OECD publishing

Please cite this paper as:

OECD (2013-06-18), “Exploring Data-Driven Innovation as a New Source of Growth: Mapping the Policy Issues Raised by "Big Data"”, *OECD Digital Economy Papers*, No. 222, OECD Publishing, Paris.
http://dx.doi.org/10.1787/5k47zw3fcp43-en

OECD Digital Economy Papers No. 222

Exploring Data-Driven Innovation as a New Source of Growth

MAPPING THE POLICY ISSUES RAISED BY "BIG DATA"

OECD

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development

18-Jun-2013

English - Or. English

DIRECTORATE FOR SCIENCE, TECHNOLOGY AND INDUSTRY
COMMITTEE FOR INFORMATION, COMPUTER AND COMMUNICATIONS POLICY

Cancels & replaces the same document of 18 April 2013

EXPLORING DATA-DRIVEN INNOVATION AS A NEW SOURCE OF GROWTH

Mapping the Policy Issues Raised by "Big Data"

JT03342004

Complete document available on OLIS in its original format
This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

DSTI/ICCP(2012)9/FINAL Unclassified
English - Or. English

FOREWORD

This report explores the potential role of data and data analytics for the creation of significant competitive advantage and for the formation of knowledge-based capital (KBC), which can drive innovation and sustainable growth across the economy and society.

The report contributes to phase one of the OECD horizontal project "New Sources of Growth: Intangible Assets", which was coordinated under the auspices of the OECD Committee on Industry, Innovation and Entrepreneurship (CIIE). The policy issues mentioned in the report will be developed further during phase two of the project to be conducted in 2013-14 under the auspice of the OECD Committee for Information, Computer and Communications Policy (ICCP).

This report was first presented to the ICCP in October 2012 and declassified by the ICCP in February 2013. It takes into account the outcome of the 2012 ICCP Technology Foresight Forum on "Harnessing data as a new source of growth: Big data analytics and policies" held on 22 October 2012 at the OECD Headquarter in Paris, France (http://oe.cd/tff2012).

The report was prespared by Mr. Christian Reimsbach-Kounatze with contributions from Mr. Brendan Van Alsenoy, both of the OECD Directorate for Science, Technology and Industry (STI). It is published under the responsibility of the Secretary-General of the OECD.

© OECD 2013.

TABLE OF CONTENTS

FOREWORD 2

SUMMARY 4

EXPLORING DATA-DRIVEN INNOVATION AS A NEW SOURCE OF GROWTH: MAPPING THE POLICY ISSUES RAISED BY "BIG DATA" 7

Introduction 7
Understanding data and the drivers of their generation and use 8
Data generation, collection and transport 8
Data storage and processing 9
Defining "big data": volume, velocity and variety, but also value 11
The increasing use and value of data across the economy 12
Online advertisement 15
Governments and public-sector agencies 15
Health care 18
Utilities 18
Logistics and transport 19
Mapping the policy opportunities and challenges 21
Privacy and consumer protection 22
Open access to data 24
Cybersecurity risks 26
Skills and employment 27
Infrastructure 28
Measurement 29
Conclusion 29

NOTES 30

REFERENCES 35

Boxes

Box 1. Data-driven science and research 14
Box 2. Data proliferation and implications for official statistics 17
Box 3. OECD Technology Foresight Forum 2012: Harnessing data as a new source of growth - Big data analytics and policies 21
Box 4. Basic principles of national application of the OECD (1980) Privacy Guidelines (part 2) 23
Box 5. Principles of the OECD (2008) Recommendation for Enhanced Access and More Effective Use of Public Sector Information 25
Box 6. Principle of the OECD Guidelines for the Security of Information Systems and Networks 27
Box 7. Transmitting data – a regulatory barrier to machine-to-machine communication 28

SUMMARY

The confluence of several technological and socioeconomic trends, including the increasing migration of social and economic activities to the Internet and the decline in the cost of data collection, transport, storage and analytics, are leading to the generation of a huge volume of data – commonly referred to as *big data* – that can be exploited to foster new industries, processes and products. Economic and social activities have long relied on data. Today, however, the increased volume, velocity and variety of data used across the economy, and more importantly their greater social and economic value, signal a shift towards a data-driven socioeconomic model. In this model, data are a core asset that can create a significant competitive advantage and drive innovation, sustainable growth and development.

In business, the exploitation of data promises to create added value in a variety of operations, ranging from optimising the value chain and manufacturing production to more efficient use of labour and better customer relationships. Even traditional sectors such as retail are changing: firms like Tesco, the UK supermarket chain, exploit huge data flows generated through their fidelity card programmes. The Tesco programme now counts more than 100 market baskets a second and 6 million transactions a day, and it very effectively transformed Tesco from a local, downmarket "pile 'em high, sell 'em cheap" retailer to a multinational, customer-oriented one with broad appeal across social groups.

Among the sectors using data, five are discussed here as areas in which the use of data can stimulate innovation and productivity growth. They include online advertisement, health care, utilities, logistics and transport, and public administration. Together these sectors accounted for roughly one-quarter, on average, of total value added in OECD countries in 2010. Overall, the benefits that the exploitation of data point to in these sectors include:

- Enhancing research and development (data-driven R&D);
- Developing new products (goods and services) by using data either as a product (data products) or as a major component of a product (data-intensive products);
- Optimising production or delivery processes (data-driven processes);
- Improving marketing by providing targeted advertisements and personalised recommendations (data-driven marketing);
- Developing new organisational and management approaches or significantly improving existing practices (data-driven organisation).

In **the online advertising sector,** click-stream data are increasingly collected to track the browsing habits of consumers. For individual firms, the exploitation of click-stream data provides new means of improving customer relationship management (CRM). It allows businesses to allocate their marketing budgets better and to target the marketing channels that reach the most valuable customers. Over the last five years the revenue generated by online advertising has grown much faster than revenue from traditional advertising channels in their first 15 years. In the first quarter of 2012, online advertising revenue of the

top 500 advertisers in the United States reached USD 8.4 billion. This is USD 1.1 billion (15%) more than in the first quarter of 2011.

The health-care sector has long wished to create unified electronic health records (EHRs). EHRs offer many advantages over paper records: reduced record management costs; reduced medical errors and improved care, diagnosis and treatments; the potential for greater use of evidence-based care; easier choice of doctor and care facilities by patients; and possible linkages to medical research and insurance. It is estimated that big data could be used throughout the health care system – from clinical operations to payment and pricing of services and R&D – with total savings of more than USD 300 billion for US health care by 2020. These estimates do not include benefits from developing timely public-health policies using real-time data, *e.g.* from web searches, to assess epidemiological trends.

In **the utilities sector,** the adoption of "smart-grid" technologies to reduce or better manage electricity consumption is leading to large volumes of data on energy and resource consumption patterns. "Smart meters", for instance, enable not only real-time collection of consumption data but also the exchange of real-time price data. Furthermore, they can send signals controlling the turning on or shutting off of various household appliances connected to the grid. While the information feedback allows consumers to adjust their energy and resource consumption to current production capacities, utilities can run data analytics to identify overall consumption patterns in order to forecast future demand and to adjust production capacities and pricing mechanisms to this demand. Overall, the use of data-driven smart grid applications could reduce CO_2 emissions, equivalent to EUR 79 billion, by more than 2 gigatonnes (billion tonnes) by 2020.

The transport sector's increasing ability to track the location of mobile devices has enabled both the monitoring of traffic to save time and reduce congestion as well as the provision of new location-based services. For example, in 2012 TomTom, a leading provider of navigation hardware and software, had more than 5 000 trillion data points in its databases, gleaned from its navigation devices and other sources, describing time, location, direction and speed of travel of individual anonymised users. TomTom adds five billion measurement points every day. Overall, estimates suggest that the global pool of personal geo-location data represented at least one petabyte in 2009, with growth of about 20% a year. By 2020, this data pool could provide USD 500 billion in value worldwide in the form of time and fuel savings, or 380 megatonnes (million tonnes) of CO_2 emissions saved. These figures do not include value provided by other location-based services.

The use of data is not limited to the private sector. **The public sector** is also an important data user and a source of data that can generate benefits across the economy. Some evidence shows that by fully exploiting public sector data, governments could reduce their administrative costs. For Europe's 23 largest governments, some estimate potential savings of 15% to 20%. This is the equivalent of EUR 150 billion to EUR 300 billion in new value. These estimates do not include the additional benefits that would arise from greater access to and more effective use of public-sector information (PSI), as called for by the OECD's 2008 Council Recommendation, currently under review.[1] Such benefits can be obtained from weather forecasts, traffic management, crime statistics, improved transparency of government functions (e.g. procurement) and educational and cultural knowledge for the wider population. Estimates suggest that the European market value related to PSI was around EUR 32 billion in 2010.

Policy implications

To unlock the potential of big data, OECD countries need to develop coherent policies and practices for the collection, transport, storage, provision and use of data. These policies cover issues such as privacy protection, open data access, skills and employment, infrastructure, and measurement, among others.

Privacy protection – ensuring trust and innovation in the Internet economy.[2] New data sources, new actors and the increasing ease of linking and processing data raise questions for privacy protection frameworks. It becomes necessary to consider today's broader uses of personal data with a view to more effective protection of privacy and the realisation of the economic and social benefits of trustworthy and innovative uses of personal data.[3] As cross-border flows of data are now critical to national and global economic and social development, privacy protection regimes should support open, secure, reliable and efficient data flows, while lessening privacy risks and enhancing responsible behaviour in the use of personal data.

Open access to data – leading by example. The linking and use of data across sectors drive innovation, socioeconomic development and growth. An example is the use of anonymised mobile telephone traffic data for automotive navigation systems or for public road maintenance. However, many data sources do not share their data as they lack the appropriate economic incentives. Frameworks for the sharing of data should be reviewed, developed and adapted to the new landscape. Governments can lead by example by taking due account of and implementing the principles articulated in the OECD Council Recommendation, *Enhanced Access and More Effective Use of Public Sector Information* (OECD, 2008).

Employment – increasing the availability of needed skills. There are considerable mismatches between the supply of and demand for skills in data management and analytics (data science). This may slow the adoption of big data analytics and lead to missed opportunities for job creation across the economy. Meeting the demand for data analytics skills and expertise at all levels and in all industries calls for a multidisciplinary approach to education, training and skills development in science, technology, engineering and mathematics (STEM) as highlighted by the *OECD Skills Strategy* (OECD, 2011c).

