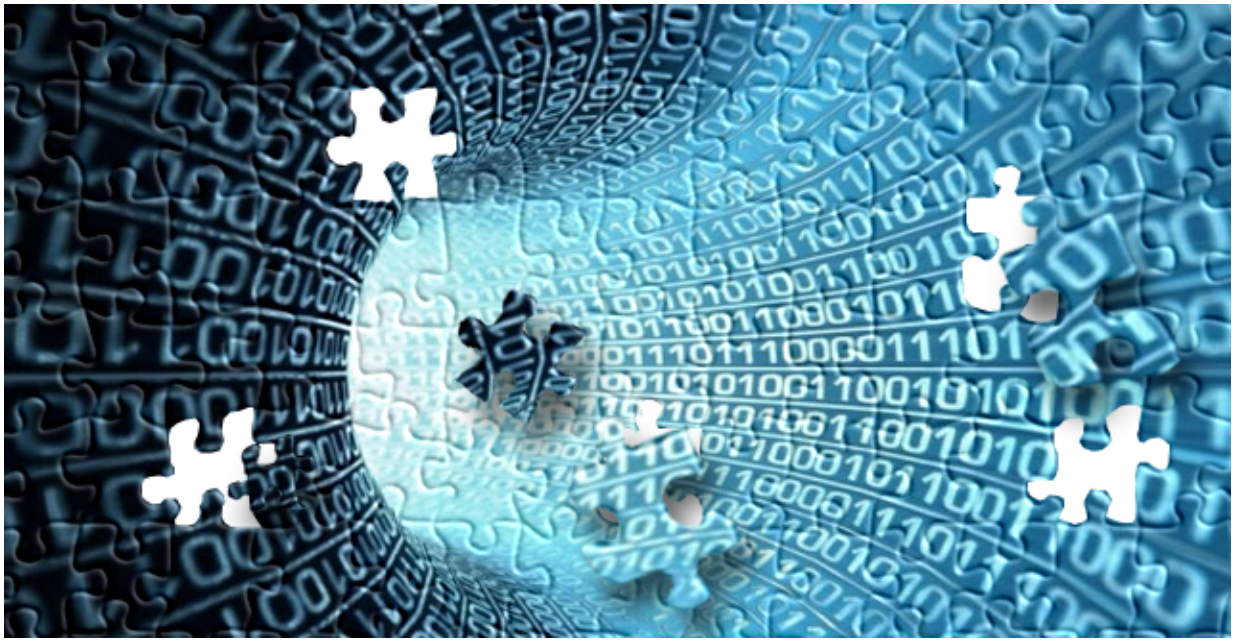

Data-Driven Business Models: Challenges and Opportunities of Big Data

Monica Bulger, Greg Taylor, Ralph Schroeder
Oxford Internet Institute
September 2014

RESEARCH COUNCILS UK
NEMODE
New Economic Models in the Digital Economy

Executive Summary

This report draws on interviews with 28 business leaders and stakeholder representatives from the UK and US in order to answer the following questions:

- How is (big) data being used; what is a 'big data business model'?
- What are the main obstacles to exploitation of big data in the economy?
- What can and should be done to mitigate these challenges and ensure that the opportunities provided by big data are realised?

There is optimism about profit potential, but experts caution that big data does not automatically lead to profits

- 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.
- Many of those organisations at the forefront of the data economy have had data analysis at the centre of their business for years or even decades, serving as a testament to the enduring benefits that being data-focused can bring. What's new is the scope of opportunity offered by big data, along with the cost effectiveness for businesses of all sizes.
- There is an awareness of the prevailing hype around big data. Many experts cite cautionary tales of how a focus on big-data to the exclusion of other considerations may lead to disaster. Data analysis in ignorance of the context can quickly become meaningless or even dangerous.
- Collecting and storing data is becoming much cheaper, but it not free. A cost-benefit analysis should be applied to all elements of a data strategy. That notwithstanding, businesses should take into account potential future uses of any data that might be collected, especially if a purely myopic cost-benefit analysis does not look favourable.

There is a real diversity of big data business models representing an interdependent data ecosystem.

- An important first step to realising the potential benefits of big data for business is deciding what the business model(s) will be.
- The data economy supports an entire ecosystem of businesses and other stakeholder organisations. These are often dependent upon each other's products and services so the vitality of the sector as a whole is crucial.
- Big data businesses can essentially be categorised as data users, data suppliers, and data facilitators. These are not mutually exclusive and many firms engage in a range of activities.
- *Data users* are organisations that use data internally—either for business intelligence activities such as forecasting demand, or as an

input into other products and services such as credit scores or targeted advertising.
    - They either generate data internally or acquire it from third parties (or both).
    - The key question for this group is “what data do we have and how can this data be used to create value within the business?”
    - A key challenge is assembling the physical infrastructure and skill base necessary to exploit big data.
- *Data suppliers* are organisations that supply data as a product for others to use.
    - This includes organisations that generate data of intrinsic value to others, and those that serve in a 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, some organisations are finding that, in the ordinary course of business, they generate data that is of value when some third party puts it to a use other than that for which is was originally intended (for example, many firms have data about their customers that is of value to marketers).
    - Governments are important suppliers of data, with most of the businesses in our sample using some form of government-produced data.
    - Key questions are what data is available, what uses might that data have (and for whom), and how should data be delivered to maximise its value?
- *Data facilitators* are organisations that help others to exploit data.
    - This includes businesses that provide outsourced analytics services and those offering either infrastructure or consultancy on data strategy.
    - This function is especially important during the transition phase because many potential users of big data do not have the skills to implement a big data strategy internally.
- Many business models focus on the use of data within and between businesses. However, as consumers are themselves increasingly faced with an abundance of data there is a growing market for businesses that provide data-driven products and services to end users.
    - The growing market for personal health and fitness devices, along with smart home technologies are pro-typical examples.
    - The whole spectrum of business models also apply in the consumer-facing segment: consumers will increasingly demand data analysis tools and services, data focused products, and help and advice for managing data their challenges.

There are key *practical* and *political* obstacles to the wider use of data

- The most commonly cited obstacles can be essentially divided into two categories: practical obstacles concerning data availability and quality, along with the necessary resources for analysis; and political obstacles that shape the environment in which data is used.
- An important practical obstacle is the quality of data sets. Some experts say analysts spend as much as 90% of their time cleaning data. Data, especially government data, is often provided in non-machine readable or non-standardised formats requiring manual re-entry.
- Past experience highlights the importance of being forward-looking in anticipating future uses (and users) of data. Legacy datasets that were not stored with appropriate human-readable metadata are now essentially useless because nobody knows what the data mean. A similar point holds for the format and physical infrastructure in which data is stored.
- Although technology has revolutionised data availability, there are still problems in providing access to data to those in an organisation who are best placed to exploit it. Tools to facilitate the use of data by non-specialists are an exciting prospect, but are still not mature enough to solve the problem.
- A key political barrier to data use is the extent to which people are protective of 'their' data. This often applies to a reluctance to share data within an organisation as much as to an unwillingness to share data between organisations, and speaks to the need for an organisation-wide policy and strategy for data use.
- There is widespread appreciation of the importance of privacy, but managers bemoan the lack of standards and clear policy guidance in this area.

Making the most of big data means having a clear business model and making it central to the business.

- Data should be central to the business. The biggest success stories have either essentially reinvented their entire business around the use of data or are 'born' data users.
- A clear profit model is essential. Experts warn that optimistically collecting data in the hope that it will somehow prove profitable is naïve. Managers and data scientists should be clear on the plan for generating value or efficiency from data before the data strategy is implemented.
- The most successful firms understand the limitations of the technology behind their big data operation and recognise the importance of combining analysis with a sound understanding of the context, a good intuition for the industry, and a critical attitude towards insights derived from data.

- Having the right skills is crucial. Very few individual posses the right combination of statistics, coding, and business skills, so cultivating talent in teams is essential.

The role of government and policymakers

- There was a prevailing sentiment that government intervention should be limited, providing a minimal regulatory structure while allowing for economic opportunities and delivering public services.
- Our experts expressed a hope that moving forward big data policies could be transparent, clear, fair, and consistent. In particular, the regulatory framework should be one that encourages businesses to innovate and compete on their own merits rather than 'picking a winner'.
- The largest obstacle to big data use is the often low quality of open datasets. Our experts recommend standardisation of codes, formats, and change management as well as accurate metadata to describe these codes.
- Provision of training in using large datasets, or in coding/structure of particular datasets would be helpful to both those currently using the datasets and to build capacity for a future workforce.

NEMODE

NEMODE is an initiative under the Research Councils UK (RCUK)'s Digital Economy (DE) research programme to bring together communities to explore new economic models in the Digital Economy. The project, which began in April 2012, has funding of £1.5 million over three years.

NEMODE focuses on New Economic Models, one of four RCUK DE Challenge Areas (the other three being IT as a Utility, Communities and Culture, and Sustainable Society). It aims to create a network of business people, academics and other experts who will identify and investigate the big research questions in new economic models that have arisen as a result of the digital economy's fast-paced growth.

The project aims to inform policy on issues pertaining to the digital economy and engage with SMEs to stimulate new ideas and the creation of new markets. NEMODE will also inform business sectors and large companies about the changes necessary to enable them to take advantage of the opportunities created by today's technologies. http://www.nemode.ac.uk

Oxford Internet Institute

The Oxford Internet Institute was established in 2001 as a multidisciplinary department of the University of Oxford, dedicated to the study of individual, collective and institutional behaviour on the Internet. Grounded in a determination to measure, understand and explain the Internet's multi-faceted interactions and effects, researchers are drawn from fields such as political science, economics, education, geography, sociology, communication, and law.

Closely engaged with industry practitioners and policy makers, the Oxford Internet Institute is a trusted place for empirically grounded, independent research on the Internet and society, as well as a space for genuinely neutral informed debate. http://www.oii.ox.ac.uk

This publication should be cited as: author names (2014). *Engaging Complexity: Challenges and Opportunities of Big Data.* London: NEMODE.

Cover image: Infocux Technologies, 2013
https://secure.flickr.com/photos/infocux/8450190120/

Cover design, image remix, report design: Monica Bulger

Correspondence should be addressed to:
Dr. Greg Taylor
Oxford Internet Institute
1 St. Giles
Oxford, OX1 3JS
+44 (0)1865 287210
enquiries@oii.ox.ac.uk

Authors

Monica Bulger is a Fellow at the Berkman Center for Internet and Society at Harvard University and a Research Associate at the Oxford Internet Institute (PhD, 2009) contributing policy research to national and multi-national groups such as UK Ofcom, U.S. Department of Justice, UNICEF, BBC, and the European Commission. Her research focuses on information use practices among industry practitioners, scientists, scholars and students to better understand how emerging technologies affect everyday practice. She has earned fellowships from the National Science Foundation, the Web Science Trust, National Writing Project and Oxford University Press.

Greg Taylor is an economist (PhD, 2010) and Research Fellow at the Oxford Internet Institute, University of Oxford who specialises in the microeconomics and industrial organisation of online markets. He has worked extensively on the economics of online search and online advertising, including issues of consumer trust, privacy, competition policy, and data-driven ad targeting—all important considerations in the data economy. He has also worked on auctions and market design, both of which have an important role to play in the organisation of economic activity in decentralised markets involving the pricing and trade of large volumes of data.

Ralph Schroeder is Professor at the Oxford Internet Institute and has been principal and co-investigator on ESRC projects and more than a dozen other funded projects, . He is currently co-investigator on a major two-year project funded by Sloan Foundation on ‘Accessing and Using Big Data to Advance Social Science Knowledge’. He has published several books and over 100 papers, including a forthcoming book with Eric T. Meyer on ‘Knowledge Machines: Digital Transformations of the Sciences and Humanities’ (MIT Press). His publications include a number of writings on economic sociology and on innovation.

Expert participants

Romina Ahmad, Technical Project Manager, Government Innovation Group, Data.gov.uk, Cabinet Office

Jeremy Barnes, Co-founder and Chief Technology Officer, Datacratic

Susan Bateman, Head of Data Science, Government Innovation Group, Cabinet Office

Tim Davies, Fellow, Berkman Center for Internet and Society and Open Data Research Coordinator, World Wide Web Foundation

Nigel Davis, Analytics IT Director, Willis Group

Dr. Mark Elliot, Senior Lecturer, Centre for Census and Survey Research, University of Manchester and Disclosure Expert, Administrative Data Liaison Service

Brian Ferrario, Vice President of Marketing, Drawbridge

Devin Guan, Vice President of Engineering, Drawbridge

Jeanne Holm, Evangelist, Data.gov, U.S. General Services Administration

Tariq Khokhar, Data Scientist, World Bank

Bryan Lorenz, Vice President, Data, BrightScope, Inc.

Marwa Mabrouk, Cloud and Big Data Product Manager, Environmental Systems Research Institute (ESRI)
(shortly after her interview for this research, Marwa Mabrouk joined Google in the role of Product Manager)

Dr. George MacKerron, Lecturer in Economics, University of Sussex and Creator and Founder, Mappiness

Paul Malyon, Head of Business to Business Transactional and Marketing Products, Experian

Alastair McCullough, Business Intelligence Competence Centre Strategy Leader, Europe, Strategy & Analytics Practice, IBM Global Business Services

Daryl McNutt, Vice President of Marketing, Drawbridge
(shortly after his interview for this research, Daryl McNutt joined Adaptive Media in the role of Chief Marketing Officer)

Dr. Patrick McSharry, Senior Research Fellow in Business Analytics, University of Oxford

Dr. Vivienne Ming, Vice President of Research and Insight, Gild

Dr. Boris Mouzykantskii, Founder and CEO, IPONWEB

Dr. Phil Mui, Chief Product and Technology Officer, EVP, Acxiom

Dr. Basem Nayfeh, Chief Technology Officer, Audience Science

Dr. Cathy O'Neil, Founder, Mathbabe.org and Program Director, Lede Program in Data Journalism at the Columbia Journalism School

Chris Nott, Chief Technology Officer, Big Data & Analytics, Executive IT Specialist, IBM UK and Ireland (UKI)

Heather Savory, Open Data User Group

Bret Shroyer, Senior Vice President, Reinsurance, Willis Group
(shortly after his interview for this research, Bret Shroyer joined the predictive modeling firm Valen Analytics in the role of Solutions Architect)

Dr. Janet Smart, Reader in Operations Management, Saïd Business School, University of Oxford

Dr. Arthur Thomas, CEO and Founder, Proteus Associates

Simon Thompson, Director of Commercial Solutions, Environmental Systems Research Institute (ESRI)

Analysts in Seattle and London who requested anonymity

Acknowledgements

The authors wish to thank all those who gave their time to be interviewed for this report or provided background context for big data use. They also thank Dr. Janet Smart and Dr. Arthur Thomas who generously provided background information and assistance in developing our questionnaire. Special thanks to RCUK and NEMODE for funding the project and to our expert respondents for their feedback during the course of the research.

