Technologies for Data

Kirsty Douglas

Working paper
April 2015

This working paper contributes to Securing Australia's Future (SAF) Project 05.

Table of contents

Technologies for data

Governments, intelligence agencies, businesses, and citizens must adapt rapidly to an environment where data is both a commodity and a currency. The data accumulated from centuries of observation, and rapid improvements in digital interoperability in the last 50 years, continue to transform the ways in which people understand and describe the world, and the uses to which information is put. Countries and organisations which develop and foster technologies to ensure that data can efficiently, reliably and economically be stored, curated, managed, retrieved, accessed, shared, and protected have an opportunity to capitalise on this global data ecosystem.

Overview

- This paper addresses some of the opportunities of three interconnected types of technology for data: data analytics ('big data'); technologies for open data; and infrastructure for collaborative research data.
- The notion of an information age or a data flood is not new.
- However, the rise of digital information and communication technologies (ICT) in the late 20th century has been described as engendering an intellectual, social and economic transformation 'akin to the emergence of literacy'.
- With the advent of digital programmable computers in the mid-20th century, data outputs from simulations of complex phenomena have increased enormously.
- More recently, improvements in networking and interoperability (broadly, the capacity for systems, organisations, processes and products to work together) have provided unprecedented access to information about the behaviour of people interacting with online platforms.
- The emergence of digital ICT as a general purpose technology (a GPT is a generic technology, widely used across the economy, with many different uses and spillover effects, after (Lipsey et al., 2005) over the last 50 years has been accompanied by a huge increase in the amount of information generated, collected, stored, and shared.
- In an environment of networked sensors and ubiquitous internet access, virtually all human activity is now a source of data. These data have different characteristics – some streaming, some stable, some updated over time, some stochastic – with different properties of stability and longevity, and different importance.
- If the current 'flood' of data can be distinguished from previous so-called information revolutions, perhaps it is better characterised as an age of interoperability rather than a new information age.
- The status of data is ambivalent. One researcher's data is another's metadata; one person's technology is another's infrastructure; the browsing and purchase histories of an online community are a commodity and a resource to other users; and the size of a 'big' dataset in one field of research, such as sociology, can be trifling in another, such as astronomy.
- In the face of this pervasive interoperability and classificatory instability, regulators, users and curators of data collections, and designers and administrators of data infrastructure may find it difficult to balance preservation, storage, and use with the management of digital traces, personal information harvested from online behaviour and sensors.
- If ICT is a general purpose technology, then data are its general-purpose constituents. The impact of technologies for data stems from the interactions of data with other technology and with other data, ensuring a multiplicative effect: as 'data intermediaries', technologies for data can themselves be the mechanisms by which things interoperate.
- Merely having more data does not translate automatically to having better information. 'Big data' analysis is subject to the same limits as any statistical analysis.
- The various uses of data creates tensions that feed into a democratic deficit, whereby data that is collected and generated for some purposes is being co-opted for others, including by elected representatives.
- For example, data producers (i.e. citizens) may have trouble accessing data about themselves, while non-state entities use this data for commercial or criminal purposes, and state agencies may themselves be complicit in the corruption of infrastructure.
- Technologies to store, manage, access, analyse, and share data can optimise the use of large and complex datasets and ensure their preservation, providing Australian researchers and industries with a competitive

advantage. But to realise these opportunities, policymakers and users need to understand the limits of data and data technologies, and consumers and producers of data need to be able to trust that their information is in safe hands.

Keywords

access, analyse, big data, capture, collect, data, disseminate, generate, ICT, information, information age, interoperability, infrastructure, metadata, open data, research data, re-use, share, standards, store, technologies for data, use

---

This paper discusses the history of information and technologies for information, challenging the idea that the late 20th and early 21st centuries are distinguished from other periods for being an information or data age; the notion of a digital revolution and the development of a digital economy and its increasing entanglement with the rest of the economy (see also (Douglas, 2015); types of data problems including 'big' data, open data, research data collections, and the interlinked problems of privacy, security and surveillance; and the technologies that are needed to store, manage, access, analyse, share and regulate data to optimise the use of these collections for Australia.

The paper does not dwell on the distinctions between data and information. For these purposes, data are items of 'information considered collectively', typically 'used for reference, analysis, or calculation', or 'something given or granted; something known or assumed as fact, and made the basis of reasoning; an assumption or premise from which inferences are drawn'.[1] So data and information describe or represent the world. Since at least the 1960s, data as raw numbers and facts have sometimes been distinguished from information as processed data, particularly in the domains of knowledge management and informatics, but this distinction is not particularly useful here. Data need infrastructure, or technology, for their capture, generation, collection, manipulation, analysis, use, dissemination, and storage. This data infrastructure may include data itself. Metadata as a type of data that characterises other data is a kind of infrastructure, for example (Research Data Infrastructure Committee, 2014). Legal and regulatory frameworks, licenses, and industry standards at all points of the data lifecycle and across sectors are also technologies and infrastructure for data (Korte, 2014) (see Table 2).

In 2014 the Australian Government Minister for Communications, the Hon. Malcolm Turnbull MP, observed that 'the cost of storing, transmitting and processing or analysing data is cheaper today than it has ever been'. Over the past half century, 'the cost of digital storage has halved roughly every two years, storage density has increased 50-million fold and our ability to process that data has increased exponentially, doubling every eighteen months' (Turnbull, 2014) in accordance with advances in the microchip industry roughly in line with Moore's Law (see also(Douglas, 2015)). In 1980, storing a single gigabyte of data cost about $US440,000. In 2014, at about 5 cents per gigabyte, you could store around 8.8 petabytes[2] (8.8 million gigabytes) for $440,000 (Korte, 2014). Turnbull observed that this explosion in the amount of data potentially available in machine-readable form to consumers, business and industry, and government presents unprecedented opportunities. He noted that in 2013 the amount of stored information globally, at around 1200 exabytes[3], 'is the equivalent of giving every person living on Earth today 320 times as much information as is estimated to have been stored in the Library of Alexandria' (Turnbull, 2014); (Mayer-Schönberger et al., 2013). (Insofar as such a claim is testable, it raises the interesting notion of a 'Library of Alexandria' as a somewhat coarse-grained unit of information storage.)

The ubiquity of data collection and generation, and improvements in networking and interoperability, provide challenges and opportunities to governments, security agencies, regulators, industry, businesses, researchers, and citizens (these are not mutually exclusive categories).

---

[1] Data, *n.* and datum, *n.*, *Oxford English Dictionary online*, Oxford University Press, accessed 9 September 2014.

[2] A petabyte is 10^{15} or a thousand million million bytes.

[3] An exabyte is 10^{18} or a million million million bytes.

Technological solutions and cultural changes together may help realise opportunities and mitigate problems. With the right tools, including standards and regulatory settings, large and complex datasets can help to address grand challenges, open new industries, optimise existing industries and business models, change the workplace and how people spend their leisure time, improve health and safety, anticipate and mitigate disaster, and foster whole new fields of research. But data and data technologies present their own challenges, including problems of security and privacy, changing social capital, as well as questions of ownership, understanding, provenance, trust, preservation, access, equality, and cost, since of course, these Alexandrian libraries are not evenly distributed at 320 per 'person living on Earth today'. Instead, they are concentrated in the files of government agencies and archives, universities, corporations, and private collections, buried in literature not widely or digitally available, or on the computers or notebooks of individual researchers, their availability utterly dependent on access and know-how. Even when stored in machine-readable form, these machines do not necessarily speak to each other – nor will potential consumers have the resources to make sense of the data. And there is a further power imbalance between the people whose behaviour generates much of the networked personal information that supports the 'big' data economy, and those state agencies, corporations, and criminals who collect and use such data (e.g. (Vastag, 2011).

These issues are considered in the context of the major themes of *Technology and Australia's Future*, the fifth of a series of reports on securing Australia's future. In particular, technologies for data are considered for the illumination they provide to project question 5: 'What are the opportunities, barriers and determining factors for new or different uses of modern ICT broadly across Australia's security, cultural, democratic, social and economic systems?'

Information in a digital age

The data accumulated from centuries of observation, accelerated by digital ICT innovation over the last half century, continue to transform the processes and nature of knowledge discovery and the uses to which information is put. The development of computers capable of building detailed simulations and solving huge numbers of equations very rapidly has enabled researchers to discover and investigate fields of study previously impervious to experiment and direct observation. Data outputs from sensor networks and from simulations of complex phenomena have increased enormously, and in turn feed back into simulations, generating more data (Douglas, 2015). Widespread use of electronic shopping, payment, and mobile communication systems, and increasing dependence on them, generates still more data about collective and individual behaviour. When data is generated by technology or infrastructure users, or when the data generated is paid for in whole or mostly with public funds, issues of ownership, access, and the right to use arise. This issue is reflected in the growing open data movement, which endorses the principle that certain sorts of data should be free to access, use, modify, and share, 'subject, at most, to measures that preserve provenance and openness'.[4] Whole systems depend on electronic data and data technologies, and it is virtually impossible to engage socially and economically in the developed world without leaving a digital trail (see (Kessler, 2013) (Gibbs, 2014). This dependence and ubiquity have enormous policy implications.

In their ubiquity and persistence, information and communication technologies (technologies for data) are general purpose technologies analogous to electricity infrastructure (Lipsey et al., 2005). Data themselves, the flow of knowledge or information about the world, could be considered akin to electricity, not just for the quality of being transmittable over a wire, but for their ambivalent, context-dependent state as both technology and 'fuel' or content. Take the case of metadata ('data about data'): some information can be data or metadata, depending on the enquiry. Information which is metadata is part of the structure of a data technology. For example, while conveying little to

[4] The Open Definition version 2.0 http://opendefinition.org/od/.

no information about an envelope's contents, an address is an essential postal technology: to the letter's recipient, it is postal metadata (Griffiths, 2014). But to a demographer, geographical details are primary information, the point of the enquiry, not its context, and to a philatelist, stamps and envelopes and the information they embody are the principal material artefacts of interest (Figure 1). So depending on context, it's all data, and it's all technology. Rather than attempting an arguably futile distinction between metadata and data, it may be more useful to categorise it all as 'data' and to be very clear about which data is being regulated.

Figure 1: 'Metadata is the material on the front of the envelope' (Griffiths, 2014): information technology, postal metadata, or philatelic primary text?[5]

Many factors affect data use and the adoption of technologies for data. These influences include data ownership, standards and metadata, access, management, regulatory frameworks, privacy and security. The availability of electronic data and of technologies for storing, managing, sharing, using, and analysing data has encouraged research which is increasingly data-intensive (Research Data Infrastructure Committee, 2014). Its availability has also opened fields of enquiry in industry and business.

This rapid acceleration in the quantity and availability of electronic data, and the greatly increased opportunity for data collection, generation, manipulation and analysis has been referred to as 'the data deluge', 'big data', 'the fourth paradigm', 'the petabyte age', and 'the data revolution' (for example, (Anderson, 2008), (Hey, 2010), (Hey et al., 2009), (Gray, 2009), (Strawn, 2012), (Wigan et al., 2013) See also (Douglas, 2015)). However, as other commentators have suggested, an abundance of information does not distinguish the present from the past. At least since the advent of language, human beings have been inundated with information about the world, and technologies for controlling information have been tools both for emancipation and to entrench existing power for almost as long (Gleick, 2011) (Scott, 1998Kindle edition) (Standage, 1998) (Wright, 2007). It is perhaps more apposite to describe the beginning of the 21st century as an age of interoperability, in which systems and infrastructure for data generation, collection, analysis and exchange are networked and can understand each other.

