave new world of social media research. *Big Data and Society*, 1–11.

Taylor, L., & Schroeder, R. (2014). Is bigger better? The emergence of big data as a tool for international development policy. *GeoJournal, 80*, 503–518. doi:10.1007/s10708-014-9603-5

Taylor, L., Schroeder, R., & Meyer, P. (2014). Emerging practices and perspectives on big data analysis in economics: Bigger and better or more of the same? *Big Data & Society, 1*. doi:10.1177/205395171453687

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

Publishing with Cogent OA ensures:

- Immediate, universal access to your article on publication
- High visibility and discoverability via the Cogent OA website as well as Taylor & Francis Online
- Download and citation statistics for your article
- Rapid online publication
- Input from, and dialog with, expert editors and editorial boards
- Retention of full copyright of your article
- Guaranteed legacy preservation of your article
- Discounts and waivers for authors in developing regions

Submit your manuscript to a Cogent OA journal at www.CogentOA.com