Table of Contents

Chapter 1
Big data's evolving economic importance

Increasingly, big data is proving to be an essential element of economic growth. An ever-expanding body of evidence points to the crucial role of big data across sectors in informing strategic decisions. A joint study of 1144 IT and business professionals by IBM and the University of Oxford's Saïd Business School found a 70% increase between 2010 and 2012 in those who reported that using big data afforded them a "competitive economic advantage". Perhaps more promising is big data's potential, with McKinsey Global Institute (Manyika, et al., 2011) predicting a massive contribution to the economy: a 60% potential increase in operating margins for retailers, $300 billion annual value to US healthcare, and 1.5 million new positions for 'data-savvy' managers. According to a 2013 study, the geospatial industry is currently estimated to contribute $150-270 billion in annual revenues for products and services[1]. A recent report of big data and public sector information in the UK describes favourable outcomes for integrating data into daily life: "data allows us to adapt and improve public services and enhance our whole way of life, bringing economic growth, wide-ranging social benefits and improvements in how government works" (Shakespeare, 2013, p.5).

[1] Prepared by Oxera Consulting in the UK on behalf of Google http://www.oxera.com/Latest-Thinking/News/January-2013/Oxera-quantifies-the-benefits-of-Geo-services-to-g.aspx

Across the private, public, and civil society sectors, big data is becoming an essential component of business operations. The UK and US governments confirm the pervasiveness of big data use in their extensive listings of companies who use government-supplied open[2] public datasets. The US Open Data 500[3] lists approximately 480 companies across 16 sectors including technology, finance, insurance, healthcare and business as well as those who do not feature as prominently in public discussions of big data such as food/agriculture, housing, and education. When considering the significance of big data for the economy, this list reflects important trends: many on the list are international and most of the top companies for each sector depend upon government-supplied open data. Likewise, the UK's Open Data initiative at data.gov.uk has published detailed case studies of 118 companies across the private, public and civil society sectors, demonstrating the significant contribution of big data, particularly open public data, to the UK and global economy.

The definition of big data is a contentious issue. In this report, we have in mind a definition similar to that provided by Tim Davies, Fellow at the Berkman Center for Internet and Society and Open Data Research Coordinator at the World Wide Web Foundation: 'big data' is data on a significant level of complexity and scale that cannot be managed with conventional analytic approaches. This remains (deliberately) vague and, for the most part, we allow subjects to talk about big data as *they* understand it. The objective of this study is to construct a picture of how practitioners see data affecting business and policy practice, rather than to impose an *a priori* definition of big data from the top down. Alastair McCullough, Business Intelligence Competence Centre Strategy Leader, Europe for IBM, for example, describes it thus: "[Big data is about] dealing with information management challenges that don't naturally fit within what we might otherwise see as traditional approaches to handling these problems".

A natural approach might be to define big data by some measure of its 'bigness'. However, as discussed further in Chapter 2, size benchmarks for big data continue to change as storage technologies become cheaper and analysis techniques more sophisticated. While initially defined as data that could not fit on a spreadsheet, the definition has progressed to data that cannot fit on a single computer, and both definition and data measurement will likely continue to expand. Big data tends to be available in real time using a combination of multiple types of data sources with the potential for predictive power not previously possible.

1.1 Benefits of big data

Economic benefit is a clear promise of big data (Schroeck, et al., 2012; Manyika, et al., 2011). Across academic and industry studies, gains in productivity, competitive advantage, and efficiency are found (Bakhshi, Bravo-

[2] In this report, we use a definition for 'open data' provided by Paul Malyon, Head of Business to Business Transactional and Marketing Products at Experian. Please see section 3.1.2.
[3] US Open Data 500 GOVLAB. Retrieved from http://www.opendata500.com/

Biosca & Mateos-Garcia, 2014; Brynjolfsson, Hitt, Kim, 2011; Experian, 2013). Economic gains reportedly evident in risk prediction for insurance companies, resource management for cars and jet engines, and better predictions of customer behaviour for retailers such as Starbucks, Tesco and Burberry seem to portend a promising future (Soudagar, 2013; Swabey, 2013; Wheatley, 2013). Along with increased profitability and efficiency, big data is predicted to increase demand for a highly skilled, data-literate workforce (Bakhshi & Mateos-Garcia, 2012; Gild, 2013; Manyika, et al., 2011; Silicon Valley Bank, 2013). As the UK and US governments increasingly provide access to large open datasets (e.g., NASA Landsat, USGS weather data, census data, UK maps, transport conditions, crime statistics) and additionally improve online access to government services (e.g., health, transport, law enforcement), there is a sense that big data will also promote civic engagement. Additionally, use of big data in scientific research is informing understandings of climate trends, diffusion of epidemics, and agricultural yields, potentially providing predictive power for management of existing resources and isolation of potential threats.

1.2 Data evolution

> "I don't think anybody really talks about small data any more…anything which is data is now big data." –*Dr. Boris Mouzykantskii, IPONWEB*

A recurring theme in the interviews conducted for this study is that experts perceive big data not as a discrete phenomenon, but rather as existing on a continuum of innovation. Indeed, since the beginnings of civilisation mathematics, statistics, and accounting practices have co-evolved with innovations in business and economic activity. Historical evidence suggests that some of the earliest developments in mathematics, for example, were tied to the desire to measure quantities of goods, wealth, the passing of the seasons, and land holdings—all important considerations for newly emerging agrarian societies. These pieces of data represented a scale and complexity that could not be managed with existing technologies and, by some definitions, might therefore be considered to have been big data. Since that time, the need to precisely identify available resources, pinpoint their demand, and accurately determine profit per unit has pushed the evolution of increasingly sophisticated metrics and a shift from reporting to predicting value.

Despite this apparent continuity, there has recently been an increase in the rate of innovation and adoption tied to advances in information technology and telecommunications. Industry experts describe a convergence of technology, capacity, and demand as providing the conditions for this dramatic change in data use. IBM's study (Schroeck, et al., 2012) identified the 4 V's of big data: volume, variety, velocity, and veracity:

- *volume*: (amount of data) as digital data grew, so did the amount of data generated; simultaneously, data storage technologies improved, enabling larger amounts of data to be stored more cheaply, thus data

that in the past may have been discarded can be saved and incorporated into analysis
- *variety*: (forms of data) in the recent past, data primarily existed in static spreadsheets, usually reporting on past events, such as completed advertising campaigns or purchases during a specific time period; data has shifted from structured (e.g., spreadsheets) to unstructured (e.g., maps, video, email), data formats that are more reflective of everyday activities
- *velocity*: (speed of data) with improvements in technology, data can now be collected, shared, and analysed more quickly, often in real-time, reducing the gap between when data is generated and when it becomes actionable in decision-making
- *veracity*: (reliability of data) data quality is a challenge regardless of scale; however, increases in volume, variety, and velocity exacerbate this issue, with burgeoning datasets that need to be rigorously evaluated for their accuracy, provenance, relevance, and consistency.

Underlying the V's are significant shifts in technology development coupled with movements in business strategy from focusing on causal relationships to predictive analysis (Gild, 2013; O'Neil, 2014; Mayer-Schoenberger & Cukier, 2013). When data storage was expensive and size of storage limited, data was collected more selectively and decisions made prior to the collection process. As storage capacities expanded and simultaneously became less expensive, larger datasets could be collected and stored, allowing for more options and flexibility in analysis. Consequently, datasets continue to expand. As Eric Schmidt, Executive Chairman of Google famously said in 2010, "There was 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days, and the pace is increasing."[4]

While much information is generated, many industry experts urge caution when estimating the usefulness of that data (O'Neil, 2014; Ebbert, 2012). In Chapter 4, we address the challenges faced by analysts as they use and merge large datasets. With such rapidly increasing output of data, it is unsurprising that data scientists are, in some instances, being overwhelmed. In Chapter 6, we explore whether big data is outstripping human capacity to analyse and make use of it.

1.3 Aims of report

In witnessing a marked shift in business practice toward more predictive, rather than causal analysis, it is important to understand the drivers, benefits, obstacles, and contexts of this phenomenon. This study aims to examine how big data is used in everyday economic contexts and how this use creates value. We sought to identify potentials for commercial use of big data, and pinpoint gaps and untapped opportunities.

[4] http://techonomy.typepad.com/blog/2010/08/google-privacy-and-the-new-explosion-of-data.html

To accomplish a better understanding of the practicalities of big data use, we identified seven key topic areas:

- how experts define ‘big data’
- the reality of big data use versus the hype
- typology of business models
- potential obstacles to the use of big data
- qualities of companies well-situated to make use of big data
- skills necessary to meaningfully interpret large datasets
- predictions of big data use and influence in the next 15 years

A key goal of this research is to identify areas where government interventions in data provision, regulation, or skills development may facilitate the full realisation of big data’s benefits.

1.4 Maximising the potential of big data: A challenge for governments

Much evidence points to dependence on public information, yet past reports have identified the (low) quality of public data as an obstacle to its use (Maude, 2012). In this study, we determine the extent to which companies use government-supplied datasets and the dynamics of this dependence.

Discussions of data use necessarily address concerns around provisions for privacy. Through our interviews, we determine how companies feel about self-regulation, their attitudes toward international regulatory frameworks, and their professional practice related to privacy provision.

Does the government also have a role in supporting national efforts to prepare for a growing demand for workers skilled in data science? We ask our experts to identify necessary skills for working with big data and evaluate whether a gap exists and recommend how it can best be filled.

Everyday use of big data in commerce can reveal areas of untapped potential as well as areas of need. This research seeks to move beyond the hype and promise and consider how governments can support the realities of everyday data use.

1.5 Methods

To better understand everyday use of big data and its future directions, the following methods were employed:

- A desk review of academic studies, industry research, and news articles to identify industry thought leaders.
- Interviews with 28 industry experts from a range of sectors using a 10-item questionnaire asking about business models, challenges in collecting and using data, characteristics of businesses poised to take advantage of big data, skills needed, and ways government can promote productive use (see “Expert Participants” section, p. 8 for participants and Appendix 2 for interview questions).

Experts were identified through their prominence in industry publications and conferences, recommendations from other experts, and renown in their respective fields. A few participants have written about commercial use of big data, and their publications are referenced in this report. 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 30-90 minutes.

Eight women and 20 men participated in the study and represented 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. In addition, researchers engaged in

informal background discussions with analysts in Seattle, Silicon Valley, and London who requested anonymity. Experts from the public sector include 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 and the World Wide Web Foundation.

Companies included in the study represent a range of sizes, from Mappiness with nearly 60,000 users 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 fourteen companies included in the study, nine are headquartered in the U.S. (four of which have offices in London), four are headquartered in the UK and one is based in Canada.

Chapter 2
What is 'big data'?

In popular culture, representations of big data range from films such as *Moneyball*, in which baseball player's performance statistics are used to accurately recruit high-performing teams, to *Minority Report* in which the lead character fights an all-knowing government and law enforcement that can predict decisions and behaviours before they occur. More recently, *Her* shows a personalised computer system that learns feelings from human interaction. But what are the practical realities of big data?

While many definitions exist for big data, most, as described in Chapter 1, address its size, scale, variety, reliability, and speed. Yet in speaking with our experts, they focused just as heavily on the technologies making this processing possible, including innovations in storage and analysis capabilities. In fact, most of our experts find the interesting part about big data isn't necessarily its size, but its analytic possibilities. Dr. Phil Mui, Chief Product and Engineering Officer at Acxiom says, "'Big data' also refers to the size of available data for analysis, as well as the access methods and manipulation technologies to make sense of the data. Big data promotes a culture of data-informed decision making. It is now much harder to justify a decision based solely on 'intuition' or 'gut' without some data analysis to back it up" (Ebbert, 2012). Chris Nott, Chief Technology Officer of Big Data and Analytics at IBM UK describes these innovations as an "evolution of

capability" and Simon Thompson, Director of Commercial Solutions at ESRI views them as "the emergence of a new normal."

Unlike most media treatment of big data, our experts emphasised that working with large datasets—and their attendant processing challenges—is not new. According to Tariq Khokhar, Data Scientist at the World Bank: "The techniques have existed for quite some time. It's more about applying some of these techniques to these problems. The innovation is the combination of having the tools, having the data, and applying a technique and tweaking techniques to work on these particular problems." Thus, while *big data* centres upon large and ever-growing datasets, it is a combination of innovation in storage capacity, processing speeds, and analysis demands.

2.1 Volume

In response to the IBM 2012 survey, over half of IT and business professionals described big data as between one terabyte and one petabyte, while 30% could not define "how big 'big' is for their organization" (p.4). Responses from our experts ranged from anything that does not fit on a spreadsheet to sizes similar to those reported by IBM. Tariq Khokhar joked, "If you can weigh it, it's big." Yet, when considering, for example, that desktop computers can handle terabytes of data and a move to larger storage options requires significant hardware, data size can be realistically estimated according to weight or physical space used for storage; large technology firms typically measure their data centres in square feet rather than terabytes. Amazon's presence in the data storage space reflects the growing demand and profit potential. Dr. George MacKerron, Lecturer in Economics at University of Sussex and founder of Mappiness, similarly evaluated the size of data based on its storage, describing his company's dataset: "it is a few dozens of gigabytes, it fits on a normal MacBook, you can run your analyses in stages on a normal MacBook." Attempts to define big data purely in terms of volume, though, are inherently problematic because constant technological change brings with it a changing frame of reference: computational tasks that appeared big a decade ago are now all but trivial.

As businesses expand well beyond data that can fit on a laptop or local server, storage requirements become a significant consideration. Even though costs for data storage have decreased, they still must be balanced against benefit and value. Further, as datasets grow, so do requirements for analysis processing, raising further the costs for using data. There is a prevailing belief that more data is 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 is valued and extracted varies for each sector. Business strategy and purpose are considered critical determinants of how truly valuable particular datasets are to a business.