Infrastructure – connecting billions of devices. When the next billion smart devices connect to the Internet and exchange exabytes of data every month, the operation of current communication infrastructures, in particular mobile networks, will be challenged. Issues that governments therefore need to address include: migration to the IPv6 Internet addressing system; opening access to mobile wholesale markets for firms not providing public telecommunication services; and numbering and spectrum policies (regulating the allocation of numbers and radio frequency spectrum as a limited resource for the maximum possible benefit of the public).

Measurement – improving the evidence base. Improved measurement could facilitate the development of policies better tailored to the scale and to the benefits and risks arising from the expanding uses of data. Today, the value of data is poorly captured in economic statistics and often poorly appreciated by organisations and individuals. It is important for governments to work with researchers and firms to understand the potential benefits and risks of applying big data analytics to various sectors in order to develop appropriate policies.

EXPLORING DATA-DRIVEN INNOVATION AS A NEW SOURCE OF GROWTH: MAPPING THE POLICY ISSUES RAISED BY "BIG DATA"

Introduction

This chapter explores the potential of the increasing generation and use of data streams as a resource for enabling the development of new industries, processes and products. While economic and social activities have long made use of data, the scale and influence of information and communication technologies (ICTs) that enable the economic exploitation of data are growing at an extraordinary pace. Declining costs along the data value chain (Figure 1) have been a significant driver of the increasing generation and use of data, as well as the accelerated migration of socioeconomic activities to the Internet thanks to the wide adoption of e-services in an increasingly participative web. The resulting phenomenon – commonly referred to as "big data" – signals the shift towards a data-driven economy, in which data enhance economic competitiveness and drive innovation and equitable and sustainable development.

Figure 1. The data value chain[1] and life cycle[2]

(1) This figure does not include the last phase, "Deletion", which is important for personal data but is considered less important in the context of "big data", where the default is to keep data for long periods if not indefinitely. However, from a policy perspective "Deletion" may deserve a more prominent role.

(2) The output of the "analytics" phase can generate additional data and feed back into the data value chain, leading to a new data life cycle.

To achieve their socioeconomic goals, OECD countries need coherent policy frameworks for the generation, collection, transport and use of data, particularly in areas such as consumer and user empowerment and privacy protection. As access to tools such as smart phones and other smart devices increases, the Internet has a tremendous capacity to enable "crowd sourcing" of consumer and user data in ways that can increase civic engagement and help citizens and consumers in their day-to-day activities. At the same time, these new sources of data, the presence of new actors with access to data, and the increasing ease of linking and transferring data on individuals all test the effectiveness of existing privacy frameworks. The potential policy implications spill over into areas such as access to data, skills and employment, competition, health, and government administration.

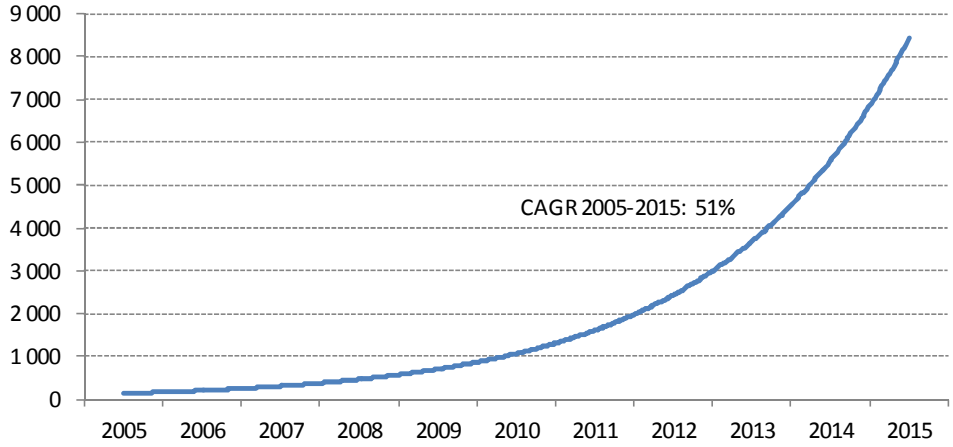
This report seeks first to provide a better understanding of the generation and use of data. It then explores the uses and value of big data across sectors and application areas, and finally describes the main policy opportunities and challenges.

Understanding data and the drivers of their generation and use

The digitisation of nearly all media and the increasing migration of social and economic activities to the Internet (through e-services such as social networks, e-commerce, e-health and e-government) are generating petabytes (millions of gigabytes) of data every second. The social networking site Facebook, for example, is said to have over 900 million active participants around the world and to generate on average more than 1 500 status updates every second (Hachman, 2012; Bullas, 2011). With the increasing deployment and interconnection of (real-world) sensors through mobile and fixed networks (*i.e.* sensor networks), more and more offline activities are also digitally recorded, resulting in an additional tidal wave of data. Measurement in this area is somewhat speculative, but one source suggests that in 2010 alone, enterprises overall stored more than seven exabytes (billions of gigabytes) of new data on disk drives, while consumers stored more than six exabytes of new data (MGI, 2011). This has led to an estimated cumulative data volume of more than 1 000 exabytes in 2010; some estimates suggest that this will multiply by a factor of 40 by the end of this decade (see Figure 2) (IDC, 2012).

Figure 2. Estimated worldwide data storage

in exabytes (billions of gigabytes)



Note: The compound annual growth rate (CAGR) describes the year-over-year growth rate at which worldwide data storage will grow over a specified period of time if it grows at a steady rate.

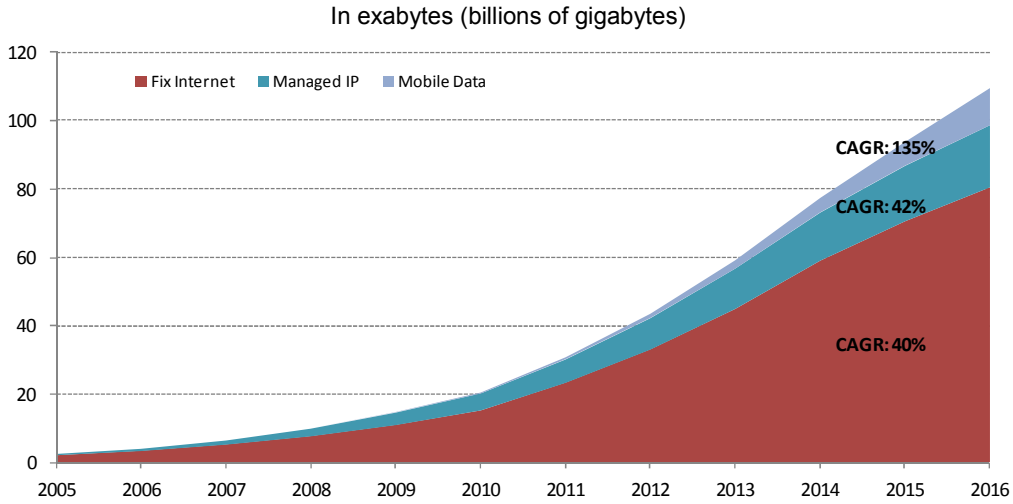
Source: OECD based on IDC Digital Universe research project.

Data generation, collection and transport

The remarkable expansion of data is largely driven by the confluence of important technological developments, notably the increasing ubiquity of broadband access and the proliferation of smart devices and smart ICT applications such as smart meters, smart grids and smart transport based on sensor networks and machine-to-machine (M2M) communication. The large decrease in Internet access costs over the last 20 years has been a significant driver. In 2011, for example, consumers in France paid around the equivalent of USD 33 a month for a broadband connection of 51 Mbit/s compared to the equivalence of USD 75 for a (1 000 times slower) dial-up connection in 1995.[4] Mobile telephones have become a leading data collection device, combining geo-location data and Internet connectivity to support a broad range of new services and applications related to traffic, the environment or health care. Many of these services and applications rely on (or involve) the collection and use of personal data. In addition to increased and more efficient Internet access, most mobile devices are equipped with an increasing array of protocols over which to exchange data locally (*e.g.* Wi-Fi, Bluetooth, Near Field Communications [NFC] with peer-to-peer data transfer capabilities). They may also capture videos, images and sound (often tagged with geo-location information).

In 2011, there were almost six billion mobile subscriptions worldwide of which roughly 13% (780 million) were smart phones capable of collecting and transmitting geo-location data (ITU, 2012; Cisco, 2012). These mobile telephones generated approximately 600 petabytes (millions of gigabytes) of data every month in 2011 (Cisco, 2012).[5] Given that mobile phone penetration (subscriptions per 100 inhabitants) exceeds 100% in most OECD countries and that wireless broadband penetration is at nearly 50%, this source of data will grow significantly as smart phones become the prevalent personal device. Cisco (2012) estimates that the amount of data traffic generated by mobile telephones will reach almost 11 exabytes (billions of gigabytes) by 2016, i.e. almost doubling every year (see Figure 3.).

Figure 3. Monthly global IP traffic, 2005-16



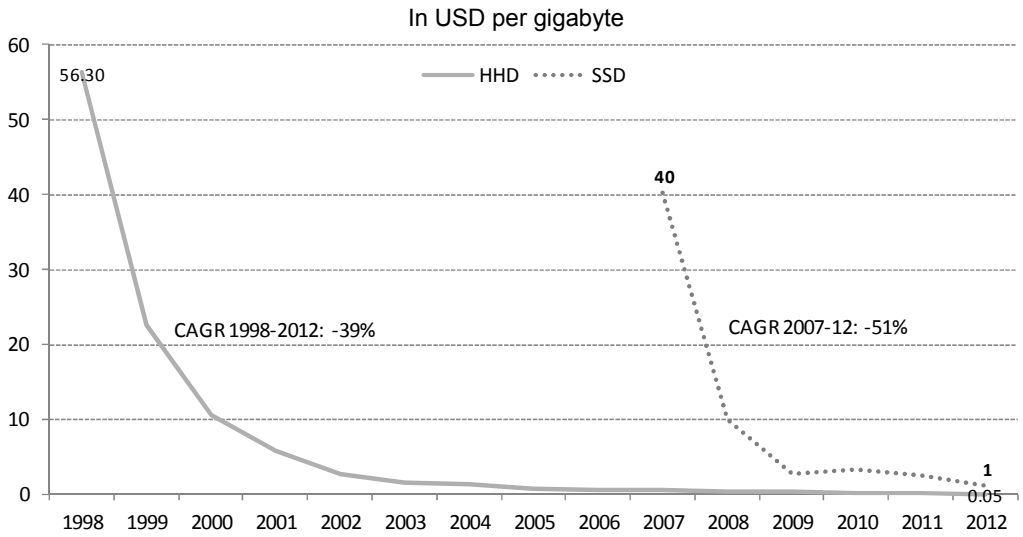
Source: OECD based on Cisco (2012).