[5] Source: 'Zeppelin mail-1934 Xmas flight Gibraltar-Brazil' by UK Government for their British Overseas Territory of Gibraltar – http://www.bephila.com/sites/default/files/pictures/Gibraltar/gibraltar-zeppelin-mail-001.jpg linked from this webpage. Licensed under Public domain via Wikimedia Commons - http://commons.wikimedia.org/wiki/File:Zeppelin_mail-1934_Xmas_flight_Gibraltar-Brazil.jpg#mediaviewer/File:Zeppelin_mail-1934_Xmas_flight_Gibraltar-Brazil.jpg

James Gleick has described the recent electronic age as a transformation akin to the 'emergence of literacy itself', implying that it is the medium of communication rather than the generation and storage of data that distinguishes one disruptive data technology from another (Gleick, 2011) (a disruptive technology is a product, process or system which, over years or decades, generates significant social and economic change and in the process may displace other products, processes or systems). In other words, people have probably always collected, generated, stored and communicated data, at least since the beginning of literacy and arguably since the beginning of language. Against the background of this use, people have developed a variety of products and processes which enable these uses. In this view, the invention of writing and reading enabled information to be 'fixed', allowing the imposition and diffusion of ideas about classification, taxonomy, reference, and definition. The electronic age has expanded these possibilities and added simultaneity (new information can be shared and communicated almost instantly) allowing an unprecedented amount of cross-referencing, changed the nature of the data being collected, and introduced a degree of interoperability unmatched in human history.

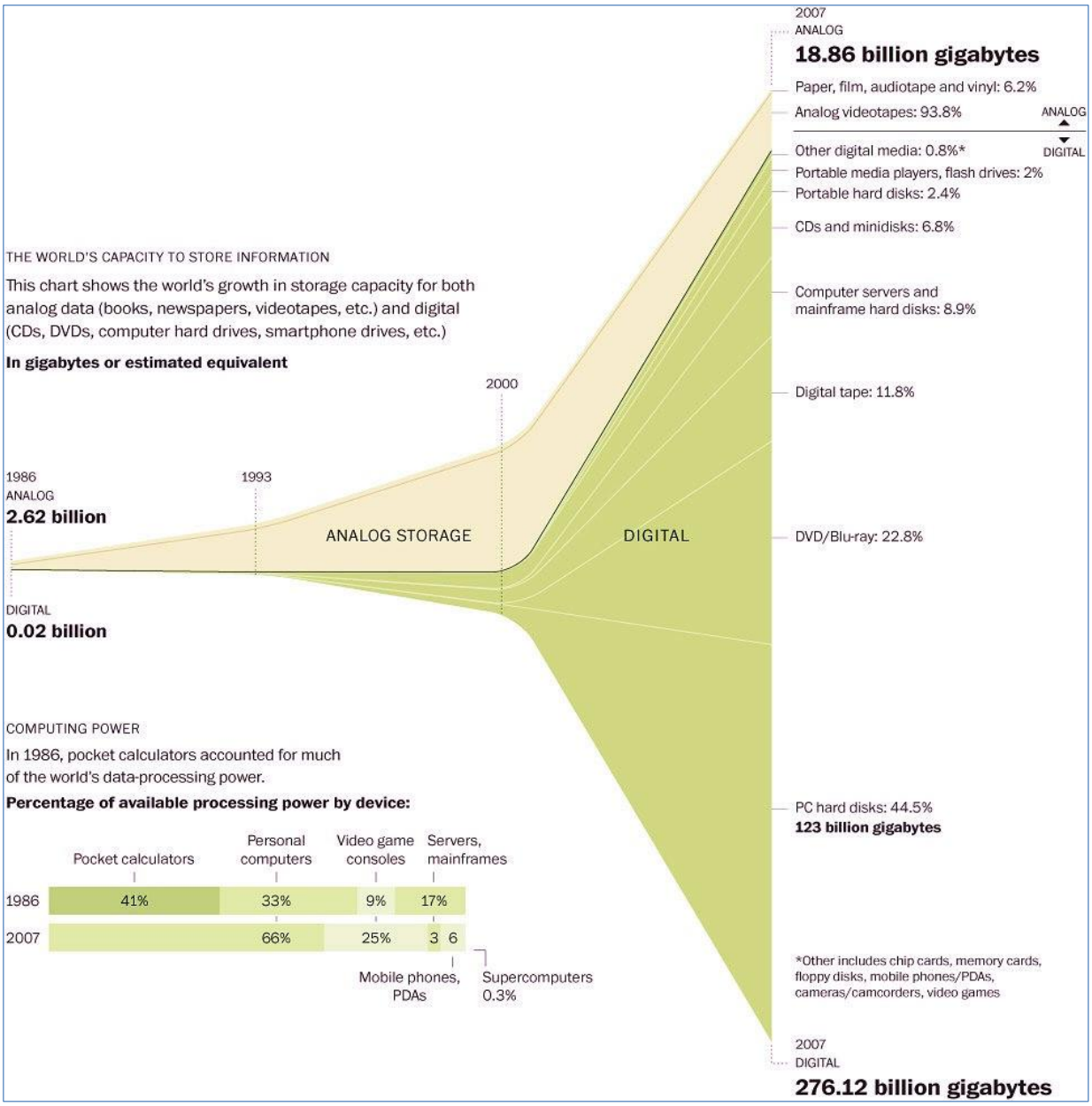
In contrast to the importance Gleick (Gleick, 2011) attributes to digital ICT, Tom Standage (Standage, 1998) suggests that the invention and global adoption of the telegraph and associated technologies was a greater discontinuity than the rise of digital communications of the last 30 years. This is consistent with the idea that changes in ICT and infrastructure, rather than the information itself, affect radical change. The real-time global communication or 'highway of thought' supported by the electric telegraph was a qualitative shift in the way information and data are communicated. Email, the internet and related technologies, argues Standage, have merely extended this capacity. Invoking another iconic and so-called disruptive information technology he claims that 'the telegraph unleashed the greatest revolution in communications since the development of the printing press' (Standage, 1998). (See also Table 1.)

Consistent with Gleick's and Standage's theses, designer and author Alex Wright (Wright, 2007) has written that 'mystical beliefs about technology are nothing new', citing 'techno-prophets' HG Wells, the philosopher and Jesuit priest Pierre Teilhard de Chardin, and communication theorist Marshall McLuhan enthusing about an emerging planetary intelligence in the mid-20th century. Wright has observed that, while 'some apostles of digitization argue that the expanding global network will do more than just improve people's lives' by changing 'the shape of human knowledge itself', few writers have acknowledged 'the contours of a broader information ecology that has always surrounded us'. This is because 'as humans had no concept of oral culture until they learned how to write, so the arrival of digital culture has given us a new reference point for understanding the analog age' that researchers are only now beginning to interrogate.

Allowing that Standage is correct on the question of qualitative change and the disruptive effect of telegraph technologies compared to modern ICT, and Gleick and Wright are correct about literacy setting a precedent for the beginning of the 'information age' roughly 5000 years ago, it is also the case, as described by Gleick, that modern digital technologies have effected an enormous quantitative shift in the amount of data available, their potential interconnections, and speed of communication (Douglas, 2007, Gleick, 2011) (Douglas, 2015). Such technologies include the Global Positioning System, networked sensors, and wireless technologies. This self-perpetuating acceleration is sometimes termed big data, the data deluge or the digital age (see, for example, (Hilbert et al., 2011), who make a case for the 'digital age' beginning in 2002, when the amount of data stored digitally reached 50% (Figure 2) (Press, 2012). Even this case is not clear-cut: for instance, one could perhaps equally validly date the beginning of the digital age from 1996, the date

at which digital data storage became more cost-effective than paper data storage (Morris et al., 2003).

Figure 2. The world's capacity to store information, 1986–2007[6]



The following section discusses the history of information and early 'technologies' for information, including writing, arguably the first complete information technology (see Table 1 for some other examples).

[6] Credit: Todd Lindeman and Brian Vastag/ The Washington Post, http://www.washingtonpost.com/wp-dyn/content/graphic/2011/02/11/GR2011021100614.html. Source: (Hilbert et al., 2011).

Table 1: Information technology highlights since the Palaeolithic[7]

Date range	Information and communication technology	Place invented, patented, disseminated or found in the archaeological record
Palaeolithic, Mesolithic	symbolism, pigment, sculpture	global
BCE		
3400–3200	hieroglyphics	Egypt
3200	cuneiform	Mesopotamia (Sumer)
3200	clay tablets, stylus (for cuneiform)	Mesopotamia (Sumer)
3200	writing[8]	Indus valley
3100	papyrus	Egypt, Southern Sudan
2000	alphabetic writing	Near East
2000	musical notation	Near East
18th to 13th centuries	Cretan hieroglyphics, Linear A, Linear B	eastern Mediterranean, Aegean
14th century	wax tablets (earliest evidence, though used until the Late Middle Ages in Europe)	Greece
1200	writing	China
1050	Phoenician alphabet	Phoenicia (Near East)
10th–5th centuries	development of Greek and Roman alphabets from Phoenician alphabet	Mediterranean, Aegean
600	Mesoamerican scripts	Mesoamerica
before 5th century	parchment (and later vellum)	Greece, Asia Minor
5th century to present	codices	Greece, Rome
4th century	reed pen	Egypt
300	ink brush	China
2nd century	paper	China (reached western Europe in 10th century CE)
CE		
6th–19th centuries	quill pens	Western Europe
1041–1048	first known movable type printing technology	China
14th century	water-powered paper mills rapidly increased the production of, and demand for, paper, lowering the cost	Western Europe (first reference Spain)
14th–16th centuries	writing slate (earliest records): use becomes widespread during the 19th century and wanes during the mid-20th century.	UK (from Welsh slate), then global
1450	Gutenberg press allowing assembly-line production of text	Germany

[7] Information from *Encyclopædia Britannica* online or Wikipedia unless otherwise specified.

[8] David Whitehouse, 1999. '"Earliest writing" found'. BBC online, 4 May. http://news.bbc.co.uk/2/hi/science/nature/334517.stm

Date range	Information and communication technology	Place invented, patented, disseminated or found in the archaeological record
1474	granting of patents by the state (Venetian statute): in Venice, new and inventive devices had to be communicated to the Republic in order to obtain legal protection against potential infringers for a period of 10 years.	Italy (Republic of Venice)
17th century	fountain pen (ink reservoir)	Europe
1725	punched cards	France
1792	semaphore telegraph	France
1796	lithography	Bavaria
1801	Jacquard loom (punch cards)	France
1806–1808	carbon paper	Britain, Italy
1810	steam press patented	UK
1822	mass-production of steel-nib pens (fountain or dip pens)	Britain
1820s	camera photography	Western Europe
1832	open source machine and proposal for punched cards for informatics and information storage – Semen Korsakov	Russia
1832	Charles Babbage describes how to design a form (Babbage, 1832)	Britain
1837	patents for Cooke and Wheatstone's electrical telegraph system (UK), Morse's system (US)	Britain, US
1839	daguerreotypy	France
1840	world's first standardised adhesive postage stamp (Penny Black)	UK
1843	first 'fax' patented	Scotland
1850	first undersea cable for overseas telegraphy	Britain, France
1851	Morse standard adopted for European telegraphy	Europe
1870s	commercial production of typewriters	Denmark, US
1876	Bell's patent for a telephone	US
1877	telephone exchange	US
1878	typewriter shift key	US
1880	public payphone	US
1888	first patent for a ballpoint pen	Britain
1889	patent issued for Herman Hollerith's tabulating machine, used for the 1890 US Census (punch-card based) (Rheingold, 2000)	US
1890s	radiotelegraphy, radio (Marconi)	Italy
1920s	microform commercialised (developed 1839, UK)	US
1920s	wireless telephony on trains	Germany
1924	first wireless fax transmission	US
1930s	Automatic teleprinter exchange services (eventually replaced Morse code for telegraphs)	Germany
1946	mobile telephones for automobiles	US