> "Often the value is reducing the size of the problem down rather than bringing more and more data to it." –*Jeremy Barnes, Datacratic*

In addition to volume, the "bigness" of big data could easily refer to its scale. When discussing datasets, Dr. Basem Nayfeh of Audience Science estimated that his company processed well over 17 billion data points per day, and growing very quickly. Similarly, Jeremy Barnes estimated that Datacratic processes up to 200,000 data points per second for some of their customers.

Alastair McCullough, Business Intelligence Competence Centre Strategy Leader, Europe, for IBM explains the increase in scale from a data storage perspective, explaining that most transactional (structured) data can be stored on a database. The next data storage stage could be said to be a data warehouse and then potentially an enterprise data warehouse. He explains that at each stage scalability and complexity increase:

> " …if you start with a database, this is a transactional home for data so you're typically storing financial or other operational transactions. If you look at a warehouse you're no longer interested in transactions per se, you're interested in the interconnectivity between the data and also in structuring the data in a completely non-transactional way to derive new knowledge from them. So the data warehouse is not transactionally-oriented, it's developed and designed to deliver outcomes based on the types of questions people ask of it; fundamentally different."
> –*Alastair McCullough, IBM Global Business Services*

Coupled with McCullough's technical description is an emphasis on purpose. As the data storage capacity expands, so do the analysis capabilities, but choice to scale up is linked to expected outcomes dependent upon the perceived value created.

2.2 Velocity

In IBM's (2012) report, *velocity* is defined as the "time between when data is created or captured, and when it is accessible" (Schroeck, et al., p.4). In Gild's *Big Data Recruiting Playbook* (2013) *velocity* is tied with outcomes "…information that occurs rapidly and needs to be addressed rapidly" of which tweets, Google searches, Facebook posts, or retail transactions are data sources (p.6). Ads for cheap flights appearing on one website immediately after a user searched for travel information on another represent the immediacy with which advertisers respond to web behaviours. Similarly, reward card users at Starbucks or Tesco may receive an email containing promotional offers immediately following an in-store transaction.

More sophisticated examples occur in natural language processing. Imagining the depth of data necessary to enable the translation of natural language in

search queries, the IBM Watson presents an extraordinary case for real-time processing of big data. According to its Wikipedia entry, Watson had access to 200 million pages of structured and unstructured content, representing approximately 4 terabytes.[5] In 2011, Watson won *Jeopardy!* an American quiz show by defeating two of its past champions.[6] Likewise, Technabling, a Scottish company, and Microsoft Kinect have both developed systems to translate sign language into text and vice-versa in real time, with a goal of enabling deaf and non-deaf users to communicate.[7]

Real time processing allows for fraud detection in the finance sector, catastrophe modelling in the insurance sector, and responsive campaigns in the retail and advertising sectors. In our interviews, experts from the advertising sector describe the improvement in response time enabled by big data versus traditional methods. Prior to improved analysis and reporting technologies, expense and technical limitations resulted in *ex post* evaluation of completed campaigns, so the focus was typically on cause and effect. Moreover, the long time lags involved often led to the introduction of confounding factors that interfered with identification of causality. In contrast, high-velocity data is a key enabler of real-time, instantaneous experimentation. Common practices such as A-B testing allow businesses to immediately identify the effects of a change and respond accordingly. In the *Harvard Business Review,* Wes Nichols, CEO of MarketShare, an analytics company, describes this shift as an "unprecedented ability" for marketers "to fine-tune their allocation decisions while making course corrections in real time."[8] Big data now enables, for example, High Street stores to determine location-specific in-store specials or coupon delivery to specific postal codes or area codes. These agile campaigns have flexible strategies that adjust as new data indicates success or failure of particular approaches.

2.3 Variety

Perhaps the clearest difference represented by big data versus traditional approaches is the variety of formats included in analyses. Our experts explained the differences between structured and unstructured data, yet emphasised that analysis of the latter is still relatively experimental and that analysts are in the early stages of exploring its potential. *Structured data*— for example, transactional data (sales, payroll, accounting, purchasing)—can easily translate into the columns and rows of a spreadsheet. *Unstructured data* can be anything that is difficult to quantify, that does not easily fit into columns and rows. Examples include Youtube videos, satellite imagery, social media interactions, or email messages. Unstructured data includes a variety of formats and are collected from across devices (e.g., laptops, mobile phones, sensors, satellites).

[5] http://en.wikipedia.org/wiki/Watson_%28computer%29
[6] http://researcher.ibm.com/researcher/view_project.php?id=2099
[7] http://www.wired.co.uk/news/archive/2013-07/18/sign-language-translation-kinect
[8] http://hbr.org/2013/03/advertising-analytics-20/ar/1

The value of varied data formats lies in the potential for a larger scope in which to understand or predict particular behaviours. Participants from IBM provided a few scenarios for illustration. In the first, if an organisation wishes to investigate internal cyber fraud for a client, analysts can employ analytics techniques to reduce false positives and prioritise cases for further investigation, first assessing internal files that logged network behaviours in an attempt to use insight to more accurately identify anomalies, or behaviours that seem out of the ordinary. Once suspicious actions are identified, analysts can then combine these findings with external sources such as social media to build a richer profile to better determine involvement.

In a second scenario, a retailer could develop a better understanding of its customers by interacting across different channels, including social media, internet, mobile, and inside the store. In one case, customer loyalty cards are used to link in-store purchases to a credit card and online purchases. Another retailer encourages customers to engage with its robust social media offerings in an effort to create a personalised experience that extends from online interactions to in-store shopping.

2.4 Reliability (Veracity)

While data diversity offers much potential, our interviews uncovered significant concerns about the reliability of these data. When asked about obstacles and challenges in data collection and use, data quality was most frequently mentioned. Datasets from multiple sources may use different values to measure the same variable. Often, as in the case of Nielsen audience data, methods of collection are proprietary, so specifics about sampling and response prompts are not transparent. In other cases, proxy measures are used to approximate something else, for example web browsing to approximate interest or gym membership to represent attitudes toward health.

In the scenarios provided in the previous section to demonstrate the potential afforded by data variety, it is clear that while the aim is to enhance customer experience or pinpoint specific anomalous behaviours, increasing availability of data formats and sources can create noise or distractions in the data. In fact, “bigger” does not necessarily mean “better.”

> “If you’re working with multiple large datasets, you’re not necessarily taking two or three big datasets and putting them all together and making everything even bigger. That’s not necessarily the smart way to get value out of data. What happens if you make it bigger and bigger and bigger, is that you don’t focus on what you’re trying to do. I think the smart way to use big datasets is often to extract different pieces of a jigsaw puzzle that you put together separately.”
> *–Heather Savory, Open Data User Group*

In her book *On Being a Data Skeptic* (2014), Dr. Cathy O'Neil illustrates challenges associated with choosing and relying upon different types of data. She compares the recommendation engines of Amazon and Netflix. Amazon relies upon transactional data to recommend purchases and rentals. The transactional data includes what those who made purchases or rentals similar to a particular customer also purchased or rented. The recommendation system used by Amazon additionally uses viewing history to predict interest in further purchases. Netflix, on the other hand, uses a recommendation engine that relies on what those who reviewed a film also reviewed, instead of relying upon a customer's rental history. O'Neil points out that Netflix therefore relies on a limited population of those likely to review, potentially weakening the strength of its predictions.

Several dimensions of data reliability must be considered when choosing data sources and defining purpose of analysis. In Chapter 4, we address these issues in more detail.

2.5 Processing power

> "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." –*Dr. Basem Nayfeh, Audience Science*

When discussing big data, a majority of our experts emphasised the convergence of technologies allowing for increased data storage at cheaper prices, improved processing/analysis, increased generation of data from unstructured sources, and the analytical capability to now process them. These improvements led to a flattening of processing time and processing costs, enabling real-time reporting of, for example, in-store transactions or satellite weather data.

Discussed in further detail in Chapters 3 and 4, improved processing reduces the costs of testing theories. Increased storage capabilities mean that decisions about which data points will be used can be made after collection; since data are now less expensive to store, there has been a shift from traditional planning. In the recent past, data were often discarded or decisions about which data were collected were carefully made in consideration of storage limits.

> "The amount of data which could be accessed, not necessarily processed, but at least accessed at a reasonable price point gets bigger and bigger. In terms of number of terabytes of information it certainly gets much, much bigger."
> –*Dr. Boris Mouzykantskii, IPONWEB*

Dr. Nayfeh observes that "What had been deleted or shoved somewhere onto a backup now became valuable because we could actually process it." Nayfeh further explains that with less data, analysts must be cleverer in their

algorithms because they are "trying to conjure up what's happening based on a small amount of signals."[9] Moreover, the analysis that can be conducted is only as good as the decision of which data to process in the first place. Recent advances in storage and processing capabilities mean most of these decisions can now be made after data collection and more experimental analyses are therefore possible. More processing may mean more signals, but our experts report that these signals are often weak and therefore challenging to find—though without the improved processing power, these weaker signals would not likely surface at all.

Alastair McCullough noted the improved analytics possible with stronger processing, describing it as advancing beyond "the capability to slice, dice, build, pivot, collapse and represent data" to derive the kind of insights that drive business decisions.

2.6 Linkages

> "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."
> *–Dr. Mark Elliot, Centre for Census and Survey Research, University of Manchester*

Bringing together large datasets 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. The opportunities created by finding these connections in datasets is described in further detail in Chapter 4.

Prior to recent advances in data processing and storage, data were often in silos, collected and analysed for single purposes due to the costs of storage and testing models. Descriptions from our experts liken the ongoing innovations in processing and analysis to an opening of floodgates in terms of what is possible. Analysis of large datasets enables the identifying of relationships between datasets, for example, weather and spending patterns or car purchases and marital status. Bryan Lorenz of BrightScope describes the benefits of linking large datasets to drive business decisions, such as in what geographic areas a company should target its sales force at or comparing where a product is weak or strong in relation to the performance of competitors. In this sense, big data represents an improvement over

[9] *Signals* refer to the meaningful, relevant information content of the data. *Noise, by contrast,* usually refers to confounding factors that obscure the relevant patterns, or to data that may be related but does not improve analytical power. In either case, noise makes finding and interpreting signals more difficult and error prone.

traditional analyses—enabling analysts to detect patterns and trends that were previously hidden `in the cracks' between datasets.

2.7 Predictive power of big data

> "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." –*Paul Malyon, Experian*

Prediction is not a new phenomenon for commerce. The difference big data represents is the method by which prediction occurs: In traditional methods, the emphasis was on the *why* of behaviours or phenomena—for example, why are more units of coffee sold in one region over another. Drivers for answering the question of *why* were then used to predict what would happen next. This represents a fundamentally 'theory first' approach to data analysis. Relatively small amounts of data are used to construct, at least implicitly, a conceptual or theoretical understanding of the underlying process or behaviour, which can then serve as the basis for prediction. For example, electronics retailers might predict a spike in demand for televisions during major sporting events based upon past experience and a conceptual idea of how and when consumers are likely to want to use these devices.

Big data has brought a shift in emphasis. The large volumes and varieties of data often mean that prediction can be decoupled from understanding of the underlying conceptual process. 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 well 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, and were selected in complete ignorance of any health data. This kind of prediction, though, comes with its own risks. Without an understanding of the underlying mechanism, predictions are vulnerable to changes in behaviour or the environment that render the model invalid. Recent failures in Google's flu prediction are a case in point (Lazer, et al., 2014).

2.7.1 Case study: Predictive analysis at Willis Group

With large datasets, behaviour or phenomena can be analysed en masse and trends identified that predict future actions. Bret Shroyer, Senior Vice President of Reinsurance for the Willis Group described the changes in data types and analysis outcomes for reinsurance decisions. He said that in the past, reinsurance would focus on 'core placement data': the premiums and losses for a given contract. In the past five years, the Willis Group has assembled a larger body of data to help predict covered losses. In addition to the core placement data, Willis now includes characteristics of the insured's

homes, past losses associated with the properties, past losses associated with the insured parties, how long insured and information about losses that will not be covered by the contract, since these losses may potentially predict losses that will be covered.

> "We've started aggregating data across multiple clients to try to draw some correlation, where we might have thin data from one client we can use data from our consortium or our pool that will help us to better predict the price for a new client or that client with thin data."
> *–Bret Shroyer, Willis Group*

Drawing from varied data sources over time enables stronger predictions for covered losses, potentially enabling better-informed decisions.

2.7.2 Case study: Predictive models for climate change

On a related topic, Dr. Patrick McSharry, Senior Research Fellow in Business Analytics at the University of Oxford, describes linking and combining data on hurricanes to forecast the potential for property losses due to hurricanes under different weather scenarios.

> "Rather than taking a catalogue of events from the past, you generate events that you think are going to be more representative of what's going to happen in the future."
> *–Dr. Patrick McSharry, University of Oxford*

For Dr. McSharry, increased processing power allows for more computationally complex models, providing the capacity to iterate and test models more cheaply than with traditional systems. To measure future risk posed by hurricanes for insurance companies, he combines historic data on hurricanes making landfall in a particular region hurricanes making landfall in a particular region with dynamic climate models.

2.8 Overhype

Few would object to improved personalisation if it meant, for example that the barista at a major coffee chain knew a patron's preferences. However, most would object if the cashier at Boots or Walgreens inquired whether a recent prescription was effective. Personalisation, the holy grail for most advertisers and retailers, is exactly the site of conflict between businesses and their target markets.

"Recently, a flight attendant walked up to a customer flying from Dallas to Houston and said 'What would you like to drink? And, oh, by the way, I am so sorry we lost your bag yesterday coming from Chicago.'"
–Ian Ayres, Super Crunchers*, 2007*

The example from *Super Crunchers* indicates a linked and fluid system by which an airline would be aware of each passengers' unique flight experience and needs, which could be viewed as a courteous benefit or invasive add-on. Yet although these examples of personalisation are commonly portrayed in discussions of big data, they may not be accurate. How many times has a flight attendant addressed a passenger's misfortunes from an earlier flight?

Perhaps the current capacity for linking data and predicting behaviours has been overstated. Dr. Boris Mouzykantskii asserts that in truth, analysis is still a far way off from predicting behaviours or tailoring advertising to an individual. He explains that the public's assumptions about big data are "misinformed simply because they don't yet deeply understand what exactly is happening and what sort of data is collected."