The growth in mobile data is not only due to the growing number of mobile telephones, which are expected to account for half of total mobile traffic in 2016 (Cisco, 2012). Other smart devices are proliferating even faster[6]. Smart meters, for example, increasingly collect and transmit real-time data on energy (OECD, 2012a), and smart automobiles are now able to transmit real-time data on the state of the car's components and environment (OECD, 2012b).[7] Many of these smart devices are based on sensor and actuator networks that sense, and may be able to interact with, their environment over mobile networks. The sensors and actuators exchange data through wireless links "enabling interaction between people or computers and the surrounding environment" (Verdone et al., 2008, cited in OECD, 2009a). More than 30 million interconnected sensors are now deployed worldwide, in areas such as security, health care, the environment, transport systems or energy control systems, and their numbers are growing by around 30% a year (MGI, 2011).[8]

Data storage and processing

While the above-mentioned technological developments mainly drive the generation and transport of data, use of the data has been greatly facilitated by the declining cost of data storage, processing and analytics. In the past, the cost of storing data discouraged keeping data that were no longer, or unlikely to be, needed (OECD, 2011b). But storage costs have decreased to the point at which data can generally be kept for long periods of time if not indefinitely. This is illustrated, for example, by the average cost per gigabyte of consumer hard disk drives (HDDs), which dropped from USD 56 in 1998 to USD 0.05 in 2012, an average decline of almost 40% a year (Figure 4). With new generation storage technologies such as solid-state drives (SSDs), the decline in costs per gigabyte is even faster.

Figure 4. Average data storage cost for consumers, 1998-2012

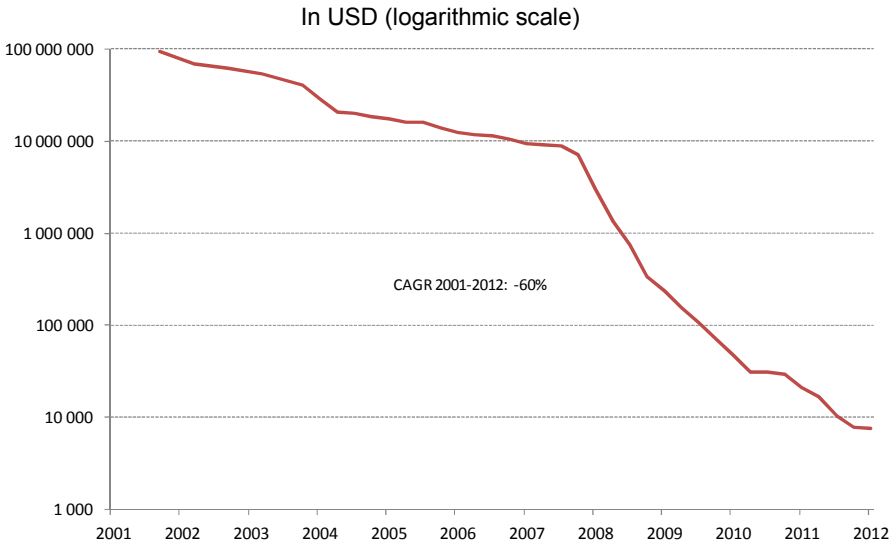


Note: Data for 1998-2011 are based on average prices of consumer-oriented drives (171 HDDs and 101 SSDs) from M. Komorowski (www.mkomo.com/cost-per-gigabyte), AnandTech (www.anandtech.com/tag/storage) and Tom's Hardware (www.tomshardware.com/). The price estimate for SSD in 2012 is based on DeCarlo (2011) referring to Gartner.

Source: OECD based on Pingdom (2011).

Moore's Law, which holds that processing power doubles about every 18 months, relative to cost or size, has largely been borne out. This is particularly noticeable in data processing tools, which have become increasingly powerful, sophisticated, ubiquitous and inexpensive, making data easily searchable, linkable and traceable, not only by governments and large corporations but also by many others. In genetics, for example, DNA gene sequencing machines can now read about 26 billion characters of the human genetic code in less than a minute, and the sequencing cost per genome has dropped by 60% a year on average from USD 100 million in 2001 to less than USD 10 000 in 2012 (Figure 5).

Figure 5. Sequencing cost per genome, 2001-11



Source: OECD based on United States National Human Genome Research Institute (www.genome.gov/sequencingcosts/).

Cloud computing has played a significant role in the increase in data storage and processing capacity. It has been described as "a service model for computing services based on a set of computing resources that can be accessed in a flexible, elastic, on-demand way with low management effort" (OECD, 2012c).[9] In particular, for small and medium-sized enterprises (SMEs), but also for governments that cannot, or do not want to, make heavy upfront investments in ICTs, cloud computing enables organisations to pay for supercomputing resources via a pay-as-you-go model.[10]

Open source software (OSS) applications that cover the full range of solutions needed for big data, such as for storage, processing and analytics (including visualisation), have also contributed significantly to making big data analytics accessible to a wider population. Many big data tools developed initially by Internet firms are now spreading across the economy as enablers of new data-driven goods and services. For instance, Hadoop, an open source programming framework for distributed data management, was inspired by a paper by Google employees Dean and Ghemawat (2004). It was funded initially by Yahoo!, deployed and further developed by Internet firms such as Amazon,[11] Facebook,[12] and LinkedIn,[13] then offered by traditional providers of databases and enterprise servers such as IBM,[14] Oracle,[15] Microsoft,[16] and SAP[17] as part of their product lines, and is now used across the economy for data-intensive operations in companies as diverse as Wal-Mart (retail), Chevron (energy) and Morgan Stanley (financial services).

New participants are entering the data market to trade and exchange data or purchase data-related services. Increasingly specialised data analysts and data brokers offer data for uses such as targeted advertisement, employment background checks, issuing of credit and law enforcement. The number of firms offering data has grown significantly in recent years. At the time of writing, privacyrights.org listed 180 online data brokers registered in the United States alone. Data brokers range from specialised business-to-business companies to simple localisation services.[18] They include companies such as LexisNexis, which claims to conduct more than 12 million background checks a year, and BlueKai Exchange, which claims to be the world's largest data marketplace for advertisers, with data on more than 300 million consumers and more than 30 000 data attributes. According to its website, BlueKai Exchange processes more than 750 million data events and transacts over 75 million auctions for personal information a day.

Defining "big data": volume, velocity and variety, but also value

All the trends described above are present along the data value chain in Figure 1. It is no surprise that these large-scale trends have led some market players to see big data as a new paradigm (Autonomy, 2012; Zinow, 2012). However, the literature offers no clear definition of "big data". Existing definitions tend to focus on volume. Many authors simply describe "big data" as "large pools of data" (McGuire *et al*., 2012). Loukides (2010) defines it as data for which "the size of the data itself becomes part of the problem". The McKinsey Global Institute (MGI, 2011) similarly defines it as data for which the "size is beyond the ability of typical database software tools to capture, store, manage, and analyse".[19] The problem with such definitions is that they are in continuous flux, as they depend on the evolving performance of available storage technologies.

Furthermore, volume is not the only important characteristic. The speed at which data are generated, accessed, processed and analysed is also sometimes mentioned, and analysts have come to use readily available data to make real-time "nowcasts" ranging from purchases of autos to flu epidemics to employment/unemployment trends in order to improve the quality of policy and business decisions (Choi and Varian, 2009; Carrière-Swallow and Labbé, 2010). The Billion Price Project (BPP), launched at MIT and spun off to a firm called PriceStats, collects more than half a million prices on goods (not services) a day by "scraping the web". Its primary benefit is its capacity to provide real-time price statistics that are timelier than official statistics. In September 2008, for example, when Lehman Brothers collapsed, the BPP showed a decline in prices that was not picked up until November by the official Consumer Price Index

(Surowiecki, 2011) (Box 2). Data analytics are also used for security purposes, such as real-time monitoring of information systems and networks to identify malware and cyberattack patterns. The security company ipTrust, for instance, uses Hadoop to assign reputation scores to IP addresses to identify traffic patterns from bot-infected machines in real time (Harris, 2011).

In some cases, big data is defined by the capacity to analyse a variety of mostly unstructured data sets from sources as diverse as web logs, social media, mobile communications, sensors and financial transactions. This requires the capability to link data sets; this can be essential as information is highly context-dependent and may not be of value out of the right context. It also requires the capability to extract information from unstructured data, i.e. data that lack a predefined (explicit or implicit) model. Estimates suggest that the share of unstructured data in businesses could be as high as 80% to 85% and largely unexploited or underexploited. In the past, extracting value from unstructured data was labour-intensive. With big data analytics silos of unexploited data can be linked and analysed to extract potentially valuable information in an automated, cost-effective way.

The potential for automatically linking sets of unstructured data can be illustrated by the evolution of search engines. Web search providers such as Yahoo! initially started with highly structured web directories edited by people. These services could not be scaled up as online content increased. Search providers had to introduce search engines which automatically crawled through "unstructured" web content.[20] Yahoo! only introduced web crawling as the primary source of its search results in 2002. By then Google had been using its search engine (based on its PageRank algorithm) for five years, and its market share in search had grown to more than 80% in 2012.[21]

These three properties – volume, velocity and variety – are considered the three main characteristics of big data and are commonly referred to as the three Vs (Gartner, 2011).[22] However, these are technical properties that depend on the evolution of data storage and processing technologies. Value is a fourth V which is related to the increasing socioeconomic value to be obtained from the use of big data. It is the potential economic and social value that ultimately motivates the accumulation, processing and use of data. It therefore appears appropriate to go beyond the purely technical aspects of volume, velocity and variety to look at the socioeconomic dimension of big data as a "new factor of production" (Gentile, 2011; Jones 2012).