Date range	Information and communication technology	Place invented, patented, disseminated or found in the archaeological record
1946	first programmable digital general-purpose electronic computer, the Electronic Numerical Integrator and Computer (ENIAC), unveiled	US
1950s	IBM mass-produces electronic programmable business computers	US (other companies doing the same in the UK)
1951	Magnetic tape first used to record computer data, from UNIVAC	US
1958	mass-produced modems	US
1964	Xerox: long distance xerography – first commercialised version of the fax machine	US
1969	ARPANet – precursor to the internet	US
1969	the 'mother of all demos' – introduces Windows, hypertext, computer mouse	US
1970s	microcomputers – rise of the personal computer	US
1973	Motorola introduces the first handheld mobile telephone (weighing 1.1 kg and measuring 23 cm x 13 cm x 4.45 cm)	US
1979	first automatic analogue cellular telephone systems	Japan
1983	Patent granted for radiofrequency identification (Walton, 1983)	US
1989	Tim Berners-Lee's proposal for a World Wide Web	Switzerland (CERN)
1989–1994	GPS: 24 satellites for Global Positioning System deployed (Alexandrow, 2008)	US
1990s	wireless local area network	several, including Australia
1990s	mass-production of digital compact cameras and video recorders	US, S Korea, Japan
1996	Google PageRank proposed (Google Inc. founded 1998) (Brin et al., 1998)	US
2000	first car GPS system patented	US
2000s	convergence of cameras, video recorders, wireless platforms and mobile phones to make smartphones	US, S Korea, Japan, China

The ages of information

This section discusses the concept of an information age (and synonyms), and 'data floods' of the past. The data deluge or flood, also known as the information explosion or, more recently, 'big data', refers to increasing data generation, collection and exchange per unit time. The phenomenon pre-dates the electronic age.[9]

From Gutenberg's press to the Overland Telegraph to Wikipedia, the history of information is the history of technologies for information and for data (see examples in Table 1). Perhaps the disruption at the dawn of the current 'information age' could be traced to the use of the Electronic Numerical Integrator and Computer (ENIAC), the first electronic general-purpose digital programmable computer, in 1946 in the Manhattan Project, which convinced the US Defense Department that its unprecedented computing power held the key to a range of military

[9] The *Oxford English Dictionary*'s earliest recorded use of the phrase 'information explosion' comes from an Oklahoma newspaper, the *Lawton Constitution*, in 1941. In *The Information* Gleick points to earlier uses.

applications; or to 'the father of information theory', mathematician and engineer Claude Shannon's breakthrough work for Bell Telephone Laboratories in the 1940s, which coalesced in 1948 into his paper 'A mathematical theory of communication' (Shannon, 1948), and the first recorded use of the word 'bit' (a portmanteau contracting 'binary' and 'digit') to describe a basic unit of information; or to Intel co-founder Gordon Moore's observation in 1965 that the number of transistors on an integrated circuit doubles approximately every two years; or to Tim Berners-Lee's 1989 proposal at CERN for the information management system which led to the World Wide Web (Ceruzzi, 1986); (Moore, 1965); (CERN, 2008); (Gleick, 2011) (see also (Douglas, 2015)). With hindsight, defining moments are a dime a dozen.

Gleick suggests that the alphabet is the 'founding technology of information'. Following from this first innovation in communication – the written word – codices, almanacs, calendars, paper, maps, chapbooks, newspapers, pamphlets, dictionaries, wax cylinders, photography, ballpoint pens, telegraphy, typesetting, forms and surveys, telephones, microform, smart phones, tablets, computers, the Global Positioning System, the internet, and cloud technology are, as Gleick insists, only 'the latest innovations devised for saving, manipulating, and communicating knowledge'. Information 'has always been there. It pervaded our ancestors' world, too', and technologies of information 'all played their parts in weaving the spiderweb of information to which we cling. Each new information technology, in its own time, set off blooms in storage and transmission' (Gleick, 2011).

Technologies for data collection and dissemination have been important mechanisms for centralising states, for democracy, and for science and research throughout human history. For example, there is a long history of governments and bureaucracies collecting information on their citizens in the form of censuses (Egypt in the third millennium; Greece around 1600 BC; mentioned in the Bible; India 300 BC; Roman Empire). Scott refers to the census as a marker of modernism (Scott, 1998). Such endeavours required numeracy, literacy and the means to record. Inspired by the need to modernise the US Census during the late 19th century, Herman Hollerith's tabulating machine adapted an older engineering technology – punch cards – to transform data collection and helped lay the foundations of IBM's later market dominance in the business information industry (Douglas, 2015, Rheingold, 2000). Technologies which helped disseminate and democratise knowledge stimulated their own dissemination: Gutenberg's press helped drive the democratisation of knowledge in early modern Europe. Its invention and dissemination preceded an increase in public literacy and numeracy, which in turn increased the market for printed materials. Other exercises in the construction of information networks and technologies for data could be seen as forms of crowd-sourcing or citizen science: compiling of the first edition of the Oxford English Dictionary took more than 70 years and involved hundreds of unpaid volunteers (Oxford English Dictionary (OED), 2013). Nineteenth-century armchair scientists whose networks of field scientists, collectors and informants spanned the known world relied on information and communication technologies including the mail and the telegraph: Joseph Banks, Charles Darwin, George Cuvier, Richard Owen, Joseph Hooker, Charles Lyell and many of their contemporaries built their scientific reputations from the laboratory or 'armchair', acting as a sort of information fulcrum, collecting the observations and specimens sent them by a vast network of informants and protégés.

The telegraph (Standage's 'Victorian internet') and undersea cables allowed virtually instantaneous communication between distant places and transformed a number of scientific disciplines and industrial processes, including revolutionising railways and the calculation of time (Standage, 1998). In colonial South Australia, as Superintendent of the Electric Telegraph Department, Charles Todd conceived and oversaw the construction of the Overland Telegraph from Adelaide to Darwin, which then connected Australia to India and thence England by undersea cable, reducing the isolation of Australia's colonies from each other and from other parts of the British Empire. Todd was also Government Astronomer, which gave him responsibility for colonial meteorology. He used a network of meteorological stations along the telegraph, and informants at stations in South Australia

and the Northern Territory, to collect detailed meteorological records and send the data to the Adelaide Observatory via telegraph. This innovation began to revolutionise weather forecasting in Australia, and Todd worked closely with his Government Astronomer colleagues in the other Australasian colonies to produce a synoptic view of Australia's weather patterns (Douglas, 2007). So the telegraph became an essential tool for the collection and communication of scientific data. One hundred years after the European discovery of the platypus, on 29 August 1884, Professor Archibald Liversidge, Chair of Chemistry and Mineralogy at the University of Sydney, relayed naturalist William Caldwell's message to the British Association for the Advancement of Science, then meeting in Montreal: 'Platypus oviparous ovum meroblastic'. Historian of science Roy MacLeod has called this episode 'a turning point in the history of science in the Pacific', and it testifies to the importance of the telegraph as a tool for sharing scientific information (MacLeod, 1994).

Microform was the great medium for data / knowledge storage for much of the 20th century, and remains an important source of historical data today (see HG Wells and Vannevar Bush's predictions about technologies for knowledge dissemination, which assumed microfilm and pneumatic tubes as foundation technologies for processes that are now digital and wireless). So entrenched was microform that when speculative fiction writer and social commentator HG Wells (in 1936 (Wells, 1938)) and Director of the US Office of Scientific Research and Development Vannevar Bush (in 1945 (Bush, 1945)) imagined something like the World Wide Web, they assumed it would develop from microfilm. Both writers anticipated a microfilm-based centralised repository of human knowledge or collective memory, either worldwide (Wells's 'World brain') or decentralised (Bush's 'memex'), with mechanical delivery systems via pneumatic tube. Rather than predict the future of material technology or processes, it could be said that the two men correctly foretold a need or a use or an outcome, while remaining thoroughly enmeshed in their contemporary technical and economic contexts. Bush's essay did indirectly shape information and communication media by influencing a young Douglas Engelbart, who had Bush's 'memex' in mind when he began the work that would eventually contribute to the invention of the computer mouse, word processor, and hyperlinks, but the processes and material products were worlds away from those Bush envisaged (Rheingold, 2000).

The rise of a fourth paradigm?

In a similar vein to Gleick, Standage and Wright, Robert Darnton recently characterised the following 'currently accepted wisdom' about information and digitisation, which appears 'on talk shows and Op-Ed pages every day', as false, misleading, or trite (Darnton, 2014):

1. *We live in the information age*
   'Every age was an age of information in its own way'. For example, in 18th-century Paris, the flow of talk, images, printed material, and songs 'shaped a collective consciousness' that 'swept through the streets … in waves of 'popular emotions'', or riots, as potent as 'anything spread today by Twitter'.
2. *All information is available online*
   'We have digitized only a small portion of the books in our libraries … Millions more cannot be located or have disappeared, and most information never made it into books, to say nothing of modern databases'. France's Archives nationales and one hundred provincial archives contain more than 2000 miles of documents. 'Still more can be found in municipal archives, various university archives, and private collections. Most of it has never been read, much less scanned'.
3. *The future is digital*
   'More books are produced in print each year than the year before – an increase of 6 percent in the United States in 2012 … To imagine a future in which the digital destroys the analog is to misunderstand current trajectories and the history of communication in general. New media do not extinguish old ones, at least not in the short run. Instead, they enlarge and

> enrich the information landscape … It is wrong to imagine digital technology flattening out every other mode of communication'.

In support of Darnton's claims, Hilbert & Lopez's data (Hilbert et al., 2011) shows that while digital storage has outstripped analogue since early this millennium, more information is stored in paper books today than was the case 20 years ago when the adoption of digital storage technologies began to accelerate (Figure 2). Vastag also notes Hilbert's observations that, while human activity generates enough data 'from TV and radio broadcasts, telephone conversations and, of course, Internet traffic' to fill the world's total digital storage capacity every eight weeks, 'most of the digital traffic is never stored long term, evaporating into the ether' like Darnton's lost books (Vastag, 2011).

In 2007 Microsoft computer scientist Jim Gray defined four successive stages (or 'paradigms') of western science (Douglas, 2015). The first stage of scientific knowledge was experimental and *ad hoc*. The second arose during the Renaissance, and involved testing theory with empirical data derived from experimentation. In the third, scientific knowledge is shaped by computer modelling and simulation (presumably in addition to theory and physical experimentation). Gray called the fourth stage, which he described as 'post-theory', data-intensive research ((Gray, 2009) (Hey, 2010); (Hey et al., 2009); (Strawn, 2012). See also (Anderson, 2008) and Box 1).

Gray's four paradigms obscure and mischaracterise the complexity of western scientific tradition. Considering the increasing weight (since at least the sixteenth century) given to empirical approaches reliant on experimentation and method, which undermined the then dominant Aristotelianism, with its dependence on logic and syllogism, it might seem that Gray's second paradigm, theory, preceded his first, experimentation (Sgarbi, 2013). Bearing in mind that many nineteenth-century scientific projects (exemplified by Charles Darwin's efforts leading to the publication of *On the Origin of Species by means of Natural Selection*, or Joseph Hooker's pioneering work on geographical botany) relied on amassing and synthesising a vast amount of data from an international network of informants, it might appear that a kind of 'big data' science, the fourth paradigm, preceded Gray's computational science, his third paradigm. This assumption too is overly simple. Gray's four paradigms should be considered more like a braid, weaving through the history and into the future of human information gathering and knowledge making, reinforcing and supporting each other.