> "The online industry shot themselves in the foot. They basically overhyped internally their ability to learn from the data."
> *– Dr. Boris Mouzykantskii, IPONWEB*

Dr. Mouzykantskii further explains that, despite exponential growth in computing power, more data is currently collected than can be analysed. 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. For example, a person who conducts web searches for five-star hotels and expensive cars will be put in a luxury segment and likely receive advertising for expensive watches. Likewise, a person who searches for nappies and infant toys will be put in a parent segment and likely receive advertising for family cars and baby food. Although the digitisation of transactions has provided data that greatly facilitates this kind of estimation, it remains some way short of the ideal of truly personalised advertisements.

While even this limited level of personalisation may seem invasive, Simon Thompson of ESRI uses the now infamous example of Target in the U.S. to illustrate how crude data analysis remains. In 2012, Target, a retailer of grocery and home goods in the U.S., sent coupons for baby clothes and cribs to a fifteen year old before her family knew she was pregnant. The predictive analysis that resulted in the coupon mailing was based on buying habits of those enrolled in Target's baby registry. Using their purchase and search patterns, analysts at Target created a list of 25 products that could indicate a woman was pregnant (regardless of registry enrolment), such as unscented lotions and pre-natal vitamins.[10] This incident has been used as an example of how invasive data analysis has become. Yet Thompson points out that, in actuality, the incident shows how far data analysis must still go: Target's analysis system obviously did not know the girl's age.

[10] http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Despite many misassumptions about big data, experts are convinced that it represents an evolution in mathematics, business strategy, computational processing, data storage, and business analytics. Marwa Mabrouk describes big data as being in a hype phase and not simply a fad. Dr Mark Elliot echoed a few of our experts' belief that big data is "akin to the birth of the Internet and the birth of the personal computer."

For the most part, our experts balanced positivity with caution, noting that the value of any data is a combined result of business strategy and domain expertise. They warned that big data alone is not a panacea for economic challenges, but rather represents a significant shift in business practice that, as addressed in Chapter 5 can serve those companies well situated to maximise its benefits.

Chapter 3
Business models for big data

Apparent in our analysis is that the successful long-standing businesses we studied engage in diverse uses of data, incorporating multiple data forms from a variety of sources and offering a range of services, include platform as service and analysis as service. Long-established companies such as IBM and Willis Group demonstrate a history of iterating. IBM presents a particularly diverse trajectory, expanding from storage and backbone services to power analysis, analytical software, and strategic advisement. Similarly, Willis Group has expanded from its work as an insurance and re-insurance broker (covering infamous disasters such as the Titanic and Hindenburg) to predictive analysis that incorporates data from a range of sources including Landsat (from NASA) images, weather data, and customer information.

These examples serve to demonstrate how the complexity of big data and the possibilities it offers inspire multi-dimensional and evolving business models. While some of the businesses included in our research, such as Gild, have straightforward business models (e.g., identifying and recruiting high quality talent), many business models in the advertising, tech, and financial sectors are multi-layered. These organisations often have more than one way of taking advantage of big data so there is no such thing as "the" big data business model. Experts interviewed for our study mentioned fluidity, adaptability, and flexibility as essential for businesses in these sectors to survive. Those who thrive seem to combine the provision of proprietary

analysis platforms with bespoke analysis services. However, client services are only part of the large web of data use.

3.1 Understanding the business of big data

There is no single blueprint for 'how to make money from big data'. Yet the companies we studied offer promising and profitable models from which much can be learned. In our analysis of big data use broadly in commerce and more specifically within the companies included in our case studies, we identified two dimensions for analysis:

- data types and sources
- business models for how data are monetised

With innovative uses constantly emerging, examples within these categories are not comprehensive, but organise practices into key decision stages at which critical choices about data use affect costs and profit.

3.1.1 Data types and sources

Data are defined by how they are acquired. Across sectors, data are drawn from multiple sources. While sectors such as retail may rely on transactional data more heavily than the health or insurance sectors, they share a dependence on diverse datasets from a mix of open and proprietary sources.

This diversity of sources is reflected both across our sample and in the broader data economy. Retailers such as Tesco and Starbucks use a combination of transactional data collected in-store and online, demographic data, and geographic data, among other data sources. These data inform in-store campaigns, promotions sent via email or text to customers, and even decisions about where to open new stores. Gild, a tech hiring company, analyses data from over 80 open source, professional, and social media sties, including GitHub, Google Code, Stack Exchange, and LinkedIn. Gild scours the web for developers, uses predictive technology to automatically evaluate their abilities by examining actual coding work. Likewise, Willis Group draws from a variety of data sources, including satellite imagery, river gauging stations, earthquake sensors, and storm forecasts to analyse risks presented by catastrophic events.

> "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."
> *–Nigel Davis, Willis Group*

How can one categorise these varied sources of data? Although diverse and evidently complex, experts from finance, retail, media, and advertising

describe data in three categories according to acquisition points and ownership:

- *First party:* data collected about a business' customers, for example, transactional data, demographic data collected directly from customers; data owned by the business (data owned by the company)
- *Second party:* data collected in collaboration with another company, for example, when running a campaign through Google AdWords; data collected by one company on behalf of another; unclear who owns the data, though generally the company collecting the data stores it (e.g., Google) (derived data, ownership uncertain)
- *Third party:* data that is collected by someone else, for example government datasets, Acxiom and Experian credit scores, Nielsen audience ratings (data owned by someone else)

For the most part, *first party* data is closed. Data collected by companies about their customers is generally for commercial purposes, for either in-house analysis or sale to third parties. It is not publicly shared or freely available. In raw form, this data might contain personally identifiable information (PII), linking an individual to, for example, purchases or occupation or income data. In our interviews, this data was referred to as "private," "proprietary," "closed," "not open," and "protected."

An analyst in Seattle described the extensive lengths to which PII data are protected, saying that even within a company access is limited or restricted. A post to Quora in 2011 by Joe Sullivan, Chief Security Officer at Facebook, describes the types of safeguards that are consistent with our interview findings:

> "We take our role as stewards of people's information very seriously and have invested heavily in protecting the data trusted to us.
>
> There is no 'skeleton key.' In fact, we have advanced internal tools that restrict access to information to only those employees who need it to do their jobs (e.g., investigating user reports). There is a cross-functional group of employees who work on these safeguards and appropriate access to these tools.
>
> Most employees do not have access and, those who do, must sign an agreement and complete a training program before using our internal tools. Finally, we track the actions performed through internal tools.
>
> Each use is logged and requires the employee to explain the purpose of his or her use, and we audit all of this regularly."[11]

We found these safeguards to be common among those companies in our study working with PII data, though most participants work with data that has been "cleaned" of PII, reporting data in aggregate. Importantly, "PII" does not have a single agreed-upon definition across companies or sectors and approaches to safeguards therefore vary.

[11] http://www.quora.com/Freddy-S-Wolens

Second party data is also typically closed. These data are shared between companies per contractual agreement specifying a scope of time and behaviours. These agreements usually include access to data collected by one party via a website and shared with another and usually involve a campaign (e.g., selected search terms) in which the second party reports response results (e.g., how many page views, how many clicks, how many times a particular search term was entered). Those we interviewed said that this data is very useful, yet ownership is often uncertain given that is collected and stored by one company on behalf of another.

At the *third party* level, diversity of sources and data types can markedly increase. Third party data can include data collected from the web (web scraping), or data collected and processed by data vendors such as Experian or Acxiom. In interviews, Nielsen's audience ratings were used as a well-known example of third party data: often methods for collection are unclear, so the reputation of the company is critical and while the party purchasing the data receives results, sampling procedures and details of the data collection are not transparent. Third party data vendors package data according to client demand, providing information on specific target groups.

Third party data includes open or public data as well as data of a closed or commercial nature.

The sources from which data are drawn are an important first component of a business model taxonomy because different sources of data feed into or emerge from different business models. First party data can be viewed as an output that may have value in its own right and can serve as a basis for proprietary competitive advantage. Most firms produce some form of first party data, but having a big data business model requires that this production activity be coupled with a monetisation strategy or a plan for how the data can be used to generate value elsewhere in the business. In contrast, third party data is an (often costly) input. Business models then revolve around how to use such an input to create new products or enhance business efficiency in novel and innovative ways.

3.1.2 A note on open data

> "We often use open data as a basic building block or bedrock on top of which we layer other sources of data."
> *–Paul Malyon, Experian*

The above quote from Paul Malyon is indicative of a broader trend: A consistent finding in our interviews was the crucial role that open data plays in facilitating a broad range of business activities, with businesses adding commercial or first party data along with data collected via second party exchanges to create proprietary products or provide business intelligence. 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.

> "In terms of open data, the most common way to think of it is
> - information owned by the state or public sector,
> - information that is available for free,
> - reuse carries very little licensing, at most attribution, but carries no restrictions on onward use,
> - tends not to feature personally identifiable information,
> - at the moment, is mainly government information or government owned data."
>
> *–Paul Malyon, Experian*

In 2011, Deloitte Analytics estimated that the UK government made available 37,500 open datasets to businesses.[12] Similarly, the US government reports 90,925 datasets available on their www.data.gov website. Datasets are available from a range of governmental departments across sectors including labour, health, transport, environment, arts, education, sports, and science.

> "One of the biggest providers and creators of administrative data is government."
>
> *–Tariq Khokhar, World Bank*

While a portion of government data is sold in the UK,[13]much is freely available as open data. It is difficult to quantify the exact value of such data to businesses, but Ordnance Survey estimated a £100 billion contribution to the UK economy from its datasets. Further, in a September 2013 blog post, Heather Savory, Independent Chair of the Open Data User Group in the UK estimates that "were Ordnance Survey data to be provided under a less restricted commercial licensing regime, external third party revenue (in the UK geospatial and data sectors) would increase by around £370 million." She reports that more data are becoming open in response to user requests. Recently, the Met Office (weather data) and Land Registry (land and property ownership data) have released new portions of their collections as open data.

> "We're looking at where are the instances where big and open data might actually happen so we can begin to categorise it."
>
> *–Romina Ahmad, Data.gov.uk*

[12] http://www.deloitte.com/assets/Dcom-UnitedKingdom/Local%20Assets/Documents/Market%20insights/Deloitte%20Analytics/uk-da-open-growth.pdf

[13] Some public data are provided by trading funds such as Ordnance Survey (maps) or the Hydrographic Office (navigational products and services) in return for a fee. The UK Government funds the Ordnance Survey to provide its data for public sector use and to deliver some open data. Other commercial users have to pay again to use this important core reference data. So, traditionally, these datasets are considered public, but not open, with datasets available for a fee with licensing restrictions.

Susan Bateman, Head of Data Science at the Cabinet Office reports that a priority for her team is to think of ways to make the data more relevant for businesses.

Bateman and colleague Romina Ahmad of data.gov.uk describe efforts to maximise open datasets by considering the range of potential commercial applications. Environmental data proved a useful example because datasets can be applied to a range of purposes such as geospatial analysis or climate modelling. These methods, in turn, have broad applicability in insurance, retail, construction, advertising, and many other sectors.

Jeanne Holm, Evangelist at Data.gov works with businesses across the U.S. to encourage use of government supplied open datasets. Holm reports two primary uses of big data in commerce. First, for larger companies such as Walgreen's, a pharmaceutical retailer in the U.S., open data such as government-funded clinical trials are used to improve predictive power for performance of their inventory.

Secondly, Holm describes a trend for start-up companies and entrepreneurs to use open data to address civil society-focused issues, such as improved access to services or ratings of local service providers.

> "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."
> *–Jeanne Holm, Data.gov*

3.2 Business models: How data are monetised

The business of big data is complex, with most companies engaging in multiple dimensions of collection, processing, and sale of data. Emerging in this market are "as a service" models in which companies provide software, platforms, analytics, or consulting "as a service," engaging their customers in multiple ways. Thus, companies such as ESRI or IBM that may also offer standardised platforms or software also develop bespoke systems and analytics for clients "as a service." Likewise, companies such as Experian or Acxiom sell datasets and offer tailored analytics "as a service."

While a range of services and analytics exist around big data in commerce, here we identify core business models that reflect how data are monetised across sectors.

3.2.1 Monetising first party data

3.2.1.1 Inform business decisions

Before discussing strategies 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. In these cases, first party data (and, potentially, second or third party data) is used as an input into the management process.

The effect of big data has been to radically amplify this practice. In the most advanced organisations, first party data is 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 alerts the company to maintenance issues allowing them to provide a higher quality of service, and also informs 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 informs almost every significant decision.

3.2.1.2 Data brokers

One obvious way to monetise proprietary first party data is to treat it like any other 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 headquartered in the UK, 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, 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 and generate second party derived data in which both parties are engaged in aspects of the data collection but the website owner is responsible for collection and reporting. For the website owner, the sale of this data is part of a business model and for the second party business, the data is used to inform strategic decisions regarding campaigns, or additional analyses are performed and the derived data is sold again.

Data brokerage is, again, not new. Customer lists have long been viewed as valuable and marketable proprietary assets, whilst various companies have provided real time stock price data from the floor of the New York Stock Exchange since the late 19th 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'.

This may be a significant untapped source of future value for big data. A broad variety of organisations that are not data companies per se nevertheless find themselves in possession of data that may be of value to others. Since companies collecting data in the ordinary course of business incur the cost of doing so irrespective of whether that data is made available to third parties or not, any additional use to which that data is put has the potential to be a pure efficiency gain.

3.2.2 Data analytics as a service

> "People want answers; they don't want more data."
> *–Dr. Vivienne Ming, Gild*

Many of our experts emphasised that the value of data lies not in its intrinsic merits, but rather in the actions resulting from analysis. 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 service*. In this model the analytics firm takes as an input its own proprietary data, data supplied by its client, or some third party source of data, and produces as an output a data summary, analysis, insight, advice, or some other product derived from that data.

For many of the firms in our sample, data analytics is part of their business model. Analytics often result in reporting insights on a client's targeted audience *segment* based on aggregated behavioural data for particular groups. Acxiom and Experian, for example, draw on massive datasets to profile consumers and thereby facilitate targeted advertising or provide consumer credit scoring. Similarly, Audience Science, an advertising company providing a management platform for advertisers, and Datacratic, a machine learning and artificial intelligence company providing software for digital marketing, use first, second, and third party data to develop audience segments for clients.

Again, we see the common theme that additional value can be unlocked when data from multiple diverse sources is combined. Compiling user data from multiple devices, for example, enables more sophisticated analysis of audience behaviours. Drawbridge, a technology company based in Silicon Valley, analyses application content to link an individual user across multiple devices. This data enables an understanding of audience behaviour between screens and provides insights for cross-screen audiences, resulting in the creation of detailed segments for advertiser and retailer use.