The increasing use and value of data across the economy

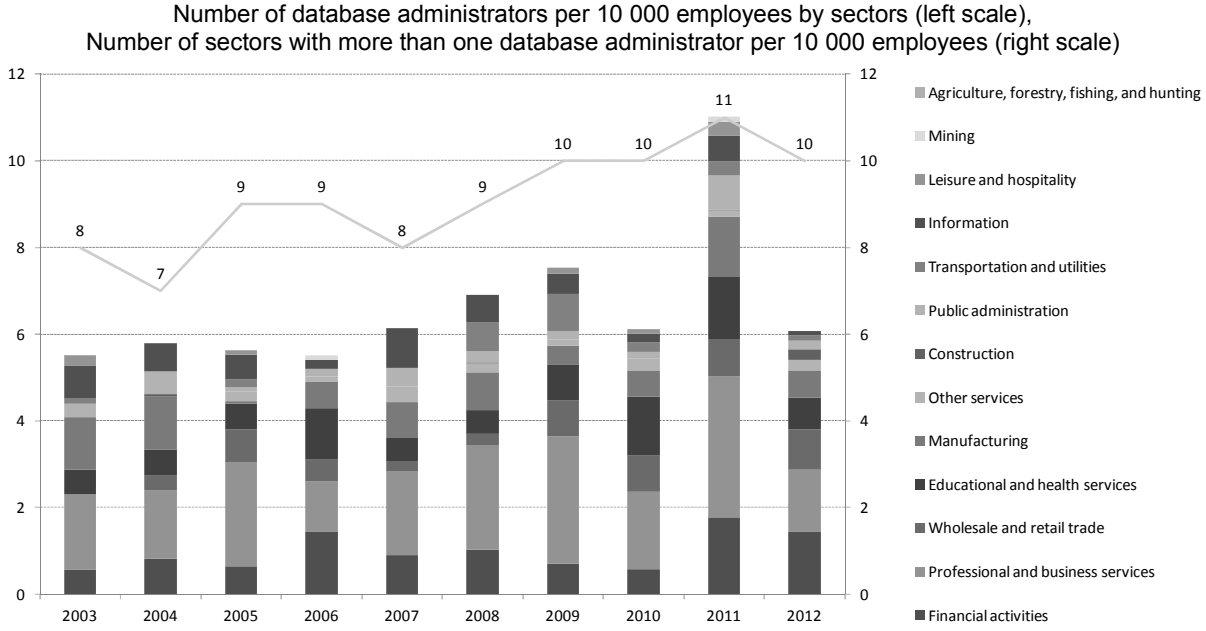
As data storage and processing become increasingly sophisticated, ubiquitous and inexpensive, organisations across the economy are using large data flows for their daily operations. Brynjolfsson et al. (2011) estimate that the output and productivity of firms that adopt data-driven decision making are 5% to 6% higher than would be expected from their other investments in and use of information technology. These firms also perform better in terms of asset utilisation, return on equity and market value. Growing investments in data management and analytics partly reflect the increasing economic role of data. For example, the market value of relational database management systems alone was worth more than USD 21 billion in 2011, having grown on average by 8% a year since 2002. Of perhaps greater interest for big data is the demand for non-relational (noSQL) database systems and business intelligence (BI) and analytics software, which has increased significantly in recent years as data analytics continue to evolve, in particular for data-driven decision making.[23]

The amount of data involved may differ significantly across sectors, as some are more data-intensive than others. According to MGI (2011), data intensity (measured as the average amount of data per organisation) is highest in financial services (including securities and investment services and banking), communication and media, utilities, government, and discrete manufacturing. In these sectors, each organisation stored on average more than 1 000 terabytes (one petabyte) of data in 2009. A similar ranking

can be deduced from the estimated number of data management and analytics professionals (data scientists) per 1 000 employees in each sector. The underlying assumption is that sectors employing more data scientists per 1 000 employees are more data-intensive (see Figure 6).[24]

According to population surveys in the United States, the number of sectors employing one or more database administrators per 10 000 employees has increased over the last nine years. In 2012, the five industries with the largest share of database administrators were: financial activities (22 database administrators per 10 000 employees); professional and business services (12); wholesale and retail trade (6); manufacturing (6); and information (5 together with public administration and other services). The share of database administrators in these sectors has also increased significantly in recent years, with a remarkable peak of more than 160 database administrators per 10 000 employees in the United States in 2011.[25] Most of the data-intensive sectors also tend to have a high ICT intensity (ICT expenditure as a share of output); however, the mining sector had a negligible number of database administrators.[26]

Figure 6. Data intensity of the United States economy, 2003-12



Source: OECD based on the Current Population Survey (March supplement), United States, 2012.

Differences in data intensity suggest that the value of data may differ significantly among sectors (OECD, 2012d).[27] Empirical studies confirm this context dependency not only at the firm level, but also at the employee level (Spiekermann et al., 2001; Acquisti et al., 2011). This makes any assessment of macroeconomic effects much more difficult, and shows the need for case studies to understand the effects in particular sectors or parts of the data value chain.[28]

The following sections briefly present the potential value of data in five sectors. These sectors have been identified in the literature and in previous OECD work as areas of high potential for the use of data as a source of innovation and productivity growth (Cebr, 2012; MGI, 2011; Villars et al., 2012; OECD 2009b; 2012a; 2012b; 2012c). The sectors are: (online) advertisement, public administration, health care, utilities, and logistics and transport. Some of these sectors have been chosen because they have been under-exploiting their data, although they are data-intensive (public administrations, utilities to some extent). Other sectors are less data-intensive today but will face growing amounts of new data, such as click-stream data (online advertisement), geo-location data (transport), smart meter data (utilities), and health records (health care), which, if fully exploited, could generate additional benefits. Together these

sectors account on average for roughly a quarter of total value added in ten OECD countries[29] for which data are available. Overall, the promise of big data lies in one or more of the following innovation-related areas:

- Use of data for the creation of new products (goods and services). This includes using data as a product (data products) or as a major component of a product (data-intensive products);

- Use of data to optimise or automate production or delivery processes (data-driven processes). This includes the use of data to improve the efficiency of distribution of energy resources ("smart" grids), logistics and transport ("smart" logistics and transport). It also includes:

- Use of data to improve marketing, for instance by providing targeted advertisements and personalised recommendations or other types of marketing-related discrimination (data-driven marketing) as well as the use of data for experimental product design (data-driven product design) (Brian, 2012); and

- Use of data for new organisational and management approaches or for significantly improving existing practices (data-driven organisation and data-driven decision making) (Brynjolfsson et al., 2011).

- Use of data to enhance research and development (data-driven R&D). This includes new data-intensive methods for scientific exploration by adding a "new realm driven by mining new insights from vast, diverse data sets" (EC, 2010) (see Box 1).

Box 1. Data-driven science and research

Measurement has always been fundamental to science. The advent of new instruments and methods of data-intensive exploration has prompted some to suggest the arrival of "data-intensive scientific discovery", which builds on the traditional uses of empirical description, theoretical models and simulation of complex phenomena (BIAC, 2011). This could have major implications for how discovery occurs in all scientific fields. Some have challenged the usefulness of models in an age of massive datasets, arguing that with large enough data sets, machines can detect complex patterns and relationships that are invisible to researchers. The data deluge, it is argued, makes the scientific method obsolete, because correlations are enough (Anderson, 2008; Bollier, 2010).

New instruments such as super colliders or telescopes, but also the Internet as a data collection tool, have been instrumental in new developments in science, as they have changed the scale and granularity of the data being collected. The Digital Sky Survey, for example, which started in 2000, collected more data through its telescope in its first week than had been amassed in the history of astronomy (*The Economist*, 2010), and the new SKA (square kilometre array) radio telescope could generate up to 1 petabyte of data every 20 seconds (EC, 2010). Furthermore, the increasing power of data analytics has made it possible to extract insights from these very large data sets reasonably quickly. In genetics, for instance, DNA gene sequencing machines based on big data analytics can now read about 26 billion characters of the human genetic code in seconds. This goes hand in hand with the considerable fall in the cost of DNA sequencing over the last five years (Figure 4).

These new developments, scaled across all scientific instruments and across all scientific fields, indicate the potential for a new era of discovery and raise new issues for science policy. These issues range from the skills that scientists and researchers must master to the need for a framework for data repositories which adheres to international standards for the preservation of data, sets common storage protocols and metadata, protects the integrity of the data, establishes rules for different levels of access and defines common rules that facilitate the combining of data sets and improve interoperability (OSTP, 2010).

Online advertisement[30]

Data generated when consumers use the Internet can create value and give firms opportunities to improve their operations and market their products more effectively. This data-driven marketing is enabled, for example, by the click-stream data collected using some combination of software code such as web-bugs[31] and cookies[32] that allow advertisers to track customers' browsing habits. For individual firms, the exploitation of click-stream data provides new means of improving the management of customer relationships. In the past, when a customer interacted with a firm offline, the information trail was scattered and limited. A firm could only collect scanner data from the checkout for customers using loyalty cards to infer what broader range of products might interest that customer. With click-stream data, firms now possess much more information. For example, firms now have information about the website that directed the user to the firm, whether the user used a search engine, what search terms were used to reach the firm's website. This allows businesses to allocate their marketing budget more effectively and to target websites that reach their most valuable customers. Furthermore, firms can find out exactly what the user looks at on a web page. This enables them to improve users' online experience based on empirical evidence and statistical methods such as A/B testing[33] rather than simply web developers' experience and subjective impressions.[34]

The collection of data is not limited to the firm's website. By using service providers such as social networking sites and advertising networks, firms can also collect data generated elsewhere. Such data are increasingly available through data markets and can be combined with data from sources such as census data, real estate records, vehicle registration and so forth. These enhanced user profiles are then sold to advertisers looking for consumers with particular profiles in order to improve behavioural targeting. For example, comScore, a data broker based in the United States, collects data on the websites visited by over 2 million panellists worldwide, including the search terms they use on search engines and their online purchase and shopping history. comScore then repackages this information to sell reports and data services that illuminate e-commerce sales trends, website traffic and online advertising campaigns. Such reports are sold to Fortune 500 companies and media companies.