The idea of a simplification or analogy (a model or simulation) as a basis for experimentation to reveal information about a larger system existed in classical antiquity (Douglas, 2015). Indeed, arguably, human beings only ever understand complex phenomena by modelling them. The development of computer simulations to study phenomena of great size and complexity has transformed modelling to revolutionise scientific discovery and research (Winsberg, 2010); (Hey, 2010); (Douglas, 2015). The 'third paradigm' has not supplanted theory, experimentation, and non-computational modelling and simulation, but has enhanced them, increasing the amount of data generated, captured and collected, which in turn feeds increasingly complex models and theories, which are tested empirically – perhaps a more appropriate characterisation of Gray's fourth paradigm.

While Gray's delineation of the progress of scientific knowledge as proceeding from experiment, through theory and modelling to data science is simplistic, even wrong, it neatly exemplifies Darnton's 'currently accepted wisdom' about information, in its implication of both the singularity of late modernity and its continuity with pre-electronic and pre-industrial society. Gray's model may help to suggest what is different about the current 'information age', if anything, and whether, as Gray and his colleagues have suggested, a quantitative change in the amount of data generated and collected electronically has resulted in a qualitative shift in the nature of information and research, and indeed of the scientific method (see for example (Anderson, 2008)).

Box 1: The world has changed

'The world of science has changed, and there is no question about this. The new model is for the data to be captured by instruments or generated by simulations before being processed by software and for the resulting information or knowledge to be stored in computers. Scientists only get to look at their data fairly late in this pipeline. The techniques and technologies for such data-intensive science are so different that it is worth distinguishing data-intensive science from computational science as a new, fourth paradigm *for scientific exploration'.*

(Gray, 2009)

Gray's model of western science since Antiquity goes something like this:

Middle Ages and before:
1st paradigm. [ad hoc experimentation]

Renaissance:
2nd paradigm. [theory] → [experimentation] → [empirical data] → [theory]

Digital age:
3rd paradigm. [theory] → [experimentation] → [empirical data] → [digital modelling and simulation] → [data] → [theory]

Data age:
4th paradigm. [data capture or data generation] → [processing by software] → [storage in computers and use* by scientists]

*presumably this 'use' includes the third paradigm progression of hypothesis to experiment to model to theory, although Gray does not specify

One change over the last 25 years resulting in the conditional availability of much more information is the rapid increase in the collection of data about people from diverse sources (for example, sensor networks, CCTV, radiofrequency identification (RFID) tags, the recent revelations by Edward Snowden about the US National Security Agency (NSA) collection of mobile telephone data (or 'metadata' as it is misleadingly described) and the persistence of the data trails their behaviour generates. Data in some form has always been important to researchers. Perhaps the change Gray observed back in 2007 is a sociopolitical shift rather than a scientific one. The perception of data as increasingly commercially important during the 21st century has mirrored its collection and use by security agencies, police forces, and hackers. A number of commentators and privacy advocates have spoken of the need for greater responsibility by data holders. For example, Wigan and Clarke observed that business and government exploit data 'without regard for issues of legality, data quality, disparate data meanings and process quality' (Wigan et al., 2013). This attitude 'results in poor decisions, the risks of which are to a large extent borne not by the organisations that make them but by the individuals who are affected by them'. According to Wigan and Clarke, the 'threats harboured by Big Data extend far beyond the individual … into social, economic and political realms. New balances must be found to handle these power shifts'. Is the fourth paradigm inevitably accompanied by a decline in privacy and a rise in surveillance and misbehaviour? Wigan & Clarke think the horse has bolted, but propose that the development of a 'private data commons' may provide a suitable framework for managing personal information in an interconnected data ecosystem (Wigan et al., 2013). Tim Berners-Lee and colleagues are among others also working on data accountability, which lays responsibility for data governance in the hands of the collectors and custodians.

Another concern is the question of what one is supposed to do with 'all this data' (Hand, 2007) (Vastag, 2011). If technologies for data cannot communicate with each other, the ability to analyse what has been collected or generated and relate it to other data is severely impeded (see Box 2). Some commentators warn of potential flaws and errors which, while are not intrinsic to 'big' data, are endemic to all statistical and data analysis, and are not mitigated by the availability of more data (Vastag, 2011) (Harford, 2014) (Lazer et al., 2014), (Marcus et al., 2014), (KNC, 2014), (Leek, 2014). Having more data available does not improve analysis in a simple linear way, and in fact the risk of identifying spurious correlations increases with more data (see (Lazer et al., 2014) on Google Flu

Trends). Some champions of 'big data' have claimed that the availability of sufficiently large datasets will render theory obsolete (see for example Chris Anderson's influential essay in *Wired* in 2008: 'The end of theory' (Anderson, 2008)). Gray's ideas about the fourth paradigm could uncharitably be read this way, and some of his supporters certainly understood him thus. But subject-area knowledge is important for dealing meaningfully with datasets no matter their size. The air-time these critiques attract may reflect where in the hype cycle 'big data' currently lies (for example, (KNC, 2014)), but it may also be a function of widespread confusion about the benefits and treatment of data, and technologies for data, across government, the research and private sectors, the commercial domain, and wider society.

The age of interoperability

Rather than an 'information revolution', this modern data-intensive environment could be considered as a revolution in networking and interoperability.

Over the past 20 years or so gradual changes have combined to transform the way human beings think about and deal with information (Hilbert et al., 2011). But it is not the data itself that distinguishes the era so much as the increased capacity to pull together and characterise huge amounts of data from different sources, and to store, manipulate and analyse it better than ever before. For example, at CSIRO, the transformative potential of interoperability is recognised in a number of very different sectors from primary healthcare to mining:

'In healthcare, this revolution holds the promise of linking disease with the molecular imbalances at its very heart, thereby ushering in an era of customised medicine. In retailing, there is the possibility of matching people, with unprecedented sophistication, to products that fill their needs. In the minerals industry, this revolution provides an ability to explore the environment more carefully and exploit it more sensitively and cheaply … through the better use and reuse of data and the integration and optimisation of equipment' (Thwaites, 2013).

There is a long way to go to exploit this potential, particularly in developing standards for both equipment and data recording, and for protecting personal information where appropriate.

Box 2: Improving interoperability for Australian geosciences

Much of the geoscience data collected in Australia over more than a century is being made accessible on the internet through AuScope, a non-profit company established six years ago by Australia's governments and universities, CSIRO and the minerals industry. Using high-speed computer links, the seven state and territory geological surveys and Geoscience Australia … are now connected into the AuScope Grid and the data they hold can be browsed from anywhere in the world using the AuScope Portal. Information of interest can then be manipulated and explored using tools available online in a Virtual Geophysics Laboratory.

(Thwaites, 2013)

For the last 70 years supercomputers, the fastest computers available at a given time, have produced 'big' data output from running simulations, as well as analysing 'big' datasets (Douglas, 2015), but now 'big' data outputs are derived from many sources which for the most part talk to each other, generated by the economic and social behaviour of consumers and ICT users as well as researchers, archivists, and governments (Strawn, 2012). However, practical, cultural, and commercial barriers persist, despite the apparent appeal of integrated data and interoperable technology. Privacy concerns and regulation may limit the degree to which personally or commercially sensitive data is available. And incentives remain for equipment manufacturers to design and data-holders to use proprietary technology which limits interoperability, and to quibble over standards. The effects of these barriers to limit the opening, sharing and use of data has spurred open data advocates, government agencies like CSIRO, and organisations like the Australian National Data Service (ANDS) and the international Research Data Alliance to encourage the development of international standards for data measurement and infrastructure (e.g. (Thwaites, 2013) (Wilkinson et al., 2012)).

Technologies for data in an age of interoperability

Box 3: The cost of knowledge

The reason big data has recently become important is that disk storage is now cheap enough that scientists can afford to store massive amounts of data. When disk storage was introduced in the 1950s, the cost was about one dollar per byte. But keeping the cost the same, disk capacities have doubled every year and a half, even faster than Moore's law for the number of transistors on a chip. We passed the gigabyte-per-dollar threshold in the last decade. Later this decade, a terabyte of disk storage may cost one dollar!

(Strawn, 2012)

It cost around $440,000 to store a single gigabyte of data in 1980, while today it costs about five cents. As of 2013, more than a billion gigabytes of data were stored in the cloud alone.

(Korte, 2014)

These figures illustrate a more general point about how technologies change: the reduction in price of a single component among many may have a big impact on the larger system.

Data is a general-use constituent of a general use technology (information and communications technology). The impact of technologies for data comes about from their interactions with other technology and with data (e.g. storage technologies and generation technologies, ensuring data generation has a multiplicative effect: technologies for data are themselves mechanisms to improve interoperability.

Early methods for extracting patterns from data include Bayesian analysis (1700s) and regression analysis (1800). Their modern iterations are known as data mining or analytics – discovering patterns in large datasets through techniques of machine learning, statistics and analysis. The intent is to extract information from data and transform it into structured material that is understandable and can be used for other purposes. However, commentators have observed that 'the modern obsession with measurement has made us none the wiser. We collect, store, process, and analyze more information than ever before – but to what end? Aristotle's wisdom has never been more relevant than it is today: the more we know, the more we know we don't know' (Fung, 2010 Kindle edition). 'Data mining' used to be a pejorative term, and remains so to some statisticians, for whom it implies using data to justify a preconception rather than test a hypothesis (Grover et al., 2008). Data are not neutral, and increasing the size of the dataset will not overcome the issue of neutrality. The questions and the analyses help to shape the answers. Researchers need to ask the right questions in the right way, using the right equipment, in order to extract meaning from information.

As economist Tim Harford puts it, ' "big data" is a vague term, often thrown around by people with something to sell' (Harford, 2014) There are many working definitions of 'big data'. The Oxford English Dictionary records an early use in 1980 (Tilly, 1980), referring to statistical methods in history and historiography, but it is not until 1997 that the first use of the phrase appears in the Association for Computing Machinery's ACM digital library (Cox et al., 1997), according to (Press, 2012)). Cox and Ellsworth define 'the problem of *big data*' as occurring when 'data sets do not fit in main memory (*in core*), or when they do not fit even on local disk' and observe that 'the most common solution is to acquire more resources.' Their 1997 definition was closer to a description of very large research data collections than to the current popular notion of big data as indistinguishable from business analytics: they predicted that input datasets were 'expected to scale with the ability of supercomputers to generate them'. The term began to be formalised from the late 1990s and, consistent with Cox and Ellsworth, referred to the enormous datasets generated by astrophysics, genomics and internet search engines, 'and to machine-learning systems that work well only when given lots of data to chew on'. Harford distinguishes between huge research datasets, as produced for example by the Large Hadron Collider, and 'the 'big data' that interests many companies' and which underpins the internet economy, which he suggests could be called ''found data", the digital

exhaust of web searches, credit card payments and mobiles pinging the nearest phone mast' (Harford, 2014).