The potential roles for analytics as a service are diverse. Innovative start-ups are specialising in using data to provide advice or insight with respect to virtually every business functions. A unique example is Gild, a recruiter of highly skilled workers, which uses data collected from select websites to create profiles of potential candidates which are then analysed to find a strong fit for their clients.

As processing and analysis technologies become less expensive and as consumers are more frequently coming into direct contact with data, big data analytics are also rapidly becoming available for personal use. Thus, a new breed of consumer-facing analytics firms are emerging. 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. While Mappiness is currently available at no charge, other applications within the quantified-self movement, such as Fitbit charge users for initial equipment or subscription to an analysis service.[14] BrightScope, Inc., offers retirement plan ratings and investment analytics. They combine open data supplied by the U.S. government, such as the Form 5500 and associated retirement plan audit reports, with quantitative metrics thus providing retirement plan ratings to aid customers in their investment decisions. An important part of the business model of firms such as Amazon and Netflix is the provision of data-driven recommendations to consumers that both enhance the customer's experience and improve customer retention.

3.2.3 Consultancy and advisement

> "We're looking at the ability to derive statistical insights and analytical insights, so typically quantification, risk, value, profitability, revenue; and people are looking for business trending, seasonality, and fluctuation insights to come out of such analysis."
> *– Alastair McCullough, IBM Global Business Services*

Realising the potential of big data for businesses remains a challenge. While the potential to accomplish the analytical expectations Alastair McCullough

[14] For an overview of the quantified self movement, see http://www.technologyreview.com/featuredstory/424390/the-measured-life/
For several examples of quantified self business models, see descriptions at http://techcrunch.com/tag/quantified-self/

describes is clearly evident, movement from envisioning a business model to fulfilling it requires expertise in the technologies, data analysis, and business and organisational strategy for big data. However, just as some companies are not well-positioned to perform their own analysis, others lack the in house expertise to position themselves to take advantage of big data in the first place. Therefore, a valid business model in this emerging area is the guidance of companies in realising the potential of big data for their particular circumstance. This type of advisory activity has at least two facets.

Firstly, from a technical side, strategic advisement can address the actual technical structuring of data within a company, its information architecture. In addressing how knowledge is structured, McCullough describes a process of evaluating data management from a technical perspective, considering storage options such as a data warehouse or the need for bespoke versus standardised processing platforms, typically making use of one or more carefully-researched and client fieldwork-based and tested technical Reference Architectures to drive design of these platforms in a detailed and substantive manner. Strategic advisement may also mean resolving politics and process within a company around data structuring and management, particularly assessing performance data to identify inefficiencies in data management and in the coherent and effective (re-) organisation of the enterprise to address cross-functional operations encompassing people, process and information technology to support real-world big data and analytics delivery. McCullough describes the additional challenges of legacy systems which report data that are only machine readable and therefore present interpretation challenges, or systems that discard rather than store essential data. These examples illustrate the everyday challenges of big data management—especially for large organisations where data has not been the historical focus—and the very human component of navigating related decisions.

Secondly, decisions related to the incorporation of data into overall business strategy additionally prompt advisement models. Datacratic sells a prediction and optimisation service that its customers use in real-time to analyse customer-provided data combined with second and third party web data.. Likewise, Audience Science advises its customers on marketing strategies using data from its advertising platform. Proteus Associates, a consulting firm in Oxford, engages in risk assessment to advise biomedical firms on strategic directions. Likewise, Willis Group provides risk assessments to insurers by drawing upon several data sources in combination with the insurers' property portfolios.

Understanding how big data can add value to a business remains challenging. Even established businesses report that their use is evolving and questions relating to monetising emerging datasets remain open, particularly related to incorporating data generated by social media. Thus, strategic advisement around the applications of big data is a growing area.

3.2.4 Monetising data processing

Evident in this chapter is that near every aspect of big data is attached to a potential profit model. The expanding technologies around the generation, management, processing, and storage of big data each provide a framework for business opportunities. However, sophisticated data analysis operations require equally sophisticated hardware and software infrastructure, and a range of companies have emerged with capability to supply this 'physical' infrastructure.

Drawing further upon the example of IBM's strategic advisement, often the technical dimensions of resulting recommendations require bespoke processing solutions. The business model for IBM is multi-dimensional and includes the supply of storage and processing platforms to complement its advisement offerings. IPONWEB, a provider of infrastructure and technology, develops bespoke platforms for the online advertising industry. These platforms are comprised of the technical backbone to support massive data processing as well as customised algorithms to extract and analyse relevant datapoints.

As described earlier, ESRI provides a software analysis platform for geospatial data while also developing bespoke solutions for its clients. Across sectors, Audience Science and BrightScope also offer platforms for their clients as well as customised modifications to optimise customer outcomes.

As increasingly larger datasets are generated, data storage space presents an additional business model. Companies such as Amazon, Google, and Rackspace globally offer storage space as part of their multi-dimensional business models.

3.2.5 Summary

Clearly, the dimensions of big data that represent a significant shift over traditional methods—its size, speed, variety, the ability to connect across datasets and follow individual users across media—are the very elements that present challenges in unlocking data's potential. Strong profit models are already pursued and realised, yet even well-established businesses admit a steep learning curve and continue to explore unanswered questions related to the untapped business potential of big data. A key finding in our interviews is the persistent nature of big data as it moves into nearly all sectors, causing a marked shift in business models, particularly the practice of extracting value from data. These challenges have engendered a variety of organisations that support the deployment of data through the provision of advice, infrastructure, or outsourced services.

In sum, we have observed three distinct types of big data business models. The first 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. Important questions facing such companies are what data do we create, what data do we need to obtain externally, and how can this data be

used to create value within our business? These businesses require the physical and human resources to take advantage of the data. However, success stories such as Tesco illustrate the benefits that can be realised when a business places data at its centre.

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 is 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, the potential efficiency gains from this kind of data reuse are large. Irrespective of whether a firm is a specialist data supplier or not, key questions are what data is available, what uses might that data have and for whom, and how should data be delivered to maximise its value?

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 internal expertise to do so without assistance.

Chapter 4
Obstacles to using big data

Despite its lauded potential, using big data is a messy process, with solutions still in the piecemeal stage for leading innovators. Conceptually, connecting datasets holds much promise, but the realities involve extensive cleaning and normalising of data, with variables as simple as telephone numbers being represented in multiple ways (e.g., with and without area code). Additionally, there is the problem of missing data and the challenge of quantifying phenomena such as reputation or risk, which may defy measurement. While tools exist for processing, most experts say they are constantly building and customising tools to address new processing needs.

Many of these issues are commonplace in industry, but not much discussed outside of it. Through our interviews and literature reviews, we explore these shared challenges, identifying those that potentially affect productivity and profitability or may create obstacles for businesses, expected or otherwise.

4.1 Data quality

While datasets are increasingly available across sectors, many are developed for a task within government or a particular organisation and do not seamlessly lend themselves to analysis by a third party. 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.” These observations reflect the practical reality that a large proportion of time is spent in preparing data for analysis, rather than actual analysis.

To some extent, this reflects the necessary and natural cost of data-centric business activity. However, many of our experts drew attention to circumstances in which better management of data handling, storage, and processing, or better planning for the ongoing role of data in a business can mitigate these costs.

4.1.1 Consistency

Why is the proportion of time spent on analysis eclipsed by preparation and “cleaning” of data? The health of a dataset is critical to reliable analyses. Analysts report that different methodologies may be used to measure the same thing, for example wealth or poverty or energy consumption. Dr. Vivienne Ming, VP of Research and Insight at Gild describes the challenge of “de-duplication,” that arises when merging multiple datasets that measure the same element.

On a practical level, the actual values residing in the columns and rows must work within whatever analysis platform is used. For example, when using the number “100,000,” it must be consistent with other values in the system, which may mean adding a space “100 000” or removing the comma “100000.” While seemingly simple, consistency in how values are represented is a critical and foundational concern, from a formatting perspective as well as the larger issue of how data are collected and quantified.

4.1.1.1 Change management

Cleaning datasets can be automated in some cases, once specific issues are found. However, even when organisations share a contractual agreement in which one party delivers data to another, seemingly small changes may go unreported and create serious data consistency challenges for analysts. Paul Malyon, Head of Business to Business Transactional and Marketing Products for Experian acknowledges the difficulty in anticipating and managing these changes: “With commercial data, you tend to find out in advance if a supplier is expecting to change their data at all, but you can get caught by unexpected things no matter what.”

These changes can seem very small, for example a software update that changes how a URL string is reported, or moving data from field 305 to field 307. Dr. Boris Mouzykantskii, Founder and CEO of IPONWEB clarifies the challenges these small changes represent in relation to studying web behaviours. For illustration, he describes how each browser update for Internet Explorer, Firefox, Google Chrome, etc. can result in several changes to the data. Mouzykantskii refers to the challenge this presents as *change management*.

> "You have subtle semantic changes because the data which you store is the result of some other process and these processes keep changing…and therefore the meaning of the data keeps changing as well."
> *– Dr. Boris Mouzykantskii, IPONWEB*

Currently, change management is a costly, yet essential element of data analysis, especially when considering the scale (IPONWEB, for example, processes over a few hundred terabytes per day). In addition to the large proportion of time required to identify data changes, change management requires expertise in datasets, analysis, and business purpose.

4.1.2 Metadata

Yet Mouzykantskii addresses a larger issue, that by separating data from its context, its meaning is potentially lost. Data are not simply a string of numbers. Disconnecting context and meaning can lead to misinterpretation in analysis, which could lead to minor misunderstandings or, given the high-stakes and dependencies on data analytics, can have serious consequences for businesses. For example, in the Target pregnancy case recounted above, it only became clear that Target's marketing strategy was problematic once the (decontextualised) data on purchasing patterns were reunited with their original context—in this case, the juvenile age of the data subject.

> "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."
> *– Dr. Boris Mouzykantskii, IPONWEB*

The promise of new sources of data, particularly from interactive media, seems to herald possibilities for richer analysis. Yet, as Simon Thompson, Director of Commercial Solutions for ESRI observes, much of these data are not consistently organised. Meta tags—ways of marking data sources to identify content type or other contextual factors—are used inconsistently in most visual media. Since, for example, Youtube is publicly curated, tags are not standardised. Thus, if searching for a specific speech and using tags such as the speaker name, the subject, and location, many videos that are potentially relevant would not be found and likewise those that do appear in

search results may not be the strongest or most relevant examples. Currently, there are no consistent or reliable methods for organising and mining most visual data.

> "We're 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."
> *– Nigel Davis, Willis Group*

Most user generated content, such as social media posts or forum posts continue to present a measurement challenge. Unlike page views or clicks on links, textual and visual content are not easily quantified. Yet Google and Yahoo! serve advertising in their email platforms based on texts within personal messages. An ongoing challenge is to assign values to image data, whether generated from satellites, sensors, or personal events.

Even when data is professionally generated entirely in-house, problems can arise unless a clear strategy is enforced for clear and consistent documentation of the data and its original meaning or context.

4.2 Data reliability

As mentioned earlier, the strength of analysis of large datasets is dependent on the quality of the data. Two key issues affect data reliability: collection methods and definition of measures.

4.2.1 Data collection methods

For first party data that is gathered and used internally, collection methods are typically known, yet when businesses engage in second and third party arrangements, collection methods are not always transparent. In fact, for companies such as Nielsen and ComScore who provide data and analysis based on their audience panels, part of the proprietary dimensions of their business may be the formation of these panels. Businesses therefore rely upon these data sources without a complete understanding of how the audience panels are formed or responses measured.

A few of our experts raised concerns about the media industries' shared reliance on these datasets, asking who exactly the panel members were and what is really known about them. Daryl McNutt 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.

No matter how sophisticated the analytic process or large the dataset, it is important to remember their origins in raw data collected for a specific purpose, representing a specific sample over a specific time. The efficacy of these analyses relies upon the trustworthiness of the initial collection point.

> "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."
> *– Daryl McNutt, Adaptive Media*

As mentioned in Chapter 2, Dr. Cathy O'Neil, Program Director of the Lede Program in Data Journalism at Columbia University, used Netflix to illustrate how data collection can affect results. In comparison with Amazon's use of transactional data to recommend other items a user might like, historically, Netflix relied upon member reviews. A potential problem with this method was that the recommendation engine for Netflix relied only on data from those likely to review rather than the larger body of Netflix user transactional data. Netflix has since improved the accuracy of their recommendation engine, which now draws on user behaviours as well as metadata about the shows (Vanderbilt, 2013).

When considering Mouzykantskii's point that data are often separated from their descriptors, reliability at the level of collection is especially important. As data moves away from its original source and apart from its metadata, issues in the data collection become potentially forgotten and the data may be assumed to be adequately representative of all users rather than a specific subset.

4.2.2 Proxy measures

When considering the challenges of measuring the immeasurable, statisticians often refer to a popular, unattributed quote that "Not everything that counts can be counted and not everything that can be counted counts." Even with vast amounts of data available and in the face of the increasing capacity to link these data, there remain elements that cannot be directly measured. O'Neil (2014) uses the example of the "amount of influence of various politicians" or the "value to a company of a good (or bad) reputation" (p.4) as elements that cannot be measured, Tariq Khokhar, Data Scientist at the World Bank describes the difficulty of measuring poverty, and a few of our experts mentioned the challenge of measuring interest.

From a business perspective, *proxy measures* are used to indirectly approximate one phenomenon by measuring another. O'Neil (2014) explains that *interest* is often measured by how many pages a person viewed and/or how much time was spent viewing those pages, rather than more direct measures such as an attitudinal survey. Proxy measures are frequently used in business as well as research and are an accepted practice, yet our experts emphasize (1) the importance of remembering when a measure is a proxy and not a direct measurement and (2) being careful of what measures are chosen to be proxies. For the latter example, O'Neill recommends blending quantitative and qualitative data to reach a more reliable approximation. She further recommends honesty and transparency in reporting by clearly

identifying when and how proxies are used and just as clearly admitting to uncertainty in measuring and analysing these phenomena.

4.3 Data availability

While massive, big data is not infinite. At the edges of big data are elements that have not been measured, cannot be measured, have outdated measures, or gaps in measurements. Tariq Khokhar of the World Bank reports that the biggest challenge facing his institution is lacking or unavailable data. In this case, the challenge is that household surveys become quickly outdated and data take years to collect, curate, and analyse. Yet commercial companies face similar challenges. Bryan Lorenz, Vice President of Data for BrightScope reports challenges in using historic government data. The retirement plan data are available, but as .pdf files rather than machine-readable formats, so must be processed using optical text recognition software and human analysts. Cathy O'Neil (2014), Program Director of the Lede Program in Data Journalism at Columbia University describes concepts that are not easily measured, such as the value to a company of having a good or bad reputation (p.4). While each of these conditions present unique problems, they reflect the overall challenge of data availability.