Overall, the revenue generated by online advertisement has grown much faster, especially in the last five years, than traditional advertising channels did in their first 15 years. In the first quarter of 2012, online advertising revenues of the top 500 advertisers in the United States, for example, reached USD 8.4 billion, according to the latest IAB Internet Advertising Report (BusinessWire, 2012). This is USD 1.1 billion (15%) more than in the first quarter of 2011. In 2011, AdWords generated more than USD 20 million a month on average from the top 20 websites. This was largely due to the increasing ability to target potential customers and measure results. However, the added value is not limited to advertisement revenue. There are also benefits for consumers. According to McKinsey (2010), consumers in the United States and Europe received EUR 100 billion in value in 2010 from advertising-supported web services. This is three times more than current revenue from advertising and suggests that the consumer value created is greater than advertising revenues would indicate.[35]

Governments and public-sector agencies

The public sector is an important source and user of data. It is in fact one of the economy's most data-intensive sectors. In the United States, for example, public-sector agencies stored on average 1.3 petabytes (millions of gigabytes) of data in 2011,[36] making it the country's fifth most data-intensive sector. However, evidence suggests that the public sector does not exploit the full potential of the data it generates and collects, nor does it exploit the potential of data generated elsewhere (MGI, 2011; Cebr, 2012; Howard, 2012; OECD, 2012e; 2012f). However, improved access to and re-use of public-sector data (PSI) offers many potential benefits, such as improved transparency in the public sector, more efficient, innovative or more personalised delivery of public services, and more timely public policy and decision making.[37]

Estimates suggest that better exploitation of data could significantly increase efficiency, with billions of savings for the public sector. According to MGI (2011), full use of big data in Europe's 23 largest governments might reduce administrative costs by 15% to 20%, creating the equivalent of EUR 150 billion to EUR 300 billion in new value, and accelerating annual productivity growth by 0.5 percentage points over the next ten years.[38] The main benefits would be greater operational efficiency (due to greater transparency), increased tax collection (due to customised services, for example), and fewer frauds and errors (due to automated data analytics). Similar studies of the United Kingdom show that the public sector could save GBP 2 billion in fraud detection and generate GBP 4 billion through better performance management by using big data analytics (Cebr, 2012).

These estimates do not include the full benefits for policy making to be realised from real-time data and statistics. Box 2 describes how such data could be used to better inform the policy-making process.[39] One area of growing interest in this context is internal security and law enforcement. CitiVox, for example, is a start-up that helps governments exploit non-traditional data sources such as SMS (text messages) and social media to complement official crime statistics. Current clients are governments in Central and South America, where a significant share of crimes are not reported.[40] By providing citizens digital means to report crimes, CitiVox's system allows individuals to remain anonymous. At the same time, policy makers and enforcement agencies can mine the incoming data for crime patterns that would not be detected (or not fast enough) through official statistics.

Furthermore, the above estimates do not include benefits achieved through the provision of public-sector information, which is defined by the OECD Council Recommendation on *Enhanced Access and More Effective Use of Public Sector Information* (OECD, 2008) as the wide range of commercially useable "information, including information products and services, generated, created, collected, processed, preserved, maintained, disseminated, or funded by or for the Government or public institution". Beneficial outcomes for economic and social life range from the weather to traffic congestion to local crime statistics to more transparent government functions, such as procurement or educational and cultural knowledge for the wider population in open journals and open data repositories as well as e-libraries.

As the potential of PSI has become more widely recognised, some governments have turned to "open data" initiatives that could accelerate the impact and role of PSI.[41] These initiatives are becoming a valuable means of developing complementary goods and services and have encouraged the emergence of "civic entrepreneurs" that provide social services based on public-sector data.[42] By providing access to and re-use of open government data, governments promote innovative service design and delivery, without the need to build new end-to-end solutions. For instance, citizens increasingly use available PSI to develop mobile phone applications (apps) that facilitate access to existing services and provide new services (m-government).[43] Moreover, through collaboration with online communities, data quality can be improved and the integrity of government data double-checked.

Investments in PSI in the United States have been estimated at tens of billions of USD (Uhlir, 2009). Preliminary modelling suggests that over three decades, the benefits of open access to archives could exceed the costs by a factor of approximately eight (Houghton et al., 2010). Another study, Measuring European Public Sector Information Resources (MEPSIR) (EC, 2006) concludes that the direct PSI re-use market in 2006 for the EU25 plus Norway was worth EUR 27 billion. Based on EC (2006), Vickery (2012) concludes that "the direct PSI-related market would have been around EUR 32 billion in 2010".

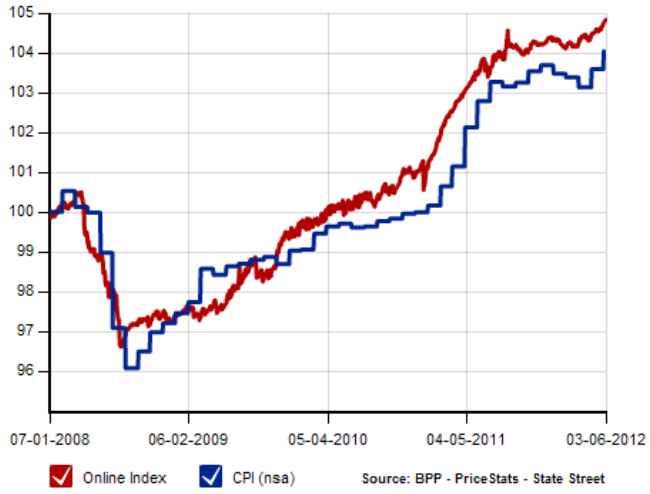
Box 2. Data proliferation and implications for official statistics

Torrents of data streaming across public and private networks can improve the quality of statistics in an era of declining responses to national surveys and can create close to real-time evidence for policy making in areas such as prices, employment, economic output and development, and demographics. Some of the new sources of statistics are search engine data derived from keywords entered by users searching for web content. Google Insights for Search, for example, provides statistics on the regional and time-based popularity of specific keywords. Where keywords are related to specific topics such as unemployment, Google Insights can provide real-time indicators for measuring and predicting unemployment trends. Askitas and Zimmermann (2009), for example, analyse the predictive power of keywords such as "Arbeitsamt OR Arbeitsagentur" ("unemployment office or agency") for forecasting unemployment in Germany. The authors find that the forecast based on these keywords indicated changes in trends much earlier than official statistics. Similar conclusions have been drawn by D'Amuri and Marcucci (2010) for the United States and by Suhoy (2010) for Israel.

Other statistics are created by directly "scraping" the web. The Billion Price Project (BPP), for example, collects price information over the Internet to compute a daily online price index and estimate annual and monthly inflation. The online price index is basically an average of all individual price changes across all retailers and categories of goods. More than half a million prices on goods (not services) are collected daily by "scraping" the content of online retailers' websites such as Amazon.com. This is not only five times what the US government collects, it is also cheaper because the information is not collected by researchers who visit thousands of shops as they do for traditional inflation statistics. Furthermore, unlike official inflation numbers, which are published monthly with a lag of weeks, the online price index is updated daily with a lag of just three days. In addition, the BPP has a periodicity of days as opposed to months. This allows researchers and policy makers to identify major inflation trends before they appear in official statistics. For example, in September 2008, when Lehman Brothers collapsed, the online price index showed a decline in prices, a movement that was not picked up until November by the CPI (Figure 7; Surowiecki, 2011).

Figure 7. Daily online price index, United States, 2008-2012

Index, 100 = 01 July 2008



Source: bpp.mit.edu.

Currently, while methods to mine these new sources are still in their infancy and need rigorous scientific scrutiny, their rapid take-up by policy makers is a harbinger of a growing trend. Governments in the United States, the United Kingdom, Germany and France and in major non-OECD countries such as Brazil have established a partnership with PriceStats, which manages the BPP index, to contribute to and use the index. In another example, the Central Bank of Chile has explored the use of Google Insight for Search to predict present (to "nowcast") economic metrics related to retail good consumption (Carrière-Swallow and Labbé, 2010). For developing economies, in particular, where NSOs' capacity to sufficiently inform policy makers is often low, the exploitation of these new data sources through public-private cooperation provide a new opportunity to better inform public policy making for development (UN Globalpulse, 2012).[44]

Source: OECD (2012g).

Health care

The health-care sector sits on a growing mountain of data generated by the administration of the health system and the diffusion of electronic health records. Diagnostic tests, medical images and the banking of biological samples are also generating new data. There are now vast collections of medical images, with 2.5 petabytes (millions of gigabytes) stored each year from mammograms in the United States alone (EC, 2010).

To some extent what has been said about the benefits of data for the public sector is also true for the health sector, as better use of data can have significant impacts, both within the sector and across the economy. Health-sector data may improve the effectiveness, safety and patient-centeredness of health-care systems and also help researchers and doctors measure outcomes, identify previously unobserved correlations, and even forecast changes in essential clinical processes and interventions (Bollier, 2010). When population data from different sources are linked to health-sector data, some causes of illness can be better understood. An example is the analysis of environmental determinants of illnesses linked to nutrition, stress and mental health (OECD-NSF, 2011).[45]

The sharing of health data through electronic health records can facilitate access to medical care and may provide useful insights for product and services innovation, including research on new medicines and therapies. Other sources of personal health data may include remote monitoring applications that collect data on specific clinical conditions or on daily living conditions, for example to learn when a frail person needs help. Personal health data are also increasingly supplied by individuals and stored and exchanged on line through health-focused social networks. The social network PatientsLikeMe not only allows people with a medical condition to interact with, derive comfort and learn from other people with the same condition, it also provides an evidence base of personal data for analysis and a platform for linking patients with clinical trials. The business model depends on aligning patients' interests with industry interests; PatientsLikeMe sells aggregated, de-identified data to partners, including pharmaceutical companies and makers of medical devices, to help them better understand the actual experience of patients and the effective course of a disease. PatientsLikeMe also shares patient data with research collaborators around the world.