It is only in mid-2011 that use of the term in Google's search engine really takes off, and at approximately the same time, the use of 'big data industry' increases rapidly, according to Google Trends. Today big data is 'the application of data-analysis and statistics in new areas, from retailing to human resources' (KNC, 2014). It is more widely understood outside the research sector as referring explicitly to the kinds of datasets generated by the activity of human beings interacting with electronic sensors and aggregators, including the internet, but can be generally regarded as referring to any large datasets that are difficult to process and analyse using conventional methods. Given that different populations of data users and creators use the expression 'big data' to refer to distinctly different data phenomena, there is little potential to define the term in a way that is both inclusive and comprehensive (Yarkoni, 2014).[10] Big data may refer to a large volume of data, structured and unstructured ('found'), that cannot be processed using conventional database techniques, or it may refer to the technology that organisations use to capture, manage, and store these data holdings. Sometimes 'big data' just means data analytics. According to Villars, big data technologies 'describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis'. By definition, 'the rate of Big Data growth exceeds the capabilities of traditional IT infrastructures' (Villars et al., 2011). While these and other examples indicate a classificatory instability, it seems that the requirement that data volume and speed of collection exceeds processing capacity is common to most ideas of big data. By this definition, volume itself at any given time is not relevant: it is volume relative to available storage and processing power that determines whether data is 'big' (Press, 2012). By this understanding, ENIAC and its immediate successors were designed to cope with a big data problem, and in turn contributed to a new 'flood' of data (an argument which would be consistent with the timing of the Oxford Dictionary's earliest recorded use of 'information explosion' in 1941) (see also (Douglas, 2015)), indicating that technology and information are mutually constituting.

The various uses of data as currency or commodity, as an input to, tool for and output of research, as a tool for policy, control and surveillance, and as a resource for citizens and consumers, creates democratic and social tensions which both contribute to and have the potential to remedy the democratic deficit. Data that is collected and generated for some purposes is co-opted for others (for example the collection of telephone metadata by state surveillance agencies) and states are complicit in the corruption of infrastructure. Furthermore, while the availability of large and complex datasets and sophisticated analytic tools and algorithms may be transforming business and research, there is also a risk of generating 'knowledge that is "too big to know"' in the sense that researchers 'can unearth the patterns without being able to explain or understand them' (Burdon et al., 2014). Without sophisticated networked infrastructure – high-speed data networks, enormous servers, and powerful computers – as well as some standards for interoperability and data management, area

[10] Pointing to the ambiguity of the term 'big data', University of Texas psycho-informatics blogger Tal Yarkoni has suggested three discrete, but intersecting, descriptions which correspond to three ways of 'making sense' of different uses of big data in different communities, or ways 'of thinking about what makes data 'Big'. They are, roughly, (1) the kind of infrastructure required to support data processing, (2) the size of the dataset relative to the norm in a field, and (3) the complexity of the models required to make sense out of the data. To a first approximation, one can think of these as engineering, scientific, and statistical perspectives on Big Data, respectively' (Yarkoni, 2014).

knowledge, statistical training and a skilled workforce, one could be left with a costly and intrusive Library of Babel (Borges, 1964)).[11]

This section will discuss the types of technologies associated with complicated datasets, as well as costs, benefits, enablers, and impediments, with particular regard to Project Question 5, 'What are the opportunities, barriers and determining factors for new or different uses of modern ICT broadly across Australia's security, cultural, democratic, social and economic systems?' These kinds of data are 'big data' (Harford's 'found data', focusing on the commercial and social networking kind of data), open data, and data for research, acknowledging that these categories overlap in multiple ways. In the context of Project Question 5, note that most of the benefits and penalties of technologies for data will be cross-sectoral, and relevant to ICT uptake across all sectors, not just government. Furthermore, the impacts of ICT are not confined to any 'ICT sector', but transform most parts of the economy, with implications for many other technogies.

An opportunity in one context is a barrier or a penalty in another. Designers embed a particular understanding of the world into the technologies they design. This world view is 'revealed through the activities supported and encouraged' when people use the technologies (Veletsianos, 2014),

[11] The Library of Babel is the universe, conceived in an eponymous short story by Argentine writer Jorge Luis Borges as a vast library containing all possible 410-page books of a particular format. Though almost all the books in the library are nonsense, somewhere in the library is every intelligible book ever written or potentially written, and every possible version or combination of each of those books. But the inhabitants of the library have no way of distinguishing the gibberish from the coherent texts.

which are not neutral, nor are their effects unambiguously positive or negative. The benefits and harms of technology depend entirely on context.

Table 2: Some technologies for data

data technology	use	examples of type of data / user
metadata	characterisation, management, interoperability, analysis	open data (standards for access), research, government (security agencies)
cloud	storage, dissemination, management	personal, open, commercial, research, government
data centre	storage, dissemination	personal, commercial, research, government
internet	generation, dissemination, communication, storage	personal, open, commercial, research, government
sensor network	capture, collection, characterisation	personal (e.g. GPS-enabled wireless tracking), research (e.g. meteorological, ecological), government (e.g. security / metadata), commercial
loyalty program	capture	commercial
social network	capture, dissemination, communication, storage	personal, commercial, government
search engine	capture, dissemination	commercial, personal
smart phone	capture, dissemination, communication, analysis	personal, commercial, government (security / metadata), research (apps for data analysis, mod-sim)
email	communication, dissemination	personal, commercial, government (security / metadata)
online shopping site	capture, collection, analysis	personal (including financial), commercial
form / survey	collection, storage	personal, commercial, government, research
database	collection, presentation, communication, manipulation, storage, management	captured / collected data
spreadsheet	collection, presentation, communication, manipulation, analysis	financial data, numerical data
modelling & simulation	generation, manipulation, analysis	researchers, policy workers, business intelligence
computers	storage, generation, manipulation, analysis, management	personal, business, research, government
telegraph	dissemination, communication	personal, business, government
punch card	collection, storage, classification, analysis	industrial (i.e. loom, tickets), government (i.e. census), business (i.e. IBM's business machines)
microscope	characterisation	research
book	collection, storage, dissemination, communication, analysis	written information
pamphlet	communication, dissemination	written information
newspaper	communication, dissemination, analysis	written information
algorithms for statistical analysis	manipulation, analysis	commercial, research, government
visualisation technologies	communication, manipulation, analysis, dissemination	commercial, research, government

'Big' data capture, collection, and analysis

'Big' datasets arise from advances in near-ubiquitous networked digital technologies (mobile, internet, sensing devices) and promise unprecedented opportunities for understanding a wide range of issues on an unparalleled scale. However, data do not replace evidence, theory or context. As (Harford, 2014) observed, 'a theory-free analysis of mere correlations is inevitably fragile. If you have no idea what is behind a correlation, you have no idea what might cause that correlation to break down'. And even though 'the data are bigger, faster and cheaper these days', such 'found' data are 'so messy, it can be hard to figure out what biases lurk inside them' and 'so large, some analysts seem to have decided the sampling problem isn't worth worrying about'. Data scientists need to understand context in order to separate the signal from the noise. Harford concludes with the reflection that so far, big insights have not necessarily accompanied big data, and new statistical methods are needed to capitalise on big data's promise. He may be understating the value of insights arising from big data analysis: as with the early history of most general purpose technologies, its ultimate benefits may take some time to become evident (Lipsey et al., 2005). The widespread availability of personal data for commercial or security purposes also brings to a head issues of privacy, data ownership, transparency, regulation, processing costs, use, cultural interpretation, social responsibility and power. As a result of the ubiquity of personal data generation and collection, and of the interoperability of systems for generating, collecting and analysing large sets of data, people's digital behaviour can be tracked and recorded in unprecedented detail via technologies for data including social networking (e.g. Facebook), online marketplaces (e.g. Amazon, eBay), online searching (e.g. Google), webmail, 'smart' devices, and the use of wireless mobile technologies (see Table 2). New forms of data use and tools (technologies) have emerged:

- data markets, text and data-mining collaborations – e.g. Elsevier (http://www.elsevier.com/about/universal-access/content-mining-policies)
- tools for data mining, data collection, analytics: database and data management, pre-processing, statistical modelling and inference, post-processing, visualisation
- tools for storage: disks, tape, data centres, cloud.

Potential benefits of these tools include 'personalisation' of services (there is value to consumers and businesses in generating detailed understanding of customers or citizens); problem solving and predictive analytics to improve insights and decision-making, increase productivity and innovation; marketing efficiencies; health and social policy improvements from demographic data. Potential problems for the data provider or service user and wider social costs include loss of privacy, reduction in serendipity, security, social capital. Potential problems for the state include security (trade-off with increased surveillance – arms race); reduction in serendipity-driven innovation; loss of social capital from perception of exploitation of personal data. Potential problems for business and industry include bad results due to misunderstandings about the capacity of 'big' data analytics, loss of social capital and increased regulation due to privacy and security breaches (real or perceived). 'Big' data can also pose challenges to internal ICT infrastructure, policies and processes, and may necessitate wholesale institutional changes, both in government and the private sector. Conventional data management systems 'support security policies that are quite granular, protecting data at both the coarse and the fine grain level from inappropriate access', but in contrast, the software used for big data aggregation and analysis may lack such safeguards. Organisations that include sensitive data in their data analysis need to ensure that 'the same data security policies that apply to the data when it exists in databases or files are also enforced in the Big Data context' (Villars et al., 2011).

Opportunities, barriers, and 'determining factors'

The Australian Government Information Management Office's (AGIMO) *Big data strategy* (released in August 2013 (Department of Finance and Deregulation, 2013)) identifies the key *opportunities* of

'big' data to Australian Government agencies across Australia's security, cultural, democratic, social and economic systems as:

- data management as a national asset – financial savings from transparency and re-use of data (**economy**)
- personalisation of services – value in generating detailed understanding of customers and clients (**society, culture, economy**)
- problem-solving and predictive analytics – improved insights and decision-making capabilities (**society, culture, economy, security**):
    - 'It is expected that some big data projects will provide unanticipated insights into business problems,' the strategy reads. 'Industry experience suggests that these unanticipated correlations and discoveries may provide important insights ... [which] may also provide opportunities to act and respond more rapidly to information and trends as they occur.' (**economy**)
- productivity and efficiency – increased productivity and innovation from large-scale analysis (**economy**).

While data and technologies for data provide opportunities, the following key *concerns or barriers* also arise for government:

- data analytics and sensor infrastructure: these can be costly, but in and of themselves do not provide answers. If one fails to ask the right questions and make the right assumptions, data analytics will be a waste of time and resources (e.g. controversy over Google Flu data) (**economy, society, security**).
- lack of skills for data management, data mining, data collection, data analytics, statistical analysis (**economy, society, culture**)
- regulation: There is ambiguity about how 'big' data sits in traditional regulatory systems. This may in part be due to a category error (i.e. 'big data' is not a natural category). The debate about metadata collection by security agencies supports this. In an example drawn from the health sciences, genomics data is covered by double legislation. Human tissue is regulated by one set of systems and computational materials/information is regulated by another (**security, democracy, society, culture**).
- control, use and ownership: who gets to 'own' data, who buys/sells, who has the responsibility to store/secure it, what rights do people have to their own data (**security, democracy, culture**)?
- social and cultural context: i.e. genomic data on its own is useful but is more epidemiologically useful when combined with other health and demographic details. Such aggregation is accompanied by particular data management issues, including how to ensure anonymity of personal information. Some data is more 'personal' than others (**economy, security, democracy, society, culture**).
- data protection (**security, society, democracy**)
- responsibility:
    - EU's 'right to be forgotten' legislation (**democracy**)
    - data harvesters conflicting responsibilities to users, to shareholders, and also to state agencies (**security, democracy, society, economy**);
    - generational/shifting ideas of privacy/data – i.e. regulators may be non-users or late adopters of social media/technology. Anxieties/fears/concepts of shame are socially, culturally and historically produced, so aspirations of regulators might not match those of users (**culture, society, democracy**).

The opportunities and barriers identified by the AGIMO report are not unique to big data or to ICT more generally. Most of these problems arise with any technology.