Compounding these issues are technical and political issues of access. Assuming relevant data are available, companies must have appropriate technologies for access, processing, and storage. Often mentioned in relation to working with clients, either on advertising campaigns, business strategy, or audience segmentation is the issue of accessing client datasets.

> Accessing a client's "large pool of data itself is quite a challenge since you need to have a large enough infrastructure to support it."
> *– Devin Guan, Drawbridge*

4.3.1 Tools for access, extraction, processing, and analysis

Alastair McCullough, Business Intelligence Competence Centre Strategy Leader, Europe for IBM Global Business Services describes the two-fold technical challenge of needing adequate storage and processing capabilities to access client data. While storage options were addressed in earlier sections of this report, an important note is that Hadoop was frequently mentioned in interviews and industry reports as enabling the processing of large datasets. Hadoop is an open source software framework that uses distributed computation (e.g., cloud computing, calculations across multiple databases) to enable large-scale processing of data.[15] With this powerful processing framework, analysts can combine and re-combine large datasets. In the recent past, when companies managing massive datasets, such as Ford or Proctor & Gamble attempted to run queries on their data, they would inevitably face database crashes. In response to this long existing need,

[15] http://en.wikipedia.org/wiki/Apache_Hadoop

Hadoop and similar frameworks constantly run combinations of queries as background processes so that the platform can more easily manage new queries.

While these platforms offer technical solutions, the question of value remains at the centre of any commercial engagement with big data. Jeremy Barnes of Datacratic emphasises the importance of consistently asking whether the value of having the information will be worth its cost.

> "Is the value of having that extra bit of information worth the price you're going to pay for it? Or is the cost of, if you're expanding your data set by a factor of ten and needing to buy ten times as many servers, does that actually make sense? Often you can get 99% of the value from one hundredth of the data, so it doesn't actually make sense to go and try and collect everything."
> *– Jeremy Barnes, Datacratic*

Value is always at the forefront of decisions around big data use. A barrier for companies using big data is finding a balance between cost and access. Given the personnel hours absorbed by data cleaning, management, and analysis, these are not insignificant considerations when weighing costs of expanding datasets.

Further, many of our experts describe the frequent need for bespoke tools when attempting new or different analyses. As businesses shift to predictive analysis, they are often facing new questions and new processing requirements. Indeed, pushing the threshold of what is currently technically feasible is, in some sense, a fundamental characteristic of any kind of work with big data. Bret Shroyer, Senior Vice President of Reinsurance for the Willis Group, when explaining the technical needs for catastrophic modelling cites a lack of standardised tools or processes as a significant obstacle.

> "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."
> *– Bret Shroyer, Willis Group*

Shroyer's observations were consistent with a majority of our interviews. Many of our experts believed current tools, while consistently improving, to be limited when attempting analyses across datasets.

In discussing the time often spent customising and managing tools for analysis, Dr. Cathy O'Neil of Columbia University expects that technologies will continue to become more intuitive and easier to use.

When describing technologies, “possible” is a significant distance from “easy” or “intuitive,” which makes the practical business of big data analysis an ongoing challenge. Innovations in technology have certainly contributed to improved and increased capabilities, yet are still a far way from performing the way O’Neil describes.

> “As a mathematician and a scientist 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.”
> *– Dr. Cathy O’Neil, Columbia University*

4.3.2 Politics of data management

As mentioned in Chapter 3, data management is neither a standardised nor straightforward process. Company politics affect what data are shared internally, between and within departments, as well as how and which 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.

For example, decisions regarding consistency in reporting timestamps can affect matching datasets later. How a company records structural change to the dataset can additionally affect current and future users and reduce the amount of time spent sorting these changes later.

While some issues of data management are simply matters of a company’s internal organisation, others have more serious considerations. Legal frameworks for privacy and protection of proprietary data are of primary importance when negotiating data access and management. Although not standardised, industries operate under expectations for the anonymisation of Personally Identifiable Information (PII) even if unevenly implemented within and across sectors. Typically, those using PII data process it into segments or aggregations, with PII removed, although internal politics and national legal frameworks influence these decisions.

4.3.3 Linking data

> "The hard part of big data is how to connect the data together. So why is that hard? It is hard because no one can tell you what is the primary key."
> *– Dr. Phil Mui, Acxiom*

Unfortunately, values in datasets are not like a jigsaw puzzle where, with enough searching and concentration, pieces that fit together can be found. Instead, the act of combining datasets more closely resembles putting pieces together from multiple and endless puzzles with pieces that never quite fit with each other. Simon Thompson of ESRI describes this current moment for big data as a technology phase in which businesses are "inventing and innovating ways to collect and manage information" with resources devoted to databases and data storage techniques. Yet since combining these databases is not a key focus, the unfortunate result of these unique systems is to separate rather than connect information.

Datasets are owned and managed by a variety of groups (e.g., government, Internet Service Providers, analytics vendors) with different interests in sharing. Loyalty card systems used by Starbucks and Tesco, for example, serve a specific purpose within the companies and are not necessarily designed for seamless analysis by Google's system or that of a third party advertiser. Further, these loyalty programmes enable the linking of offline sales to online sales. Likewise, data collected by mobile gaming or coupon applications may not seamlessly match with data from other mobile gaming or coupon applications.

Paul Malyon of Experian views these obstacles as relatively common across datasets, whether open or commercial because data are collected for a specific task or purpose within an organisation. When selecting and using third party datasets, Malyon recommends "You have to have context for the data, so you have to know what the data is, you have to understand how often it's updated, where it's generated from. You have to be able to rely on that source." As Dr. Boris Mouzykantskii of IPONWEB observes, however, data are often separated from their context and do not carry descriptions that might help in matching datapoints across datasets.

This separation of metadata is especially challenging when attempting to track single users across multiple datasets. Dr. Phil Mui of Acxiom described a silo effect resulting from unique storage and collection systems. Without a means to match datasets, any potential value in predictive analysis is lost.

> "Even if you have all the processing power in the world, if you don't know that joebloggs@tesco.com, who made a particular purchase offline is the same as joebloggs@yahoo.com then whatever data that I collect about you is in different siloes and cannot be linked together to derive insight."
> *--Dr. Phil Mui, Acxiom*

Linking data across datasets often means attempting to track a single user across multiple devices (e.g., laptop, mobile phone, tablet). For example, retailers have difficulty linking users who research a product online with eventual in-store purchases. Marks & Spencer, a general merchandise retailer, reported that online purchases represented 13% of total sales for 2013, yet they record an average of 3.6 million weekly visitors.[16] Argos, a home goods retailer, uses a strategy by which online shoppers can reserve items for in-store purchase. Internet sales represent a 41% of total sales[17], an exceptionally high percentage, but this figure represents a mix of online purchase and in-store pick-up. A benefit of this approach is the ability to pair a customer's web behaviours, e.g., researching a product, with their in-store purchases.

In Drawbridge's recent (2014) White Paper, "Approaches and Tips for Effective Cross-Device Advertising," two approaches, *deterministic* and *probabilistic device pairings*, are described. In the first, an individual user is tracked across devices via a login, so a company like Facebook or Google can serve advertisements to a user according to his or her behaviours and preferences whether the user is on a mobile phone or laptop. Whenever the user is logged in, the companies can be relatively certain of whom they are targeting.

Attempting to match a user across devices without benefit of a login has previously lacked the "key" described by Mui, a way to connect a single user to behaviours across mobile apps and websites (accessed via laptop, desktop computer, or tablet). However, Drawbridge Vice President of Engineering Devin Guan and Vice President of Marketing Brian Ferrario describe an innovative method of matching datapoints for a single user across devices. Using *probabilistic modelling*, "where large amounts of known data need to be analysed to make assumptions about the unknown, such as in weather forecasting or stock market predictions,"[18] developers at Drawbridge have formulated a method to track a user across devices. In explaining the process, Guan describes sorting through petabytes of data to identify specific datapoints that will be unique to a single user. Guan used the example of music preferences and access patterns:

---

[16] http://annualreport.marksandspencer.com/strategic-review/our-plan-in-action/multi-channel/index.html
[17] http://www.computerworlduk.com/news/it-business/3364875/argos-multichannel-sales-buoy-struggling-home-retail-group/
[18] Drawbridge (2014). *Approaches and Tips for Effective Cross-Device Advertising.* San Mateo: Drawbridge. http://blog.drawbrid.ge/?p=122

> “I’m pretty sure that if you pick a random eight songs from a random eight singers from your iTunes list they can very uniquely identify you compared to any other person in the world.
> Same thing with any user in the mobile world, like the type of app, the access points, the location they access, and the time they access them is very unique too.”
> *--Devin Guan, Drawbridge*

Drawbridge represents an innovative probabilistic method of finding unique identifiers to follow a user across devices. As businesses such as Drawbridge identify the elusive key to connect a user across devices and datasets, they move closer to removing obstacles to matching datasets.

4.4 Overcoming obstacles

Obstacles facing users of big data are not simply technical. Apparent in our interviews is that strong leadership is critical so that issues such as data quality or change management do not become insurmountable in cost or labour. Strong leadership includes an awareness of the limitations of large datasets, especially what is required for maintenance and usability, how its use is integrated into a profit model, and the kind of teams that will maximise lucrative engagement with the data. The next chapter addresses ways in which strong management can realise the potential of big data.

Chapter 5
Realising the potential of big data

As everything from health services and auto purchase to housing and education becomes ever more quantifiable, companies that can collect and process these digitised data outputs are best positioned to profit from big data. First party data—such as those collected by Amazon, Tesco, and Visa—provides a competitive advantage because it directly captures customer behaviour, making the data source unique and proprietary. To fill in the gaps of their first party data collection, a majority of companies in our study additionally purchase third party data (and use open data) to better understand their market. Companies best positioned to realise the potential of big data therefore collect their own data and use second and third party data to extend and supplement their analysis.

As mentioned in an earlier chapter, the significant change in practice is not so much a matter of 'big data', but 'big processing.' As billions of data points are now processed in real-time, companies rely upon these data to drive real-time decision-making. The realities of this shift are likely apparent in meeting rooms across the globe. No longer are decisions reliant on incomplete or out-dated annual, quarterly, or even monthly reports. Fluid analytic platforms are leading toward real-time discussions of data, away from weekly meetings in which data were discussed only to find a key point missing that would then be incorporated at a later date, delaying action. Now, analysts bring data

processing to their meetings[19], testing, analysing, and reporting real-time data to drive decision-making.

What does real-time decision-making look like in practice? One example is Uber[20], a rideshare and taxi alternative application that pairs riders with drivers, using data-driven decision-making to resolve a gap in the existing system of hired transport. Through its mobile application, Uber efficiently directs resources where needed, providing maps of available drivers GPS-targeted to those seeking rides. It further maximises capacity by connecting people possessing cars, local geographic knowledge, and time to drive with those in need of their services. With riders in direct contact with drivers, Uber improves upon the existing system of calling and waiting for a car service or taxi. Airbnb[21] similarly functions to maximise resource distribution and leverage local capacity. Offering an alternative to hotels, Airbnb optimises the potential afforded by ratings systems to pair potential guests with hosts, enabling data-driven decisions on both ends of the supply chain.

Likewise, retail recommendation engines such as Amazon categorise buyers based on their purchase histories to suggest future purchases. Recommendations are driven by data, as are consumer purchasing decisions. Retailers who use reward cards, such as Starbucks and Tesco can deliver real-time coupons to customers' mobile phones or initiate store-specific campaigns based on real-time data. For advertisers, what used to be an expensive, time-consuming process of testing campaigns with limited feedback based on focus groups or delayed, dispersed feedback from sales can now occur in real-time. The costs of testing different variables are low in terms of time and delivery as well as storage and processing. For insurers, real-time data can enable identification of vulnerabilities based on weather data or pharmaceutical testing or disease outbreaks.

Recommendation #1: Data must be central to business model

As the above examples illustrate, what predicts whether a company is well positioned to maximise the potentials of big data is not its size, but its integration of data into its core business model. Simon Thompson, Director of Commercial Solutions at ESRI describes this approach: "Information is part of their business nervous system. The whole organisation understands the value of information." As demonstrated by Uber and Airbnb, big data should be the key focus of the business, not an afterthought or accessory. Our experts recommend that companies engaging big data have a clear plan for use, with dedicated resources for analysis, and flexibility to respond to what they find in the data.

Unfortunately, many start-up companies fail to realise the potential of big data beyond a promising buzzword. Dr. Cathy O'Neil of Columbia University warns

[19] Programs such as Vertica and Tableau enable real-time analysis of data across hundreds of machines.
[20] https://www.uber.com/
[21] https://www.airbnb.com

that many companies may not easily benefit from big data. She offers a litmus test to determine the centrality of data within an organisation:

- Do you have a data-driven culture?
- Do you collect data on your customers?
- Do you actually collect the relevant data that you can use to improve your customer's experience or to answer the question that you think will help your company?

Putting data at the centre of a business is likely to be especially challenging for large, long-established companies with no significant history of data use, where the change is likely to represent a significant organisational and cultural shift. Nonetheless, success stories such as the Tesco Clubcard illustrate the benefits that such a reinvention can yield.

Recommendation #2: Clear profit model

Clearly, big data alone does not guarantee commercial success. In fact, when considering the success of start-ups or businesses in general, a higher likelihood exists for failure (Nobel, 2011). Jeremy Barnes, Co-founder and Chief Technology Officer of Datacratic cautions against assuming that the use of big data automatically leads to profit. He notes a common failing of businesses to consider how and why they plan to use big data, describing a frequent scenario in which the first step is "using big data" and the third step is "profit," but the essential middle step, the *how* and *why*, is ignored. Jeanne Holm, Evangelist for Data.gov similarly observed that start-up companies often have interesting ideas for using big data, for example in mobile applications to identify traffic hotspots, or as ratings systems for local restaurants, but fail to develop plans for profitability and sustainability, such as using the ratings system to drive a service like Airbnb. Barnes recommends that companies tie their data use to clear goals: "People dealing with the data must understand what the expectation is for how that data is going to be valuable and orient their efforts toward that." As demonstrated in the examples of Uber, Airbnb, Amazon, and Tesco, the likelihood of success increases when businesses are structured around clear goal setting that identifies the purpose of data to drive real-time decisions.