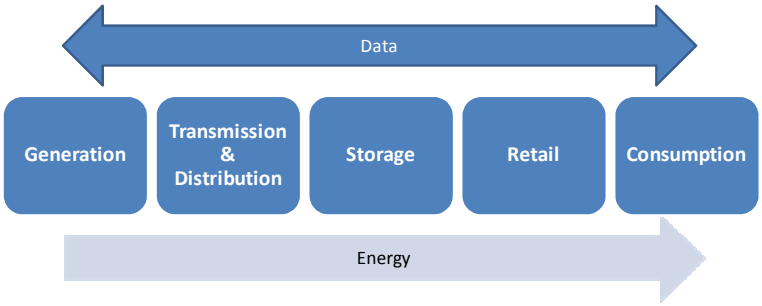
Large health providers such as Kaiser Permanente (a managed-care consortium in the United States) use these data sets to discover the unforeseen adverse affects of drugs such as Vioxx which were not detected in clinical trials but were discovered by mining the data generated as the drug was prescribed and used (MGI, 2011). The United Kingdom National Institute of Health and Clinical Experience has also used large clinical datasets to investigate the cost effectiveness of new drugs and treatments, leading to improved outcomes at a lower cost. More generally, linked data could reduce the costs associated with under- or over-treatment; they could also help combat chronic diseases by determining behavioural causes and thus guide intervention before the onset of disease (Bollier, 2010). MGI (2011) estimates that big data could be used throughout the US health-care system – clinical operations, payment and pricing of services, and R&D – at a savings of more than USD 300 billion, two-thirds of which would come from reducing health-care expenditures by 8%. These estimates, however, do not include the benefits of data analytics for enabling timely public health policies through real-time statistics such as those provided by web search data to assess flu trends in real time (Polgreen et al., 2009; Ginsberg et al., 2009; Valdivia and Monge-Corella, 2010 as well as Box 2 on the use of new data sources for official statistics).

Utilities[46]

"Smart" utilities are deployed for more efficient generation, distribution and consumption of energy, but increasingly also for other natural resources such as water. For example, "smart" grids are electricity networks with enhanced information and communication capacities that can address major electricity

sector challenges along the value chain from energy generation to consumption (Figure 8). These challenges include managing consumption peaks, which are typically CO_2 expensive, and the integration of volatile renewable energy sources during energy generation and reducing losses in energy transmission and distribution.[47]

Figure 8. Stylised electricity sector value chain with energy and data flows



"Smart" utilities rely heavily on data collected through "smart meters" at households and other consumers of energy and resources. These smart devices enable bi-directional communication across the value chain, enabling not only real-time collection of consumption data but also the exchange of real-time price data and signals to control the turning on or shutting off of various appliances in households and industries. Estimates suggest that connecting one million homes to a smart grid may produce as much as 11 gigabytes of data a day; this could create significant challenges for data management and analytics (OECD, 2009b). In order to accommodate hourly readings, a network with a minimum capacity of up to 1 Mbit/s could be needed (GE, 2007; IEEE, 2009; OECD, 2009b). While the information feedback loop allows consumers to adjust their consumption to production capacities, utilities can now run data analytics to identify overall consumption patterns and forecast demand. This can help them adjust their production capacities and pricing mechanisms to future demand.[48] Overall, according to GeSI (2008), the use of data-driven smart-grid applications could reduce CO_2 emissions by more than 2 gigatonnes (the equivalent of EUR 79 billion).

Furthermore, data collected from distribution networks allow utility providers to identify losses and leakages during the distribution of energy and other resources. By deploying smart water sensors in combination with data analytics, Aguas Antofagasta, a water utility in Chile, was able to identify water leaks throughout their distribution networks and reduce total water losses from 30% to 23% over the past five years, thereby saving some 800 million litres of water a year.

As in the case of public-sector data, opening smart meter data to the market has led to a new industry that provides innovative goods and services based on these data which have contributed to green growth and created a significant number of green jobs. Opower, for example, is a US-based start-up that partners with utility providers to promote energy efficiency based on smart-meter data analytics. The company successfully raised USD 14 million in venture capital (VC) funding in 2008 and USD 50 million two years later. Three years after its creation Opower employed more than 230 people.

Logistics and transport

The logistics and transport sector is less data-intensive but is facing growing amounts of data. These may make it possible to increase the efficiency of transporting goods and persons through smart routing and through new services based on smart applications.

Smart routing is based on the real-time traffic data that are used, but increasingly also collected, by navigation systems. Some of these systems are dedicated hardware devices, but the large majority of

personal navigation systems are expected to be operated as software running on smart phones or integrated in automobiles. These applications are very data-intensive. For example, TomTom, a leader in navigation hardware and software, had in its databases in 2012 more than 5 000 trillion data points from its navigation devices and other sources, describing time, location, direction and speed of individual anonymised users,[49] and it adds 5 billion data points every day.[50] Overall, estimations by MGI (2011) suggest that the global pool of personal geo-location data was at least 1 petabyte in 2009, and growing by about 20% a year. By 2020, this data pool is expected to provide USD 500 billion in value worldwide in the form of time and fuel savings or 380 million tonnes of CO_2 emissions saved. This does not include value provided through other location-based services.

As well as navigation system providers such as TomTom, others also provide significant amounts of data. For example, mobile network operators use cell-tower signals to triangulate the location of mobile telephone users and to identify patterns related to accidents and congestions based on data analytics. These data and inferred information are sold to providers of navigation systems, but also to third parties such as governments. For example, the French mobile telecommunication services firm Orange uses its Floating Mobile Data (FMD) technology to collect mobile telephone traffic data to determine speeds and traffic density at a given point of the road network, and deduce travel time or the formation of traffic jams. The anonymised mobile telephone traffic data are sold to third parties, including government agencies, to identify hot spots for public interventions, but also to private companies such as Mediamobile, a leading provider of traffic information services in Europe.[51]

Another area in which the use of data promises significant benefits in the logistics and transport sector is the use of smart applications based on machine-to-machine (M2M) communication. Smart automobiles, for example, are increasingly equipped with sensors to monitor and transmit the state of the car's components as well as of the environment in which the car is moving. This enables services such as OnStar and Sync, which are offered by vehicle manufacturers to car owners and include theft protection and navigation and emergency services. New business models and new forms of fees and taxes, such as dynamic road pricing based on GPS and M2M data, are also providing significant added value. MGI (2011) estimates that by 2020 the use of automatic toll collection based on the location of mobile telephones will generate from USD 4 billion to USD 10 billion in value to final consumers and USD 2 billion in revenue to services providers.

Mapping the policy opportunities and challenges

With the increasing exploitation of data across the economy comes a wide array of policy opportunities and challenges, many of which were identified at the 2012 OECD Technology Foresight Forum, Harnessing data as a new source of growth – Big data analytics and policies (see Box 3).

Box 3. OECD Technology Foresight Forum 2012:
Harnessing data as a new source of growth - Big data analytics and policies

The 2012 Technology Foresight Forum (the Foresight Forum), held on 22 October 2012, highlighted the potential of big data analytics as a new source of growth. It put big data analytics in the context of key technological trends such as cloud computing, smart ICT applications and the Internet of Things. It focused on the socioeconomic implications of harnessing data as a new source of growth and looked at specific areas: science and research (including public health), marketing (including competition) and public administration.

Participants discussed specific potential policy opportunities and challenges. They stressed the tremendous potential of big data in science and research (including for health care), retail, finance and insurance, and public-service delivery. They noted the opportunity costs of not using data and the need to measure the socioeconomic value of data use and re-use. Participants also discussed the changes needed in mindsets of individuals, businesses and policy makers to understand the "big data phenomenon" and to be able to capture the potential benefits while handling the associated risks. Among challenges, they frequently emphasised privacy and consumer protection in association with the issue of consent and the current limitations on anonymisation and de-identification due to big data analytics. They noted that big data analytics were changing the nature of digital identity and thus the relationship between identity and privacy.

Participants also drew attention to issues related to open *vs*. closed data and the related issue of data ownership and control. They discussed the implications of big data analytics for employment, and stressed the need for new skills and improved awareness across all industries and all organisational levels in order to ensure that the economy makes good use of data. In particular, they warned that big data may put white collar jobs at risk (including professional, managerial or administrative workers), just as the industrial revolution did for blue collar jobs (and workers mainly performing manual labour).

Participants considered that the ethical dimension of big data analytics is increasingly important. They cited rules of ethics such as "just because you can, doesn't mean you should". In this spirit, a speaker compared the big data phenomenon with nuclear energy in the early 20th century: "It's coming whether we want it or not. What we can do is promote the responsible use of big data".

Source: OECD, http://oe.cd/tff2012.

The following sections introduce policy issues raised by the application of large-scale data analytics across the economy. Some of these issues – related to privacy, open access to data, including public-sector information, ICT skills and employment, and infrastructure – are not new. In the case of privacy protection, problems related to "data mining" and "profiling" are long-standing. What is novel is that it is increasingly easy to infer information about individuals, even if they have never deliberately shared this information with anyone. As an illustration, Target, a United States retailer, knew that a teenage girl was pregnant before her father did (Hill, 2012). In a context in which the volume, variety, velocity and economic value of data are constantly increasing, policy issues related to intellectual property rights (IPR), competition, corporate reporting and taxation gain in importance. These policy issues are not discussed here. Specific issues related to the health sector are discussed in OECD (2012h). The challenges and opportunities of big data for national statistics agencies are examined in OECD (2012g).

Privacy and consumer protection

OECD member countries have adopted various mechanisms to protect the privacy of individuals as regards the processing of their personal data. These regulatory instruments largely reflect the "basic principles of national application" contained in the OECD (1980) Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data ("the Privacy Guidelines", see Box 4), which are currently under review.

The Privacy Guidelines define personal data as "any information relating to an identified or identifiable individual (data subject)". Any data that are not related to an identified or identifiable individual are therefore non-personal and are outside the scope of the Guidelines. However, data analytics have made it easier to relate seemingly non-personal data to an identified or identifiable individual (Ohm, 2010). Furthermore, big data applications may affect individuals using data which are generally considered non-personal (Hildebrandt and Koops, 2010). These developments challenge a regulatory approach that determines the applicability of rights, restrictions and obligations on the basis of the "personal" nature of the data involved. As the scope of non-personal data is reduced, the difficulty of applying existing frameworks effectively become more acute.

Many data-driven goods and services also raise issues for the application of the basic principles of data protection, such as purpose specification and use limitation.[52] These goods and services offer opportunities for beneficial re-use of personal data, often in ways not envisaged when they were collected. They also implicitly rely on the lengthy retention of information. As such, they stretch the limits of existing privacy frameworks, many of which take limits on the collection and storage of information, and on its potential uses, as a given (Tene and Polonetsky, 2012).