Other 'determining factors': In Australia, broadband speeds and download capacity will significantly shape the possibilities of data harvesting and analysis. Before the 2013 election, under the ALP's preferred National Broadband Network (NBN) model, AGIMO anticipated that the NBN would 'assist in providing the necessary bandwidth to transport the data and may help to enable data to be analysed in a cloud environment and in near real-time'. With regard to the need for a skilled workforce, AGIMO noted a current 'shortage of university degrees that have a curriculum focused on big data analytics. Industry is looking for more qualified professionals with skills related to big data analytics. There is a push for education providers to develop courses that provide suitable education and training in this area of expertise' (Department of Finance and Deregulation, 2013). Government agencies also need to attract workers with diverse skill sets including training in 'science, technological, research, statistical, analytical and interpretive skills, business acumen and creativity … as well as an understanding of the underlying nature of the business process or policy intent', noting that 'these skill sets are unlikely to be found in any one person, and this means that collaborative teams of specialists will need to be assembled to allow agencies to achieve optimal results from their data analysis efforts. Many observers have noted that there is currently a major skills gap for data scientists with experience in big data analytics.'

There are opportunities for technology-driven innovation in government service delivery. Data analysis may help government agencies to increase productivity and effectiveness by enabling them to tailor and target services, policies and programs with more foresight. AGIMO anticipates that increasingly, data analysis will allow tailored services reflecting specific individual and community needs and interests and reduce over-servicing. The strategy also anticipates that better policy outcomes will be delivered through improved data management, by allowing government to access and perform analysis on more information from more sources: 'Decision makers may be able to model different policy options and more accurately predict the outcomes of policies before they are implemented and use this information to inform and improve the policy development process,' the strategy reads. 'Agencies could then use this granular information to make better informed and more responsive decisions, to achieve desired outcomes in a shorter amount of time, and at lower cost to the community.'

In particular, agencies need to improve the connections between cross-agency datasets, implement better practices for the use of third-party datasets and the de-identification of data, improve management of the mosaic effect (data that in isolation appear anonymous but can lead to a privacy breach when combined), support open data, and improve data retention and cross-border data flows. Innovations like data.gov.au and myGov are already improving interoperability and access to data and services, although they can raise fears of intrusion. Organisations like ANDS are working with governments and research organisations to help these agencies to implement better practices.

To complement the strategy, in April 2014 the Australian Government Department of Finance released a 'Better practice guide' for big data in the Australian Public Service, and in June 2014 the Australian Government released a draft guide to responsible data analytics, setting out the relevant existing administrative and ethical frameworks that support responsible data analytics.[12]

Open data

The AGIMO strategy observed that 'accessible information is the lifeblood of a robust democracy and a productive economy'. To ensure value, data must be 'discoverable, accessible and usable', whatever their level of aggregation, stability, and longevity (Department of Finance and

[12] Updated in January 2015, available at http://www.finance.gov.au/sites/default/files/APS-Better-Practice-Guide-for-Big-Data.pdf; http://www.finance.gov.au/sites/default/files/Responsible%20Data%20Analytics%20Draft.pdf.

Deregulation, 2013). Globally, there is a growing open access movement, which holds to the principle that certain sorts of data should be free to access, use, modify, and share.[13]

Open access can refer to access to articles published in academic journals (Nath, 2012):

- The open access movement seeks to make scholarly publications freely available online.
- Academic journals play a role in the dissemination of peer-reviewed information, in particular from publicly funded research. Journal publishers distribute and archive printed and digital editions of journals, and administer the peer review process.
- Articles are provided to journals (free of charge or at a cost to the author), and before being accepted for publication, they are inspected (free of charge) by other researchers in the field.
- Access to the often publicly funded information in articles published in academic journals is often via a paywall, which can be prohibitively expensive for researchers who are not affiliated to an organisation which subscribes to that journal.
- Furthermore, in the past few decades, journal subscription fees have, on average, increased faster than inflation, while university and library budgets have decreased. Journal publication, traditionally the domain of academic professional societies, is increasingly provided by for-profit publishers. Before the growth of the World Wide Web, which has presented many opportunities for dissemination of information, journals played an important role in gatekeeping and quality control, which may no longer be required, or even useful.

Open access can also refer to open access to scientific data. A drive for greater transparency and more effective exploitation of public funds has also led to demand for open access to publicly-funded research data, to support research and innovation.

This paper is primarily concerned with the second sense of open access, while noting that the two go hand-in-hand: increasingly, journals are requiring that data be made available to support research conclusions. For example, in 2014, online open publisher PLOS ONE revised its data policy to require authors to 'indicate where the data are housed, at the time of submission' in order that reviewers, editors and readers 'have that information transparently available when they read the article'. Facilitating open data publication, there are a growing number of subject area repositories for data 'such as GenBank for sequences, clinicaltrials.gov for clinical trials data, and Protein Data Bank (PDB) for structures' as well as 'unstructured repositories such as Dryad or FigShare where there is no appropriate subject-domain repository'.[14]

Open data is a movement whose hour appears to have come. At TED2009, Tim Berners-Lee called for 'raw data now' – 'government data, scientific data, community data, whatever it is' to be made available on the web, so it could 'be used by other people to do wonderful things, in ways that they never could have imagined' (Berners-Lee, 2010). Government and international organisations and universities are increasingly supportive of technologies which encourage access to publicly funded data. For example, the Research Data Alliance, founded in 2013 by the European Commission, the US National Science Foundation and National Institute of Standards and Technology, and the Australian Government's then Department of Innovation, is an international coalition dedicated to building 'the social and technical bridges that enable open sharing of data … across technologies,

[13] The Open Definition version 2.0 http://opendefinition.org/od/.

[14] PLOS blog 2014,'PLOS new data policy: Public access to data', 24 February, http://blogs.plos.org/everyone/2014/02/24/plos-new-data-policy-public-access-data-2/.

disciplines, and countries to address the grand challenges of society'.[15] The US Government now houses its sharable government data at data.gov; the UK's data.gov.uk is dedicated to 'opening up government' via its data; and the Australian Government's data.gov.au 'provides an easy way to find, access and reuse public datasets from the Australian Government'.[16] The Australian Research Council (ARC) is seeking input on how to optimise access to data generated through ARC-funded research, taking into account 'the wide variety of practices across disciplines for the generation, dissemination and storage of research data'(Australian Research Council, 2013), and other organisations are considering similar requirements or inducements. In Australia, the work of ANDS and the Research Data Alliance encourages and fosters open data protocols.

In 2011, the EU released an open data strategy, and in 2012 initiated the European Open Government Data Initiative, a free, open-source, cloud-based repository of software assets for government organisations based on four principles of open government: transparency, participation, collaboration, and job creation.[17] In 2012, the UK Cabinet Office published an influential white paper on open data, 'Unleashing the potential', identifying barriers and opportunities, and calling for the public sector to:

- enhance access to data to ensure there are no inequalities in the data market
- publish all data that can be published
- enhance vehicles for redress and introduce standards for higher data usability
- build greater trust in public data by safeguarding from misuse and protecting the public's right to privacy
- shift the culture of the public sector to improve data sharing where it is in the public interest and within legislative boundaries to ensure public services are more personalised and efficient.

Looking beyond government data, the white paper also seeks to encourage activity to make the data held by businesses and charities publicly available, release more open data and support open data users, identify legislative opportunities to enable the public sector to make better use of public data, and seek opportunities to facilitate greater openness across society as a whole. Consistent with these international trends, on 13 June 2013 the G8 released its Science Ministers Statement supporting collaboration and agreement on open scientific research data, and increasing access to the peer-reviewed, published results of scientific data.[18]

Opportunities, barriers, and 'determining factors'

Opening access to publicly funded data where appropriate will provide *opportunities* for businesses, industry groups, researchers and citizens, as well as for governments themselves. The EU describes data as 'an essential raw material for a wide range of new information products and services which build on new possibilities to analyse and visualise data from different sources. Facilitating reuse of this raw data will create jobs and thus stimulate growth' (**economy, society**). Opening access to publicly funded data may provide new business opportunities (**economy**), for example, new products and services. It can increase transparency in government and business (**democracy, society, security**). It can also provide public sector workers with data to improve administration and policy across government, improving public services and targeting public spending (**economy,**

[15] Research Data Alliance https://rd-alliance.org/about.html

[16] https://www.data.gov/; http://data.gov.uk/; http://data.gov.au/

[17] http://europa.eu/rapid/press-release_MEMO-11-891_en.htm.

[18] https://www.gov.uk/government/publications/g8-science-ministers-statement-london-12-june-2013.

democracy, society). Opening access to data allows reuse as well, which is cost-effective and may encourage innovation (**economy**).

Reducing *barriers* to access means more than removing formal restrictions on collection and use. Access, reuse, and sharing of publicly funded data by the public may also be limited by privacy concerns, about both personal information and about proprietary data (**security, society, culture, democracy**). In order to access information, one needs to know it exists, so cataloguing and managing data is critically important to a functioning open data system. Uncertainty about who might own the data, and the terms under which it can be shared and used, can also be an impediment to reuse. Lack of standardisation of metadata, and inconsistent, expensive, or hard-to-use storage formats, complicated licensing, and cost are also barriers to use and reuse of public data (**economy, security**).

'Determining factors': For effective use and re-use, open data needs technologies, infrastructure and processes which promote standards and metadata, interoperability, sharing, storage, and where appropriate, de-identification. But even where technical solutions exist, there may be cultural and regulatory barriers. Regulators, users and curators of data collections, and designers and administrators of data infrastructure need to find the right balance between preservation, storage, use on the one hand and managing digital traces, personal information harvested from online behaviour and sensors on the other (**society, culture, economy, security**).

Technologies for research data

'Researchers often need more than their own data and more than their own computer … the efficient storage of large sets of data to allow future access is a whole specialist enterprise in itself' (Wilkinson et al., 2012)

The Australian Research Data Infrastructure Strategy, released by the Australian Government in November 2014, describes infrastructure for research data as 'a range of facilities, equipment or tools that serve research through data generation, manipulation, [curation,] and access. Research data infrastructure includes data itself, and relies on a skilled technical and research workforce for establishment, implementation, operation and use. It includes data itself' (Research Data Infrastructure Committee, 2014). Aspiring to a holistic Australian research data infrastructure system that '*collects* data systematically and intentionally; *organises* data to make it more valuable; and *uses* data insightfully many times over', the strategy proposes three key requirements to support Australian research data:

- sustained infrastructure to support priority research data collections, data generation and management
- appropriate data governance and access arrangements, and
- effective data infrastructure arrangements to support delivery of enhanced research outcomes.

While there are a number of funding mechanisms to support research infrastructure, large, national, collaborative research infrastructure investments in Australia have for the last decade been made through the National Collaborative Research Infrastructure Strategy (NCRIS). NCRIS represents an innovative approach to funding large-scale research infrastructure. Developed over several years, the key to the model is the recognition that medium to large-scale research infrastructure have national implications and are best supported through a strategic, collaborative processes. The strategy focuses on:

- supporting priority research areas and aligning with national research priorities
- facilitating access for all Australian researchers and internationally
- encouraging collaboration through collaborative distributed infrastructure rather than funding individual disciplines

- understanding lifecycle costs of infrastructure
- cross-sector co-investment
- strategic development of Australian science and research capacity
- cost-effective and efficient use of infrastructure.

Consultation through the 2008 and 2011 *Strategic roadmaps for Australian research infrastructure* confirmed the value of the strategic collaborative approach to funding large-scale research infrastructure (Research Data Infrastructure Committee, 2014).

Opportunities, barriers, and 'determining factors'

'International agencies increasingly recognise that data, being a pervasive and potentially long-lived information asset for all of society, needs planning and coordination' (Research Data Infrastructure Committee, 2014).