> "Keep costs in line with the amount of value the data is providing."
> *--Jeremy Barnes, Datacratic*

Recommendation #3: Pair strong business strategy with understanding of technology

Strong business strategy remains a consistent predictor for successfully engaging big data. Yet, as data storage becomes cheaper, analysis faster, and tools progressively streamline reporting and analysis, technological advancements increasingly drive strategic decisions. When determining what leads to successful integration of big data in commercial pursuits, it is difficult

to separate what is considered strong business practice from the technological advancements that continue to improve data processing. Unsurprisingly, many companies in our study, including IBM, Acxiom, Experian, Audience Science, BrightScope, and Willis Group, combine strategic advisement with a technical platform (for further details, please see Chapter 3: Business models for big data).

As real-time data driven decision-making practices mature, so does an appreciation for the necessity of human-directed analysis. Technology continues to push the envelope of what is possible, surpassing human capacity in the generation and processing of data. Yet, unlike machines, human analysts can pick up nuances, determine significance of various combinations and experiments, or direct analysis according to strategy. For example, humans can easily distinguish between “read” and “red” or Jaguar the car and jaguar the animal. According to our experts, this human factor is essential to deriving value from large datasets.

> “The idea is not just collecting the data or even summarising it but how do you put the data into action? How do you actually use it to drive actual results in systems?”
> *--Dr. Basem Nayfeh, Audience Science*

Alastair McCullough, Business Intelligence Competence Centre Strategy Leader, Europe at IBM Global Business Services observes that companies who are already aggregating and using big data, who have experience in striking the balance between the technical capabilities and human-directed strategy are well-positioned for moving forward. As examples, McCullough mentioned credit rating agencies (e.g., Experian and Acxiom), telephone companies (e.g., Vodafone and Telefonica), and retailers (e.g., Marks & Spencer, Tesco, and Walmart).

Recommendation #4: Look for Low-Hanging Fruit

The rate of innovation among big data start-ups is astonishing, as is the scale of some big data infrastructure projects. However, many of the greatest success stories have come from businesses that have found new uses for old data, or that have found relatively straightforward ways of capturing previously neglected data. Companies such as Tesco and Netflix have built success by recognising the value of data that might be generated but not captured in the course of their core activities. Retailers and media firms have led the way with loyalty cards and audience data, but there are many ways to cheaply generate data such as encouraging online customer interaction or retrofitting inexpensive sensors to existing equipment. The cost-benefit test is much more likely to be passed by data projects that utilise existing data or build upon existing transactions.

Many companies could benefit simply from using additional data generated in this fashion to better inform themselves about their business and its

customers, and to make decisions about price-setting, product design, or organisation of the production process. Off the shelf solutions now exist that make estimating consumer demand elasticity or monitoring machinery downtime feasible—even for relatively small businesses. Such data can also be used to improve the customer experience by making service more personalised, improving the responsiveness of after-sales support, or providing information such as product recommendations to help consumers make more informed choices.

Realising profit

Despite the hype, realising profit from big data relies upon strategic practices that pre-date big data and big processing. Leadership must pair vision with a clear profit model. Essential to this mix are reasonable expectations of technologies that consider the realities of big data use, particularly the attendant challenges and resource requirements. Data must be central to the business model, driving real-time decisions. Most importantly, the human factor cannot be overlooked; human-directed analysis is critical to successfully managing this massive data.

Chapter 6
Big data skills and the organisational environment

The term “data scientist” has become a catchall to describe statisticians possessing programming skills, business knowledge or a combination of both. If current predictions are to be believed, alongside the expansion of data use in commerce is a growing and largely unfilled need for data scientists to aggregate and extract meaning. In a 2013 report by eSkills UK, the number of specialist staff working in larger firms is expected to grow to approximately 69,000 people, an increase of 243% in five years. Similarly, McKinsey (Manyika, et al., 2011) predicts an unprecedented “deep analytical skills” gap in the U.S. of between 140,000-190,000 positions by 2018 (p. 11). To better understand training needs for a future workforce, we asked experts to describe the specific skills their organisation required to make use of their data.

We find that data work is fluid, often practised under pressure and frequently demanding attention to detail while simultaneously focusing on the larger purpose. As detailed in earlier chapters, analysis often involves the customisation or creation of tools, the painstaking cleaning of datasets, and the technical and analytic challenges of linking datasets. Within this challenging environment, data workers must have a strong skill set that

combines technical and business acumen, involving creativity and agility as well as strong problem-solving skills. Grit, dogged persistence and resilience in the face of these daily challenges, underlies the essential skill set of data workers, without which survival and success are unlikely.

6.1 Statistical analysis

> "There's a trivialisation of the amount of intellectual effort needed to actually make sense of some of these datasets."
> *--Jeanne Holm, Data.gov*

A strong statistical background provides the foundation for data work. Companies rely upon analytic polymaths, data workers who are fluent in statistical approaches and analytic tools and versatile in their application. Despite analytic tools that seemingly obviate the need for deep math skills, a thorough and flexible understanding of math and the scientific method is a foundational requirement for data workers. Dr. Vivienne Ming, VP of Research and Insight of Gild observes a growing focus of employers on creative problem solving skills, the ability to build models to test hypotheses and derive meaning from data.

To support creative problem solving, a company culture must encourage experimentation and allow for failure. Paul Malyon, Head of Business to Business Transactional and Marketing Products for Experian, explains, "You need a management team who are willing to take calculated risks on new data sources, new techniques of analysis and new technology. In a bubble, they need to give their permission to their team to try and fail and try again." Dr. Boris Mouzykantskii says that at IPONWEB, he encourages teams to challenge each other, to find gaps in analysis and to assume findings are not true without rigorous testing. Successful practice is therefore a combination of a strong analytic skill set and a company culture that allows for experimentation and failure.

6.2 Coding skills

Given the interdependence between analytics and tools, coding skills are a necessity for analysts as well as management. A 2014 report by the Graduate Management Admissions Council found that across 44 countries, companies rated quantitative and technical skills among their top three criteria for new hires (GMAC, 2014, p.16). Amazon reportedly seeks MBAs who can "dive into data and be technically conversant (Weinberg, 2014). We find that companies seek what Jeanne Holm of data.gov described as *merged skills*, or as Pink (2012) describes as stretching "across functional boundaries" (34-35). Chris Nott, Chief Technology Officer of Big Data and Analytics at IBM describes a data scientist as someone who can grasp a business need and apply analytic and programming expertise to explore the need and derive insight for action.

> "You don't want to simply hire somebody that is smart in using a certain data, certain types of technology, because this stuff moves too quickly. You want to hire smart technologists who are willing to pick up new technologies without getting anxious."
> *--Dr. Phil Mui, Acxiom*

We find that employers seek above average coding ability. As with analytic skills, technical skills must be accompanied by creativity and flexibility in application. This is essential for big data work because, as Dr. Mui emphasises, the pace of change is so rapid that adaptation to new challenges is a central part of the job. For this reason, many data scientists are educated to an advanced postgraduate level.

George MacKerron, Founder of Mappiness describes the relationship between these skills: "the analysis involves a certain amount of programming. It means there's flexibility; there's not much you can't do if you're willing to sit down and write some code." Baseline technical expectations include expertise with database frameworks such as Vertica, Hadoop, NoSQL, or PostGIS and familiarity with at least one mainstream programming language. To respond to daily demands, data workers must have, as MacKerron describes, an awareness of what is technically possible with existing tools, what is feasible to customise, and where external solutions are needed.

6.3 Business expertise

Business expertise includes an understanding of the market, how to derive value from data, develop models, identify niches, and fill gaps. Within a successful company, analysis occurs within a strategic framework; teams balance the quantitative and technical dimensions of data work within overarching commercial demands. A typical scenario is that those working in business strategy will direct the focus of analysis, then analysts will identify gaps in the dataset, determining whether these gaps can be filled by datasets the company has already collected or purchased. Whether additional datasets must be purchased is a business strategy decision. Layering of data to determine specific strategic directions also falls with the purview of business expertise. The analyses will then inform real-time actions.

Simon Thompson, Director of Commercial Solutions for ESRI describes "people who in terms of business questions can look at the relevance of those business questions to change the business to credibly understand how data makes an impact in terms of being able to perform analytics on it." Typically, business people are analysts with an aptitude for strategic thinking. This pairing ideally results in, as Thompson described, an ability to see how data can be used to inform strategic decisions. We find this analytic background represents a paradigm shift: traditionally, management in fields such as advertising and finance would be drawn from sales departments, but must now possess an analytics background.

6.4 Domain knowledge

While quantitative and technical skills are normally identified as essential for data scientists, Holm includes domain knowledge in her description, saying that to truly find value in big data, a person must combine these skills with knowledge of a particular market sector:

> "When you have somebody who understands biochemistry really well and is also a statistician, they're going to be able to make a lot out of those big datasets. But if you try to just take a new computer science grad and throw them at the problem, it's not going to be very useful. Or if you take just a biochemist that doesn't have some of the statistical analysis capabilities it's going to also be difficult."

The realities of data work require a merged skill set to understand which datasets are available and should be used together. Tim Davies, Open Data Research Coordinator for the World Wide Web Foundation describes the challenge of bringing together multiple datasets, requiring "not only domain expertise but knowledge of the way data is managed in that sector." Davies provides an example of working with National Health Service data in which those working on health issues, but unfamiliar with NHS codes faced a steep learning curve, yet other non-specialists but who were familiar with government data could more easily "comprehend what the data was." A Seattle-based analyst similarly reported it took a full year to know what data to check within a large company, to know what makes sense for analysis, saying it "feels like a daily sprint."

> "You have to be part detective, know how to get around problems and solve problems by finding other alternatives. It's not just about the technology but it's about how to solve the problem itself."
> *--Marwa Mabrouk, ESRI*

Building of domain knowledge is time-consuming and ongoing, yet this deep understanding enables flexibility and resourcefulness in identifying and layering relevant datasets. Devin Guan, Vice President of Engineering for Drawbridge, describes an interplay of creativity and intuition enabled by industry experience: "First, you must have the right type of training in analysing the data and applying statistical or probabilistic models to the data and have enough foundation in mathematics and data science. Beyond that, it's the creativity and experience in the industry and how much data intuition into the numbers and the data you have." Mabrouk further observes that "Data can be used in multiple ways, so you have to have enough insight to see how you can apply specific datasets to problems that maybe people didn't think of it as a way to solve it before."

6.5 Importance of teams

Experts consistently reported that skills are not to be found in one person, but in teams. In describing the diverse technical, scientific, analytic, and business

needs required for businesses to use big data, Mabrouk reports the difficulty in finding these skills in one person, “what you see is a variety of different skill sets and all of them typically come together in teams.” Nott recommends that instead of seeking this diverse skill set in a single person, to find people who can effectively contribute their expertise in collaboration with team members and widen access to information by adopting analytics tools which hide complexity where feasible. Within this context of intense collaboration, Sue Bateman and Romina Ahmad of data.gov.uk emphasise the importance of strong communication skills to bridge potential gaps in domain knowledge. Indeed, corporate recruiters participating in the 2014 GMAC study rank communication skills as a first priority.

Paul Malyon of Experian broadly envisions teams as a series of ecosystems involving the entire organization. Malyon described an arc of expertise that starts with management who are willing to engage with new data sources and techniques of analysis. He explained that those who develop products must also “understand data and the power of data.” He further emphasised the importance of positioning data workers where they can provide feedback and collaborate on how data are used. Malyon extended his skills description to include the need for account managers and sales people “who actually understand the data that lies behind the products and propositions.”

Chapter 7
Government role

Comprising the petabytes[22] of data analysed each day are individuals who contribute their personal information and behaviours, with different levels of consent to and understanding of this contribution. When Eric Schmidt observed that every two days we create more data than had been generated over the course of civilisation, he also predicted that this pace would increase. The multiple stakeholders engaging big data must grapple with issues of privacy, responsibility, and transparency. Increasingly faster technological capabilities challenge traditional practice and have thus far outpaced government response. As Dr. Arthur Thomas, Founder of Proteus Associates describes from a biotech perspective, "there is the fundamental data governance issue of how to find a balance between having data which is open enough to do scientific research versus protecting the important details about a particular person." Across sectors, experts acknowledged similar challenges.

Among those we interviewed, we additionally find a concern that industry expertise is not adequately represented in discussions of regulation and provision, resulting in a potentially limited view of data practice. On this issue, our expert perspectives varied. Some felt that the potentials of big data had

[22] One petabyte = one billion megabytes

been overstated and have thus generated uninformed panics. Others feel that decision-makers are not informed enough about the various potentials of big data processing, regarding ever shrinking privacy on the one hand and expected economic benefit on the other. Whether government should intervene in commercial big data use and at what point this intervention should occur remains an open question.

> "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."
> *--Heather Savory, UK Open Data User Group*

A majority of our experts expressed similar perspectives. Through our interviews, we explored recommendations for achieving a minimum regulatory infrastructure and promoting the economic benefits of big data.

7.1 Standardised Datasets

Nearly all of the companies participating in our study use some form of public or open data. By far, the largest obstacle to using this data is data quality. Datasets are often inconsistent, whether across agencies, within agencies, or within the same collection period. Historical data is frequently available as pdf files instead of data files, and meta information (describing column codes and what numbers represent) tends to be inconsistent or unavailable. Given private sector dependence on public data, this lack of standardisation significantly impacts productivity and profit. Our experts estimate that analysts spend between 80-90% of their time preparing data for analysis (e.g., cleaning the data, locating data across different files or databases, change management, linking data, and filtering).

Tim Davies, Open Data Research Coordinator for the World Wide Web Foundation says that while data are available, they are "scattered across many different files" and observes that "the technical challenge of aggregating that data are significant." Dr. Vivienne Ming, VP of Research and Insight of Gild finds that public data are often in poor shape, noting that high uncertainty in data quality hurts the industry as a whole by reducing confidence: "Ultimately, the biggest value from these sorts of datasets is going to come from having well maintained, highly accessible, highly shareable data."

Bryan Lorenz, Vice President of Data for BrightScope, acknowledges that providing public access to large government administrative databases is a significant achievement. Lorenz sees the next challenge as "accessing detailed data more easily," observing that currently, while datasets are searchable, the search tools tend to be generic and under-powered. This observation was consistent with other respondents, that making access to data 'possible' was far from making it 'easy.' Not all data are digitised,

searchable, or machine-readable, imposing a tedious burden on companies to make the data usable. This data preparation becomes proprietary, meaning that the painstaking cleaning and sorting of the data is repeated across companies, creating a significant inefficiency.