The increased complexity of data-driven goods and services also makes it more difficult to provide individuals with comprehensive and comprehensible information about the collection and use of personal data (see Box 4). The sheer scale of data processing lessens the ability of individuals to participate in the processing of their personal data (Cavoukian and Jonas, 2012). As the amount of personal data grows, and the number of actors involved in using them expands, it may be necessary to reconsider the appropriate roles of different types of actors. For commercial transactions, in particular, consumers' access to their personal data is being regarded as increasingly important for empowering consumers to drive innovation and enhance competition in the marketplace. This access would help consumers make better informed decisions by being able to compare prices, get an overview of their transactions history, look at the value of their own data, and thus actively participate in the data-driven economy.[53]

When the Privacy Guidelines were adopted, data flows involved a limited number of data sources, which were connected through closed networks. This environment allowed policy makers to make a single actor (the "data controller") responsible for every aspect of processing (collection, use, security, data quality, etc.). The transition from a closed network environment to an open network environment has made it increasingly difficult to maintain this approach. Instead of discrete, well-defined transfers of information, many data-driven goods and services typically involve a multiplicity of information flows, with many different actors, each of which exercises varying degrees of control. This changed environment has introduced an additional level of complexity (Burdon, 2010). For example, services such as cloud computing and social networking often involve many different types of actor, each of which influences the collection and use of information to a different degree. These developments may imply the need for more adaptable and flexible allocation of responsibilities.

Box 4. Basic principles of national application of the OECD (1980) Privacy Guidelines (part 2)

Collection limitation principle

There should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject.

Data quality principle

Personal data should be relevant to the purposes for which they are to be used and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date.

Purpose specification principle

The purposes for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfilment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose.

Use limitation principle

Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance with Paragraph 9 except:

a) with the consent of the data subject; or

b) by the authority of law.

Security safeguards principle

Personal data should be protected by reasonable security safeguards against such risks as loss or unauthorised access, destruction, use, modification or disclosure of data.

Openness principle

There should be a general policy of openness about developments, practices and policies with respect to personal data. Means should be readily available of establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.

Individual participation principle

An individual should have the right:

a) to obtain from a data controller, or otherwise, confirmation of whether or not the data controller has data relating to him;

b) to have communicated to him, data relating to him

1. within a reasonable time;
2. at a charge, if any, that is not excessive;
3. in a reasonable manner; and
4. in a form that is readily intelligible to him;

c) to be given reasons if a request made under subparagraphs (a) and (b) is denied, and to be able to challenge such denial; and

d) to challenge data relating to him and, if the challenge is successful to have the data erased, rectified, completed or amended.

Accountability principle

A data controller should be accountable for complying with measures which give effect to the principles stated above.

Although the Privacy Guidelines call for specification of purpose prior to the collection and use of personal data, they do not restrict the nature or types of purposes for which personal data may be used. This approach has left the contours of responsible data usage largely undefined. For example, one might ask: "Where does the boundary reside between, on the one hand, improving customer relationships, and, on the other, unfair consumer manipulation? When does risk optimisation become unfair discrimination?"

Open access to data

The linking and use of data across sectors can drive innovation and generate socioeconomic benefits. Examples includes the use of PSI across the economy by BrightScope or the sale of anonymised telecommunication data collected by Orange to traffic information service providers such as TomTom or MediaMobile. They suggest that open access to data can lead to significant economic benefits.

However, appropriate sharing of data across the economy requires more robust frameworks. Many sources of third-party data do not yet consider sharing their data, and economic incentives may not be aligned to encourage it (MGI, 2011). More needs to be known about pricing and licensing models, but also about ownership and control mechanisms, including intellectual property rights (IPR) regimes.[54] Objective pricing of information is notoriously complex, and identification of the different cost components may be somewhat arbitrary (Shapiro and Varian, 1998). For PSI in particular, the circumstances under which the public sector should produce value-added products from its assets continue to be debated. Many governments wish to recover costs, partly for budgetary reasons and partly on the grounds that those who benefit should pay. However, the calculation of benefits can be problematic. Moreover, as Stiglitz *et al.* (2000) have argued, if government provision of a data-related service is a valid role, generating revenue from that service is not.

The public sector has nevertheless led the way in opening up its data to the wider economy through various "open data" initiatives. The OECD (2008) Council Recommendation for *Enhanced Access and More Effective Use of Public Sector Information*, which is currently under review, describes a set of principles and guidelines for access to and use of PSI; among these, openness is the first principle (Box 5). The Recommendation refers to the OECD (2005) *Principles and Guidelines for Access to Research Data from Public Funding*, which also highlight openness as its principle. This latter Recommendation in particular specifies that "openness means access on equal terms for the international research community at the lowest possible cost, preferably at no more than the marginal cost of dissemination. Open access to research data from public funding should be easy, timely, user-friendly and preferably Internet-based". Open data initiatives are also emerging in the private sector. The Open Knowledge Foundation, for instance, has established an open data framework, which defines open data as "a piece of content or data (which) is open if anyone is free to use, reuse, and redistribute it – subject only, at most, to the requirement to attribute and/or share-alike".[55]

Box 5. Principles of the OECD (2008) Recommendation for Enhanced Access and More Effective Use of Public Sector Information

Openness. Maximising the availability of public sector information for use and re-use based upon presumption of openness as the default rule to facilitate access and re-use. Developing a regime of access principles or assuming openness in public sector information as a default rule wherever possible no matter what the model of funding is for the development and maintenance of the information. Defining grounds of refusal or limitations, such as for protection of national security interests, personal privacy, preservation of private interests for example where protected by copyright, or the application of national access legislation and rules.

Access and transparent conditions for re-use. Encouraging broad non-discriminatory competitive access and conditions for re-use of public sector information, eliminating exclusive arrangements and removing unnecessary restrictions on the ways in which it can be accessed, used, re-used, combined or shared, so that in principle all accessible information would be open to re-use by all. Improving access to information over the Internet and in electronic form. Making available and developing automated on-line licensing systems covering re-use in those cases where licensing is applied, taking into account the copyright principle below.

Asset lists. Strengthening awareness of what public sector information is available for access and re-use. This could take the form of information asset lists and inventories, preferably published on-line, as well as clear presentation of conditions to access and re-use at access points.

Quality. Ensuring methodical data collection and curation practices to enhance quality and reliability including through co-operation of various government bodies involved in the creation, collection, processing, storing and distribution of public sector information.

Integrity. Maximising the integrity and availability of information through the use of best practices in information management. Developing and implementing appropriate safeguards to protect information from unauthorised modification or from intentional or unintentional denial of authorised access to information.

New technologies and long-term preservation. Improving interoperable archiving, search and retrieval technologies and related research including research on improving access and availability of public sector information in multiple languages, and ensuring development of the necessary related skills. Addressing technological obsolescence and challenges of long-term preservation and access. Finding new ways for the digitisation of existing public sector information and content, the development of born-digital public sector information products and data, and the implementation of cultural digitisation projects (public broadcasters, digital libraries, museums, etc.) where market mechanisms do not foster effective digitisation.

Copyright. Intellectual property rights should be respected. There is a wide range of ways to deal with copyrights on public sector information, ranging from governments or private entities holding copyrights, to public sector information being copyright-free. Exercising copyright in ways that facilitate re-use (including waiving copyright and creating mechanisms that facilitate waiving of copyright where copyright owners are willing and able to do so, and developing mechanisms to deal with orphan works), and where copyright holders are in agreement, developing simple mechanisms to encourage wider access and use (including simple and effective licensing arrangements), and encouraging institutions and government agencies that fund works from outside sources to find ways to make these works widely accessible to the public.

Pricing. When public sector information is not provided free of charge, pricing public sector information transparently and consistently within and, as far as possible, across different public sector organisations so that it facilitates access and re-use and ensures competition. Where possible, costs charged to any user should not exceed the marginal costs of maintenance and distribution, and in special cases extra costs associated, for instance, with digitisation. Basing any higher pricing on clearly expressed policy grounds.

Competition. Ensuring that pricing strategies take into account considerations of unfair competition in situations where both public and business users provide value-added services. Pursuing competitive neutrality, equality and timeliness of access where there is potential for cross-subsidisation from other government monopoly activities or reduced charges on government activities. Requiring public bodies to treat their own downstream/value-added activities on the same basis as their competitors for comparable purposes, including pricing. Particular attention should be paid to single sources of information resources. Promoting non-exclusive arrangements for disseminating information so that public sector information is open to all possible users and re-users on non-exclusive terms.

Redress mechanisms: Providing appropriate transparent complaints and appeals processes.

Public private partnerships. Facilitating public-private partnerships where appropriate and feasible in making public sector information available, for example by finding creative ways to finance the costs of digitisation, while increasing access and re-use rights of third parties.

International access and use. Seeking greater consistency in access regimes and administration to facilitate cross-border use and implementing other measures to improve cross-border interoperability, including in situations where there have been restrictions on non-public users. Supporting international co-operation and co-ordination for commercial re-use and non-commercial use. Avoiding fragmentation and promote greater interoperability and facilitate sharing and comparisons of national and international datasets. Striving for interoperability and compatible and widely used common formats.

Best practices. Encouraging the wide sharing of best practices and exchange of information on enhanced implementation, educating users and re-users, building institutional capacity and practical measures for promoting re-use, cost and pricing models, copyright handling, monitoring performance and compliance, and their wider impacts on innovation, entrepreneurship, economic growth and social effects.

Cybersecurity risks

As the volume and value of data stored increases so does the risk of data breaches. According to company surveys, reported thefts of electronic data surpassed losses of physical property as the major crime problem for global companies for the first time in 2010 (Masters and Menn, 2010; Kroll, 2012). This demonstrates the increasing corporate value of intangible assets, such as data, as compared to tangible assets.

Data collected by the Privacy Rights Clearinghouse, for example, show that large-scale data breaches, i.e. those involving more than 10 million records, are becoming more frequent. Examples include the 2008-09 malicious software hack that compromised Heartland Payment Systems Inc. (an online payments and credit card company based in the United States), affecting more than 130 million credit and debit card numbers (Voreacos, 2009; Zetter, 2009), and the security breach of Sony's PlayStation Network and the Sony Online Entertainment systems in 2010-11 which resulted in the exposure of 104 million records of personally identifiable information including names, addresses, birthdates, passwords and logins, among others (Reuters, 2011; Seybold, 2011; Goodin, 2011).