Publicly funded data is a valuable national asset. Large-scale collaborative research infrastructure is an enabling investment that underpins the delivery of the multidisciplinary research required to address both national and global challenges, now and into the future (**economy, democracy, society, culture, security**). The Australian Research Data Infrastructure Strategy identifies the following opportunities for technologies for research data (Research Data Infrastructure Committee, 2014):

- maximising the availability and delivery of, and the connections between, data generated and collected by research data infrastructure to researchers and others
- identifying pathways by which data generated from outside the research sector, including within government, can also be made available to researchers more effectively
- making publicly funded research data available to a wider audience, such as government users, non-government users and the private sector
- investigating how data is exchanged and used in an environment where the researcher relies not only on data created and collected in his or her sector, discipline or institution, but also on data produced elsewhere, by other disciplines or in government and private sectors.

The Australian Government has made a number of research data infrastructure investments through NCRIS and the Super Science Initiative. Among other infrastructure investments, ANDS, through its development of an Australian Research Data Commons and its work with the Research Data Alliance, 'is transforming research data from unmanaged, disconnected, invisible and single-use to managed, connected, findable and reusable'. Two national high-performance computing centres – the National Computational Infrastructure (NCI) at the Australian National University and the Pawsey High Performance Computing Centre led by the joint venture iVEC in Perth – 'will provide petascale capacity for data analysis and modelling into the next five years' in fields including meteorology, Earth science research and astronomy. Investment in the Australian Research and Education Network (AREN) 'is improving research data transfer and collaboration across the nation and internationally, including for the large data sets produced from disciplines such as radio astronomy and environmental science' (Research Data Infrastructure Committee, 2014).

The generation, collection, curation, use and re-use of large and complex datasets requires large-scale technologies for data such as those described above, but sustaining their funding is often an area of market failure, though the benefits to industry and to citizens are widely recognised (Research Data Infrastructure Committee, 2014) (**economy**). The strategy suggests that a solution lies in a national research data infrastructure system 'that allows integration throughout the data lifecycle, from processing to collection to curation and storage to reuse', one which encourages discoverability, and promotes 'open and flexible access arrangements … while allowing funders, operators and users of research data infrastructure to capitalise on future transformative technologies'. This is necessary to optimise the use of Australia's research data, as 'policy-makers, as well as funders, designers, operators and users of research data infrastructure, will need new

approaches and solutions which take account of changing technologies and environments, including current and future national and international drivers'.

'Determining factors': Large and complex research data programs need technologies and processes for:

- collection and generation – supercomputers, sensor networks, modelling and simulation
- storage – disks, tape, data centres
- management and curation
- analysis – computers, virtual laboratories, modelling and simulation
- fulfilling interoperability and metadata requirements
- access to government data and records – archives, publicly funded data, public collections
- making sense of big, complex datasets, whether they are derived from online behaviour, the public health system, computer modelling of climate cycles, or astronomy (imaging data presents the biggest challenge).

Box 4: Things to do to data

capture / collect (e.g. sensor network, clinical trial, survey)

organise (e.g. database)

analyse (e.g. spreadsheet, algorithm)

manipulate (e.g. model & simulation)

compute (e.g. processor)

generate (e.g. simulation)

manage (e.g. metadata records)

re-use / repurpose (e.g. open data technologies, standards)

share / disseminate (e.g. internet)

curate (e.g. digitisation and database entry)

store (e.g. data centre, cloud)

Technologies for data need planning and coordination. Researchers in this environment with an overabundance of electronic data and access to large and important data collections, or responsibility for them, benefit from better integrated infrastructure. To support this data-intensive research and optimise the outcomes for researchers in all fields, funders and infrastructure designers and operators need to provide better ways to generate, organise, manipulate, share, use and re-use data throughout the data lifecycle (see also Box 4). Policymakers, as well as funders, designers, operators and users of research data infrastructure, will need new approaches and solutions which take account of changing technologies and environments, including current and future national and international drivers.

Policy issues and implications

Granting that *Homo sapiens* has arguably always (and by definition) been in an information age, is it possible to trace a shift in kind as well as volume over the last half century? i.e., has the rapid increase in the amount and accessibility of electronic data effected a shift in the notion of what it means to research or to understand something? Proponents of data as the 'fourth paradigm of scientific knowledge' would say yes (Gray, 2009) (Hey et al., 2009); (Hey, 2010); (Research Data Infrastructure Committee, 2014)). If so, what are the most important policy issues emerging from the electronic 'data deluge' and the 'age of interoperability'? These policy issues will probably relate to concerns about privacy, security, cybercrime, and surveillance, as well as information overload,

statistical literacy, models for funding of national research infrastructure, a workforce that is potentially underprepared for large-scale data management and analysis, and the wider social implications of electronic personal data and ubiquitous sensors.

AGIMO's Big Data strategy is underpinned by six 'big data principles': the strategy aims to position Australia as a leader in the public sector use of data analytics to reform service delivery, improve public policy and protect citizens' privacy. While somewhat light on the details of the benefits of data from and for research, and the infrastructure that supports them, and with little discussion of open data, the identified key benefits to Australian Government agencies apply equally to the three data systems discussed ('big', open, and research):

- data management is a national asset – financial savings from transparency and reuse of data
- personalisation of services – value in generating detailed understanding of customers
- problem solving and predictive analytics – improved insights and decision-making
- increased productivity, efficiency and innovation from large scale analysis.

Big data, open data technologies and collaborative research data infrastructure share many features, impediments and offer similar opportunities, and attempts to maintain strict boundaries will often serve to highlight shared features (Yarkoni, 2014). Again, many of these features, impediments and opportunities are common to other technologies:

- Regulation, which can protect consumers, also may hamper outcomes that may produce public good.
- Cost is a common factor for decisions made about: infrastructure, equipment, training, entire data lifecycle (includes storage, curation, management, and disposal).
- Entire systems are increasingly data-dependent. Knowledge infrastructure, including sensor networks and technologies for interoperability, facilitates and fuels this dependence.
- Economic benefits of data access can accrue at individual, social, firm, and industrial levels.
- Cultures of design and use, including those around metadata, interoperability, and standards, are all important for users and designers of big, open, and collaborative research data infrastructure.
- Concerns about privacy, security and surveillance, data protection ('the right to be forgotten'), data responsibility (of data harvesters, storers, users; but also of data 'owners' or subjects), data accountability commonly affect decisions about data.
- Related to data ownership and privacy, exclusion is a factor in the distribution of benefits and costs from data technologies.
- Knowledge infrastructures and data are invaluable tools for policy, social research, public health in the right hands, with the right questions, but data are not neutral and nor are the technologies for understanding and managing them.
- Cross-sector improvement in skills for researchers and data infrastructure specialists is needed:
    - infrastructure designers and operators
    - statistics – for analysing, manipulating, and understanding data
    - data storage, management, curation, access etc.

There are other ways of grouping these three data systems (big, open, and research):

- data as currency (data markets, data mining):
    - regulatory problems when companies use personal data to trade, but also opportunities to improve people's customer experiences through targeted advertising
    - trade-off: free online services provided in exchange for data about users. This trade-off is often implicit/invisible, and data providers/service users may not realise its

implications, becoming angry or suspicious when the extent of data collection becomes clear to them
    - the rise of technologies like bitcoin and other cryptocurrencies (literally *data as currency*);
- data for government and social policy:
    - as with data as currency, people are concerned about intrusion by government, in particular security agencies, that arises from the collection and use of 'personal' data by government agencies
    - these data have more value when their infrastructure is interoperable and they are available, where appropriate, to researchers
- data as research tool and output:
    - interoperability and availability for re-use are critical to provide research data and data for research effectively as a collective resource
    - data generated by research can be a valuable tool for government policymakers.

Taking these aspects and recent investments into account, Australia could position itself to lead in a data economy: 'Imagine having the best analytic tools processing the most complete data sets running on one of the fastest computers in the world in collaboration with colleagues connected by high-speed networks into a virtual laboratory. This is not a dream for an increasing number of Australian researchers. The … facilities that comprise the nation's investment in research infrastructure … [are] already delivering it' (Wilkinson et al., 2012).

The recent history of data collection, generation and use illuminates some recurring inquiries in the SAF05 project, which have been addressed in some other case studies, including:

- the relationship between science, engineering, and technology
    - data are mediated by information and communication technologies; there is an arms race between data security and data surveillance (supporting Project Question 5)
    - 'Big data' is made from small data, but a sufficiently 'big' dataset is qualitatively different from small collections.
    - For scientific researchers, there is a mutualism between developments in high-speed computing, sensor networks, digitisation, etc, and the need for better technologies and processes for data storage, analysis, management, sharing etc. (Douglas, 2015)
- the extent to which technological innovation creates or responds to demand
    - the example of Google: its founders had a great search engine, but how to capitalise on the fact that people use it, and how to make it financially viable and self-sustaining? What Google has is data about people and their habits, and a means to collect more. Google, Amazon, Facebook, eBay and their ilk are part of a technology Zeitgeist that helped create and feed a market for data about consumers.
    - Personalisation of services – there is value for consumers and for providers in generating detailed understanding of customers/users and improving customer/user experience.
    - Problem solving and predictive analytics – improved insights and decision-making for business, government, and research.
- the role of economics in innovation
    - data as a national asset – economic benefits accrue from increased transparency and re-use of data
    - productivity and efficiency – increased productivity and innovation from large-scale analysis of data

    - These factors emerge from sustainable holistic collaborative research data infrastructure, pervasive interoperability, and standards.
- and the roles of government and industry in fostering and directing technological innovation and development
    - For large and complex data sets of national significance, government funding may be required to optimise the use of these data, ensure they are available for future research, and to encourage interoperability and cross-disciplinary uses.
    - Privacy, security, and data responsibility: In many fields, particularly where data is commercially, financially, or personally sensitive (e.g. patents, credit cards, health data), data collection and storage are already regulated. Since the EU ruled that people have the 'right to be forgotten', business data analysts and proponents of the commercial use of big data have become increasingly concerned about the imposition of tighter regulatory standards for data collection, storage, and use.
    - Standards and metadata: formal versus de facto standards – it is difficult to force formal technical standards. Standards apply through stages of development, implementation, and acceptance. Simplicity is most likely to be accepted. The http standard works because it only has four elements to it (get post put delete). The protocol html is more complicated but tractable, but anything more complicated might be hard to enforce or encourage. There may be trade-offs in composability, efficiency and interoperability.
    - Data lifecycle – funding issues. Large-scale collaborative data infrastructure may be an example of market failure, and an area for government intervention. For optimal value, data needs to be managed throughout its lifecycle (Research Data Infrastructure Committee, 2014).

Summary

While information about the world is nothing new, developments in information and communications technologies over the last 50 years are revolutionising research and industry. These sectors benefit from the transformative potential of high-speed networks, computational power, ubiquitous sensor networks, smart tools, and the increased promotion and uptake of cloud computing technologies.

Human behaviour can now be traced via digital interactions in unprecedented detail. As a result of the pervasive interoperability, regulators, users and curators of data collections and designers and administrators of data infrastructure may find it difficult to strike the right balance between preservation, storage and use on the one hand, and managing digital traces, personal information harvested from online behaviour and sensors on the other.

Data has many uses – as currency or commodity, as an instrument for and output of research, as a tool for policy, control, democracy, equity, and surveillance, and as a resource for citizens and consumers. Data that is collected and generated for some purposes is being co-opted for others (for example, the collection of telephone metadata by state surveillance agencies, or the use of GPS data from mobile telephones to track phone user habits and target advertising to them), and there is an unequal power relation between those who are the source of data and those who capture, collect and use it. These contested uses and outcomes produce tensions that are part of a democratic deficit, whereby democratic institutions fall short of fulfilling the principles of democracy in their practices or operation. For instance, dearly-held commitments to checks on the power of the state may be in conflict with recent legislation allowing the use by Australian law enforcement of 'metadata' collected by internet service providers, together with compelling ISPs to retain this data. And data producers (i.e. citizens) may have trouble accessing data about themselves, while non-state entities use this data for commercial or criminal purposes, and state agencies may themselves be complicit in the corruption of infrastructure.