When issues of standardisation, metadata, and linking are not considered, Dr. George MacKerron, Founder of Mappiness points out “you risk making all future data incomparable with all past data and essentially ending up with a multiplicity of incompatible datasets.” Nigel Davis, Analytics IT Director at Willis Group described the utter enormity of the data companies increasingly need to use and acknowledged growing challenges posed by datasets pulled from multiple sources in which authenticity and quality must be determined without shared standards and accurate metadata. Several of our experts mentioned that lack of standards adds significantly to the workload and therefore, expense of using big data.

Davies observes “a lot of what we’ve seen in getting data online has been data dumping rather than saying this is data part of the core infrastructure of what the state does or what the company does.” He recommends that in making data publicly available, a focus should be on identifying what data are core and guaranteed to be available, and also determining which datasets are important for linking up different data. We find that many of our experts emphasised the need for currency in data and for consistent timetables for making new data available. Those working in the UK acknowledged that the government is well-advanced in providing open data and believed a beneficial next step would be to improve the linking of datasets. Many anticipated, as MacKerron says, “massive benefits from being able to join different datasets.”

Our experts recommend standardisation of codes, formats, and change management as well as accurate metadata to describe these codes. Tariq Khokhar, Data Scientist at the World Bank emphasised the importance of providing data in machine-readable, analysable form. Davies recommends that governments need to step beyond publishing data to adopt a model of open data engagement where they “support conversations around it and build capacity and collaborate with people.” From a practical perspective, Bret Shroyer, Senior Vice-President of Reinsurance at Willis Group recommends that governments provide software tools and best practice guidelines to improve access and use of public data. Marwa Mabrouk, Cloud and Big Data Product Manager at ESRI and Jeanne Holm, Evangelist at data.gov recommended collective platforms in which cleaned datasets could be shared among data workers, noting that the majority of issues with quality are shared, so fixes would be useful for many.

Chris Nott, Chief Technology Officer of Big Data and Analytics at IBM, encouraged governments to think in terms of the future directions of open data: exploring how open data is being used in many other countries and wider uses not only by companies but individuals. Addressing the increasing demand for open data, Nott suggests that different levels of service could have a pay model, with charges for up-to-date data for commercial use, for example.

7.2 Regulation

The prevailing sentiment among our experts, when asked about the potential role for government regulation, was that any intervention should be limited. The above quote from Heather Savory summarises this well: government can provide an enabling role by providing the infrastructure and framework necessary for data-intensive businesses to flourish, but should stop short of actively interfering with the details of how big data business models develop. One exception, where a more fine-grained involvement of government may be warranted, is in a promotional capacity: small businesses and entrepreneurs that may not be aware of the full spectrum of opportunities and public resources available should be encouraged to participate in the data economy and the public sector may have a role to play here.

On a more macro level, there was a general consensus that big data policies should be transparent, clear, fair, and consistent. In particular, the regulatory framework should be one that encourages businesses to innovate and compete on their own merits rather than 'picking a winner'. 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. There is a general willingness in business to respect and protect personal privacy, but our experts bemoan the lack of clear standards in this area. The law has lagged behind both the growth in personal data use and developments in technical and statistical anonymisation techniques. Moreover, there is a perceived lack of fairness, especially given the lack of standardisation of privacy practices across jurisdictional boundaries. This hints at a role for strong government leadership in establishing international standards for data practices. Indeed, a number of experts were of the view that voluntary standards or codes of conduct 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.

It is important that any regulation is designed to facilitate, rather than stifle innovation and growth in the sector. Given the growing level of expertise in industry, private-sector stakeholders should be involved in any new regulatory process. Devin Guan remarked that much privacy regulation responds to fears or misunderstood uses of data, while a few of our experts observed that policymakers are unlikely to fully understand the needs of businesses working on the cutting-edge of data science. These are problems that are best resolved by engaging the business community from the earliest possible stage.

7.3 Provision

A majority of our experts considered that governments are making a laudable effort to provide open access to datasets as well as workshops to support start-up businesses. Yet, there seems to be space for better supporting capacity building around using large open datasets. A few experts recommended training and support in large data access and use. Training could directly benefit those seeking to use the datasets as well as better preparing a workforce to work with data.

Marwa Mabrouk of ESRI works with communities to distribute the more tedious aspects of data cleaning and change management. She recommends that alongside their open data provision, governments provide a data warehouse where users can upload cleaned datasets and ideally reduce repetitious processing for the data science community. As datasets become increasingly relevant for individuals, this suggestion reflects a willingness and need to communally curate data and pool collective efforts to reduce repetitive tasks.

Chapter 8
Conclusion

Over the past year as we've conducted our research, data processing capacities have increased exponentially. One expert reported increasing from 10 to 17 billion data points processed per day. According to the EMC^2 Digital Universe study (2014), data is doubling in size every two years and by 2020 will multiply ten-fold from 4.4 trillion gigabytes to 44 trillion gigabytes (EMC^2, 2014; Williams, 2014). These unprecedented increases impact storage, processing, human analysis, skill needs, and profit. This research explored the realities for commercial use of big data, identifying who has potential advantages as large-scale analysis eclipses previous practice.

We interviewed 28 thought leaders in the big data space from a range of sectors including advertising, technology, biotechnology, insurance, public good, and education and from different sized companies and organisations. We also interviewed members of government directly responsible for provision of open datasets.

Our key findings, as elaborated in earlier chapters:

- **Big data business models use a combination of datasets**, including public and open, proprietary, and data scraped from web.
- Large international companies are **dependent upon open government data** for a variety of purposes and needs.

- **Publicly available or open data are rarely used in isolation**, usually combined with semi-public, or private, proprietary datasets.
- **Companies best situated to realise potential of big data are those for whom data is integrated into core of business, not as an add-on or afterthought**; experts recommended a clear plan for use, with dedicated resources for analysis and flexibility to respond to what they find in data.
- A distinct advantage to the ever growing capabilities of big data processing is the **improved ability for businesses to make real-time, actionable decisions.**
- **Linking data across devices** is enabling further personalisation and predictive power.
- **Datasets and analysis tools may be available, but are not necessarily easy to use;** analysts report up to 90% of their time is spent cleaning and preparing data for use; change management (monitoring changes and updates to datasets) is time-consuming; tools are often bespoke due to the rapid advances in data collection and analysis.
- **Skills are not to be found in one person, but in teams;** in practice, commercial use of big data requires strong analytic skills, understanding of scientific method, business expertise, exceptional programming abilities and an ability to adapt to a constantly changing, challenging environment.
- **Experts urge for more transparency and consistency in policy and regulation;** while most agree that government should take a limited role and focus primarily on provision of resources, many report confusion about what is acceptable in terms of privacy, storage, collection.

Our experts envision a future where, much like the Internet, big data is seamlessly integrated into our daily lives. Examples include the Internet of Things, sensors, Google Glass. They urge a consideration of what it will mean when individuals can query large datasets and control their own data, when global sharing of datasets is a taken for granted practice, when nearly everything we do is quantified. One expert predicted that everyday life will incrementally become more personalised. Another believed that having data about our world will make the whole of society more effective. The increasing linking of datasets will enable more seamless information sharing across companies and government departments. Simon Thompson of ESRI observed that “Big data has changed our sense of where we are in the world, we believe everything comes to us, we don’t have to understand how the rest of the place we’re in connects to the larger part. He asks, “how big is this thing we’re looking at? Is big data finite?”

Despite the long history of using data in business, commercial use seems to be at a tipping point where technical, analytic and economic possibilities continue to accelerate. In the face of overwhelming opportunity, elements that traditionally makes a business successful—a clear vision, realistic focus on profit, and a sustainable growth model—are ever more important.

References

Ayres, Ian. (2007). *Super Crunchers.* New York: Bantam Books.

Bakhshi, H., Bravo-Biosca, A., & Mateos-Garcia, J. (2014). *Inside the Datavores: Estimating the Effect of Data and Online Analytics on Firm Performance.* London: NESTA. Retrieved from http://www.nesta.org.uk/sites/default/files/inside_the_datavores_technical_report.pdf

Bakhshi, H. & Mateos-Garcia, J. (2012). *Rise of the Datavores: How UK Businesses Analyse and Use Online Data.* London: NESTA. Retrieved from http://www.nesta.org.uk/sites/default/files/rise_of_the_datavores.pdf

Brynjolfsson, E., Hitt, L., Heekyung, K. (2011). St*rength in Numbers: How Does Data-Driven Decision Making Affect Firm Performance?* Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486

E-skills UK. (2013). *Big Data Analytics: Adoption and Employment Trends 2012-2017*. London: E-skills UK. Retrieved from http://www.e-skills.com/research/research-publications/big-data-analytics/#November%20report

Ebbert, J. (2012, December 3). Define it – What is Big Data? *AdExchanger.* Retrieved from http://www.adexchanger.com/online-advertising/big-data/

Economist. (2011, May 28). Building with Big Data. Retrieved from http://www.economist.com/node/18741392

EMC^2. (2014). *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things.* Massachusetts: EMC^2.
Experian Marketing Services. (2013). *Digital Trends 2013.* London: Experian. Retrieved from http://www.experian.co.uk/assets/marketing-services/white-papers/digital-trends-2013.pdf

Gild. (2013). *The Big Data Recruiting Playbook.* San Francisco: Gild. Retrieved from http://www.gild.com/resource/big-data-recruiting-playbook-2/

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.

Graduate Management Admissions Council. (2014). *Corporate Recruiters Survey Report.* Virginia: Graduate Management Admissions Council. Retrieved from http://www.gmac.com/market-intelligence-and-research/research-library/employment-outlook/2014-corporate-recruiters.aspx

Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014). The parable of Google flu: Traps in big data analysis. *Science*, 343(6176), 1203–1205.

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

Maude, F. (2012). *Unleashing the Potential.* London: Cabinet Office. Retrieved from https://www.gov.uk/government/publications/open-data-white-paper-unleashing-the-potential

Mayer-Schönberger, V. & Cukier, K. (2013). *Big Data: A Revolution that will Transform How WE Live, Work, and Think.* New York: Houghton Mifflin Harcourt Publishing Company.

Nobel, C. (2011). Why companies fail—and how their founders can bounce back. *Working Knowledge Blog.* Cambridge: Harvard Business School. Retrieved from http://hbswk.hbs.edu/item/6591.html

O'Neil, C. (2014). *On Being a Data Skeptic.* California: O'Reilly Media, Inc. Retrieved from http://www.oreilly.com/data/free/files/being-a-data-skeptic.pdf

Oxera Consulting (2013). *What is the Economic Impact of Geo Services?* Oxford: Oxera Consulting. Retrieved from http://www.oxera.com/Oxera/media/Oxera/downloads/reports/What-is-the-economic-impact-of-Geo-services_1.pdf

Pink, D. (2012). *To Sell is Human.* New York: Riverhead Books.

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

Shakespeare, S. (2013). *An Independent Review of Public Sector Information.* Ref. BIS/13/744. London: Department for Business, Innovation & Skills, HMSO. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/198752/13-744-shakespeare-review-of-public-sector-information.pdf

Silicon Valley Bank. (2013). *Startup Outlook 2013 Report.* Santa Clara: Silicon Valley Bank, 16-19. Retrieved from http://www.svb.com/startup-outlook-report/

Soudagar, R. (2013, October 28). How fashion retailer Burberry keeps customers coming back for more. *Forbes.* Retrieved from http://www.forbes.com/sites/sap/2013/10/28/how-fashion-retailer-burberry-keeps-customers-coming-back-for-more/

Swabey, P. (2013, April 16). Tesco saves millions with supply chain analytics. *Information Age.* Retrieved from http://www.information-age.com/technology/information-management/123456972/tesco-saves-millions-with-supply-chain-analytics

Vanderbilt, T. (2013, August 7). The science behind the Netflix alogrithms that decide what you'll watch next. *Wired.* Retrieved from http://www.wired.com/2013/08/qq_netflix-algorithm/

Weinberg, C. (2014, July 11). B-Schools finally acknowledge: Companies want MBAs who can code. *Bloomberg Business Week.* Retrieved from http://www.businessweek.com/articles/2014-07-11/b-schools-finally-acknowledge-companies-want-mbas-who-can-code

Williams, D. (2014, April 15). EMC: World's data doubling every two years... *Techday.* Retrieved from http://techday.com/it-brief/news/emc-worlds-data-doubling-every-two-years/182718/

Wheatley, M. (2013, January 10). Data-driven location choices drive latest Starbucks surge. *Data Informed.* Retrieved from http://data-informed.com/data-driven-location-choices-drive-latest-starbucks-surge/

Appendix 1: Participating companies, government offices, and organisations

Audience Science
Acxiom
Brightscope
Data.gov and Data.gov.uk
Datacratic
Drawbridge
ESRI
Experian
Gild
IBM
IPONWEB
Mappiness
Open Data User Group
Proteus Associates
Willis Group
World Bank
World Wide Web Foundation

Appendix 2: Interview questionnaire

1. Describe sector in which you work and your particular position.

2. How do you define big data?

3. Business model: What sort of data do you use? (e.g., is it data from within your own organization, or are you taking data from external sources, such as Twitter feeds, weather, employment, publicly available for data, or data that you're buying in, or a combination?) Who processes the data? (e.g., do you have your own dedicated analysts, are you buying in services, are you a client?) If in house, who is doing it, is it a few in a dept or are directors of depts involved?

4. Where are the challenges in collecting and using data (e.g., ownership, privacy, analytical skills, data quality, etc.)? What questions do you ask the data?

5. What do you see as being the characteristics of a company in your industry that is well-positioned to benefit from big data (e.g. big or small, vertical or flat hierarchy, etc.)? *Alternatively*: how should a company in your industry position itself to benefit from big data? (examples of possible responses: Knowing other datasets available that could combine with ours…)

6. What kinds of specific skills does your organization require to make use of big data?

7. What are the main obstacles to accessing and using big data? What are the obstacles to understanding, interpreting, assigning meaning to the findings?

8. Are there uses of data analytics that, in your view, have been or are overhyped/that have failed to deliver the expected benefits? Why?

9. What, in your view, could regulators/policy makers do to help/encourage companies to benefit from data analytics?

10. What is the big deal about big data? Where do you see big data going in the next 5, 20 years?