ICT is a general purpose technology, and data are its general-purpose constituents. The impact of technologies for data stems from the interactions of data with other technology and with other data, ensuring a multiplicative effect: as 'data intermediaries', technologies for data themselves facilitate interoperability. However, merely having more data does not translate automatically to having better information.

'Big' data analysis is subject to the same limits as any statistical analysis. Data collections need to be managed. Technologies to store, manage, access, analyse, and share data can optimise the use and ensure the preservation of large and complex datasets in the national interest, as well as providing Australian researchers and industries with a competitive advantage. But if Australians are to capitalise on this 'age of interoperability', policymakers, researchers, small and medium enterprises, business and industry, legislators, and citizens need a better understanding of the challenges, opportunities and limits of data and data technologies.

The significant increase in the rate of data being created and captured, the range of disciplines and industries depending on data, and the substantial potential benefits offered by integrated data collection, generation, analysis, manipulation and re-use require coordination across research infrastructure initiatives and between stakeholders. This includes coordination within the research sector, and also with governments, non-government organisations, the private sector and the community.

References

Alexandrow, Catherine (2008). The story of GPS. *In:* Defense Advanced Research Projects Agency (DARPA) (ed.) *DARPA, 50 years of bridging the gap.* Arlington VA: Defense Advanced Research Projects Agency.

Anderson, Chris (2008). The end of theory: the data deluge makes the scientific method obsolete. *Wired Magazine* [Online], 16. Available: http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory.

Australian Research Council (2013). Australian Research Council annual report 2012–13. Canberra, Australian Research Council, Available from:http://www.arc.gov.au/pdf/annual%20report%2012-13/Annual%20Report%20PDF%20for%20web_Part1.pdf

http://www.arc.gov.au/pdf/annual%20report%2012-13/Annual%20Report%20PDF%20for%20web_Part2.pdf

Babbage, Charles (1832). *On the economy of machinery and manufactures,* London, Charles Knight.

Berners-Lee, Tim (2010). The year open data went worldwide. *TED* [Online]. Available: http://www.ted.com/talks/tim_berners_lee_the_year_open_data_went_worldwide.

Borges, Jorge Luis (1964). The Library of Babel. *In:* Yates, Donald A & Irby, James E (eds.) *Labyrinths: selected stories and other writings.* New York: New Directions.

Brin, Sergey & Page, Lawrence (1998). The anatomy of a large-scale hypertextual web search engine. *Seventh International World-Wide Web Conference (WWW 1998)*, 14-18 April 1998 1998 Brisbane.

Burdon, Mark & Andrejevic, Mark (2014). Detection devices: how a 'sensor society' quietly takes over *The Conversation* [Online]. Available: http://theconversation.com/detection-devices-how-a-sensor-society-quietly-takes-over-26089.

Bush, Vannevar. (1945). As we may think. *The Atlantic*, 1 July 1945, Available from: http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/.

CERN. (2008). Tim Berners-Lee's proposal [Online]. Available: http://info.cern.ch/Proposal.html.

Ceruzzi, Paul E (1986). An unforeseen revolution: computers and expectations, 1935–1985. *In:* Corn, Joseph J (ed.) *Imagining tomorrow: history, technology, and the American future.* Cambridge MA: MIT Press.

Cox, Michael & Ellsworth, David (1997). Application-controlled demand paging for out-of-core visualization. Moffett Field CA, NASA Ames Research Center, Available from:https://www.nas.nasa.gov/assets/pdf/techreports/1997/nas-97-010.pdf

Darnton, Robert (2014). The good way to do history: book review of 'The Allure of the Archives' by Arlette Farge. *New York Review of Books* [Online]. Available: http://www.nybooks.com/articles/archives/2014/jan/09/good-way-history/.

Department of Finance and Deregulation (2013). The Australian public service: big data strategy. Canberra, Department of Finance and Deregulation, Australian Government Information Management Office, Available from:http://www.finance.gov.au/sites/default/files/Big-Data-Strategy_0.pdf

Douglas, Kirsty (2007). *Under such sunny skies: understanding weather in colonial Australia, 1860–1901,* Melbourne, Bureau of Meteorology.

Douglas, Kirsty (2015). Digital computing, modelling and simulation. *Working paper of SAF05 project.* Melbourne: Australian Council of Learned Academies.

Fung, K (2010). *Numbers rule your world: the hidden influence of probabilities and statistics on everything you do,* New York, McGraw-Hill.

Gibbs, Samuel (2014). Why ditching Facebook feels like opting out of modern life. *The Guardian* [Online]. Available: http://www.theguardian.com/technology/2014/feb/01/why-ditching-facebook-feels-like-opting-out-of-modern-life.

Gleick, James (2011). The information: a history, a theory, a flood.

Gray, Jim (2009). Jim Gray on eScience: a transformed scientific method. *In:* Hey, Tony, Tansley, Stewart & Tolle, Kristin (eds.) *The fourth paradigm: data-intensive scientific discovery.* Redmond WA: Microsoft Research.

Griffiths, Emma (2014). Data retention laws: Tony Abbott says Government 'seeking metadata', not targeting people's browsing history. *ABC News* [Online]. Available: http://www.abc.net.au/news/2014-08-06/security-laws-abbott-browsing-history-not-collected/5652364.

Grover, Lovleen Kumar & Mehra, Rajni (2008). The lure of statistics in data mining. *Journal of Statistics Education,* 16**,** 1**,** 1–8.

Hand, David J (2007). *Information generation: how data rule our world,* Oxford, OneWorld Publications.

Harford, Tim (2014). Big data: are we making a big mistake? *FT Magazine* [Online]. Available: http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz3DIGcMzsq.

Hey, Tony (2010). The big idea: the next scientific revolution. *Harvard Business Review***,** November**,** 1-8.

Hey, Tony, Tansley, Stuart & Tolle, Kristin (eds.) (2009). *The fourth paradigm: data-intensive scientific discovery,* Redmond WA: Microsoft Research.

Hilbert, Martin & López, Priscila (2011). The world's technological capacity to store, communicate, and compute information. *Science* [Online], 332.

Kessler, Sarah (2013). Think you can live offline without being tracked? Here's what it takes. Available: http://www.fastcompany.com/3019847/think-you-can-live-offline-without-being-tracked-heres-what-it-takes.

KNC (2014). The backlash against big data. *The Economist* [Online]. Available: http://www.economist.com/blogs/economist-explains/2014/04/economist-explains-10.

Korte, Travis. 2014). Data Innovation 101. *Ideas Lab* [Online]. Available from: http://www.ideaslaboratory.com/post/93343636988/data-innovation-101 [Accessed 16 January 2014.

Lazer, David, Kennedy, Ryan, King, Gary & Vespignani, Alessandro (2014). The parable of Google flu: traps in big data analysis. *Science,* 343**,** 6176**,** 1203–1205.

Leek, Jeff. 2014). Why big data is in trouble: they forgot about applied statistics. *Simplystats* [Online]. Available from: http://simplystatistics.org/2014/05/07/why-big-data-is-in-trouble-they-forgot-about-applied-statistics/ [Accessed 7 May 2014.

Lipsey, Richard G, Carlaw, Kenneth I & Bekar, Cliffford T (2005). *Economic transformations: general purpose technologies and long-term economic growth,* Oxford, Oxford University Press.

MacLeod, Roy (1994). Embryology and empire: the Balfour Students and the quest for intermediate forms in the laboratory of the Pacific. *In:* MacLeod, Roy & Rehbock, PF (eds.) *Darwin's laboratory: evolutionary theory and natural history in the Pacific.* Honolulu: University of Hawai'i Press.

Marcus, Gary & Davis, Ernest (2014). Eight (no, nine!) problems with big data. *The New York Times* [Online]. Available: http://www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html?_r=0.

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

Moore, Gordon E (1965). Cramming more components onto integrated circuits. *Electronics,* 38**,** 8**,** 114–117.

Morris, RJT & Truskowski, BJ (2003). The evolution of storage systems. *IBM Systems Journal,* 42**,** 2**,** 205–217

Nath, Chandrika (2012). Open access to scientific information. London, Parliamentary Office of Science and Technology, Available from

Oxford English Dictionary (OED) (2013). History of the OED. Available: http://public.oed.com/history-of-the-oed/.

Press, G. 2012). A very short history of big data. *What's the big data?* [Online]. Available from: http://whatsthebigdata.com/2012/06/06/a-very-short-history-of-big-data/ [Accessed 6 June 2012.

Research Data Infrastructure Committee (2014). The Australian research data infrastructure strategy. Canberra, Department of Industry, Innovation, Science, Research and Tertiary Education, Available from:http://docs.education.gov.au/system/files/doc/other/the_australian_research_data_infrastructure_strategy.pdf

Rheingold, Howard (2000). *Tools for thought: the history and future of mind-expanding technology* Cambridge MA, MIT Press.

Scott, James C (1998). *Seeing like a state: how certain schemes to improve the human condition have failed,* New Haven, Yale University Press.

Sgarbi, M (2013). *The Aristotelian tradition and the rise of British empiricism: logic and epistemology in the British Isles (1570-1689),* New York, Springer.

Shannon, CE (1948). A mathematical theory of communication

Bell System Technical Journal, 27**,** July

October**,** 379–423

623–656.

Standage, Tom (1998). *The Victorian internet: the remarkable story of the telegraph and the nineteenth century's on-line pioneers,* New York, Walker and Company.

Strawn, George O. (2012). Scientific research: how many paradigms? *Educause Review***,** 26-34.

Thwaites, Tim (2013). Data revolution. *Resourceful: Bringing CSIRO research to the minerals industry,* 4**,** 3.

Tilly, Charles (1980). The old new social history and the new old social history. *Center for Research on Social Organization Working Papers***,** 218.

Turnbull, M (2014). AIIA Speech: Navigating Analytics Summit. *Navigating Analytics Summit* 20 March 2014 2014 Canberra. Australian Information Industry Association.

Vastag, Brian (2011). Exabytes: documenting the 'digital age' and huge growth in computing capacity. *Washington Post* [Online]. Available: http://www.washingtonpost.com/wp-dyn/content/article/2011/02/10/AR2011021004916_pf.html.

Veletsianos, G. 2014). On Noam Chomsky and technology's neutrality. *George Veletsianos* [Online]. Available from: http://www.veletsianos.com/2014/01/23/on-noam-chomsky-and-technologys-neutrality/ [Accessed 23 Jan 2014.

Villars, Richard L, Eastwood, Matthew & Olofson, Carl W (2011). Big data: What it is and why you should care. Available: http://sites.amd.com/us/Documents/IDC_AMD_Big_Data_Whitepaper.pdf.

Walton, CA. (1983). Portable radio frequency emitting identifier. *USA patent application*.

Wells, HG (1938). *World brain,* London, Methuen and Co.

Wigan, Marcus R & Clarke, Roger (2013). Big data's big unintended consequences. *Computer,* 46**,** 6**,** 46-53.

Wilkinson, Ross & Thwaites, Tim (2012). Piecing together the eResearch puzzle. *Share: the newsletter of the Australian National Data Service***,** 14**,** 1–2.

Winsberg, Eric (2010). *Science in the age of computer simulation,* Chicago, University of Chicago Press.

Wright, Alex (2007). *Glut: mastering information through the ages* Washington DC, National Academies Press.

Yarkoni, Tal. 2014). Big data, n. A kind of black magic. Available from: http://www.talyarkoni.org/blog/2014/05/19/big-data-n-a-kind-of-black-magic/ [Accessed 19 May 2014.