Anecdotal evidence also shows an increasing number of so-called advanced persistent threats (APTs). These are typical cyberespionage incidents often targeting a sector's key organisations or key competitors to steal data or different forms of intellectual property and to reduce these organisations' competitive advantage. Operation Shady Rat was an APT that compromised more than 70 companies, governments and non-profit organisations in 14 countries (McAfee, 2011). Operation Red October targeted government, military, aerospace, research, trade and commerce, nuclear, and oil organisations in two dozen countries (DeCarlo, 2013).[56] Reports and statements by officials in the United Kingdom (Esposito, 2012) and the United States (NCIX, 2012) have noted an increase in industrial cyberespionage activities. Yet, the scale of the phenomenon is uncertain as victims are reluctant to disclose information about successful attacks (Severs, 2013).

As data usage today requires information systems and networks to be more open, organisations are obliged to adapt their security policy to the more open and dynamic environment in which data are widely exchanged and used. The OECD 2002 Security Guidelines, currently under review, were designed to promote an approach to security that enables rather than restricts such openness at the technical level (Box 6). Such an approach is particularly important for seizing the benefits of a data-driven economy.

Box 6. Principle of the OECD Guidelines for the Security of Information Systems and Networks

1) Awareness: Participants should be aware of the need for security of information systems and networks and what they can do to enhance security.

2) Responsibility: All participants are responsible for the security of information systems and networks.

3) Response: Participants should act in a timely and co-operative manner to prevent, detect and respond to security incidents.

4) Ethics: Participants should respect the legitimate interests of others.

5) Democracy: The security of information systems and networks should be compatible with the essential values of a democratic society.

6) Risk assessment: Participants should conduct risk assessments.

7) Security design and implementation: Participants should incorporate security as an essential element of information systems and networks.

8) Security management: Participants should adopt a comprehensive approach to security management.

9) Reassessment: Participants should review and reassess the security of information systems and networks, and make appropriate modifications to security policies, practices, measures and procedures.

Skills and employment

A pool of qualified personnel with skills in data management and analytics (data science) is essential for the success of a "smarter" data-driven economy (OECD, 2012i). However, these skills must also be specific to some extent, as they require an appropriate mix of advanced ICT skills, skills in statistics and specific knowledge of the sector involved (see *OECD Skills Strategy*, OECD 2012j). Demand for highly specialised skills is expected to intensify as data analytics proliferate, and a shortage of data scientists is likely in the near future. MGI (2011), for example, estimates that the demand for deep analytical positions in the United States could exceed supply by 140 000 to 190 000 positions by 2018. This does not include the need for an additional 1.5 million managers and analysts who can use big data knowledgeably.

In the past, there have been considerable mismatches between the supply of and demand for ICT skills in general and for software skills in particular. Shortfalls in domestic supply (owing to a large share of students leaving compulsory education, lack of educational courses and little training in the industry), restrictions on immigration of highly skilled personnel, or difficulties in international sourcing of development and analytical tasks requiring large amounts of interaction among employees are continuing challenges, as is the relatively low number of female employees in the ICT industry (OECD, 2012i).

However, data science skills are not only obtained from formal university or tertiary institution degree courses in specific study programmes such as computer science. Scientific fields that require the analysis of large data sets also provide a good source of data scientists. In fact, a significant number of data scientists have a degree in experimental physics, molecular biology, bioinformatics or computer science with an emphasis on artificial intelligence (Loukides, 2010; Rogers, 2012). Despite the availability of these skills across OECD economies, anecdotal evidence suggest that most employees working as data scientists are located in the United States.[57]

Beyond the high level of expected demand for data scientists, the full implications of big data for employment are not yet well understood. Increased labour productivity resulting from the use of data

analytics may lead to the disappearance of some jobs that previously required human labour (*e.g.* Google's Driverless Car could replace taxi drivers). The ability to mine vast amounts of data to optimise logistics, customer relations and sales could also have a significant impact on jobs of a "transactional" nature (Brynjolfsson and McAfee, 2011). While productivity-enhancing, this structural change comes at a time when the economy is fragile and it may exacerbate the weak employment market and the bias towards higher skills and inequality in earnings.

Infrastructure

As noted earlier in the chapter, the availability of high-speed broadband access, in particular mobile broadband access, has greatly facilitated the collection, transport and use of data in the economy. It is estimated that households across the OECD area now have an estimated 1.8 billion connected smart devices (OECD, 2013). The number could reach 5.8 billion in 2017 and 14 billion in 2022. This will require governments to address the issue of the migration to a new Internet addressing system (IPv6). The current IPv4 addresses are essentially exhausted, and mechanisms for connecting the next billion devices are urgently needed. IPv6 offers one solution. It is a relatively new addressing system that offers the possibility of almost unlimited address space, but adoption has been relatively slow. Furthermore, as many data-intensive smart applications rely on machine-to-machine (M2M) communication, this raises regulatory challenges related to opening access to mobile wholesale markets to firms not providing public telecommunication services and to numbering policy and frequency policy issues (see Box 7).

Box 7. Transmitting data – a regulatory barrier to machine-to-machine communication

In the near future, the Internet will connect things as well as people. Companies will change how they design machines and devices. They will first define the data needed and then build the machine. Tens of billions of devices are likely to be connected by 2025. A new type of user of mobile networks will emerge – the million-device user (such as car, consumer electronics and energy companies, and health providers, whose vehicles and devices connect to the Internet). M2M communication will become standard.

Mobile networks are best geared to geographically mobile and dispersed users who want to be connected everywhere and all the time. However, a major barrier for the million-device user is the lack of competition once a mobile network provider has been chosen. The problem is the SIM card, which links the device to a mobile operator. By design, only the mobile network that owns the SIM card can designate which networks the device can use. In mobile phones the SIM card can be removed by hand and changed for that of another network. But when used in cars or other machines it is often soldered, to prevent fraud and damage from vibrations. Even if it is not soldered, changing the SIM at a garage, a customer's home, or on-site, costs USD 100-USD 1 000 per device.

Consequently, once a device has a SIM card from a mobile network, the company that developed the device cannot leave the mobile network for the lifetime of the device. Therefore, the million-device user can effectively be locked into 10- to 30-year contracts. It also means that when a car or e-health device crosses a border, the large-scale user is charged the operator's costly roaming rates. The million-device user cannot negotiate these contracts. It also cannot distinguish itself from other customers of the network (normal consumers) and is covered by the same roaming contracts.

There are many technological and business model innovations that a large-scale M2M user might want to introduce. However, at present, it cannot do so, because it would need the approval of its mobile network operator. Many innovations would bypass the mobile operator and therefore are resisted. The solution would be for governments to allow large-scale M2M users to control their own devices by owning their own SIM cards, something that is implicitly prohibited in many countries. It would make a car manufacturer the equivalent of a mobile operator from the perspective of the network. Removing regulatory barriers to entry in this mobile market would allow the million-device customer to become independent of the mobile network and create competition. This would yield billions in savings on mobile connectivity and revenue from new services.

Source: OECD (2012b).

Measurement

Improved measurement could facilitate the development of policies better tailored to the scale, benefits and risks of the expanding uses of data. It would mean better understanding the value added of data-driven activities, including data processing and data storage activities, identification of sectors in which data are a key intangible asset, and better recognition of the impact of framework conditions on the collection, distribution and use of data across the economy. At present, the value of data-driven activities is poorly captured in economic statistics and often insufficiently appreciated by organisations and individuals. Estimates by Mandel (2012) suggest, for example, that data-driven activities in the United States are underestimated in official economic statistics, with real GDP in the first half of 2012 rising by 2.3% rather than the official rate of 1.7%.

In the case of personal data, collection directly from individuals is often a non-explicit exchange for "free" services. The ability to combine and recombine varied data sets enables uses that were not anticipated when the data were collected, making valuation difficult for national statistics as well as for organisations and individuals. A further measurement challenge is related to the complexity of current data flows, including across borders, and the assessment of value created through the analytic techniques themselves.

Conclusion

There is already some evidence of the potential benefits of using data as a resource for new industries, processes and products and therefore for innovation and growth. The large-scale and comprehensive developments affecting all stages of the data value chain presented in this chapter underline the need to take a closer look at data as an intangible asset and a new source of growth.

However, this paper also describes issues that deserve more work in order to understand better the potential and challenges of big data. One is evaluation of the socioeconomic impact of data across the economy and another is the contribution of data to GDP growth. OECD (2012a) discusses the challenges of measuring the monetary value and impacts of personal data. In fact, the value of data of all sorts is poorly captured in economic statistics and financial reports and often insufficiently appreciated by organisations and individuals. The fact that the value of data is context-dependent shows the need for the case studies to be undertaken as part of the OECD's follow-up work on big data.

This paper has looked at important policy areas that should be addressed. A number of OECD instruments referred to here are currently under review (Privacy Guidelines, Security Guidelines, and the PSI Recommendation). The OECD will assess other areas of policy relevant to big data in greater depth during 2013 and 2014. These include the employment impact of data-driven automation, issues related to competition, and intellectual property rights.

NOTES

[1] The openness principle of the Recommendation highlights that government should: "maximis[e] the availability of public sector information for use and re-use based upon presumption of openness as the default rule to facilitate access and re-use"; "develop… a regime of access principles or assuming openness in public sector information as a default rule, wherever possible no matter what the model of funding is for the development and maintenance of the information", and "defin[e] grounds of refusal or limitations, such as for protection of national security interests, personal privacy, preservation of private interests for example where protected by copyright, or the application of national access legislation and rules".

[2] Adopted from OECD (2011b), "Terms of Reference for Ensuring the Continued Relevance of the OECD Framework for Privacy and Transborder Flows of Personal Data".

[3] The fundamental rights of freedom of speech, freedom of the press and the need for open and transparent government should be considered.

[4] This would be an average yearly decrease of 38% in the cost of shifting one bit per second.

[5] See www.ted.com/talks/harald_haas_wireless_data_from_every_light_bulb.html.

[6] The number of mobile wireless devices connected to the Internet across the globe is estimated to reach 50 billion by 2020 (OECD, 2011b).

[7] The McKinsey Global Institute (MGI, 2011) estimates that the number of connected smart devices based on M2M will increase by more than 30% between 2010 and 2015 with the number of mobile-connected devices exceeding the world's population in 2012 (Cisco, 2012).

[8] This trend is confirmed by available sales figures. According to the Semiconductor Industry Association for instance, sensors and actuators are the fastest-growing semiconductor segment with growth in revenue of almost 16% (USD 8 billion) in 2011.

[9] Big data solutions are typically provided in three forms: software-only, as a software-hardware appliance or cloud-based (Dumbill, 2012a). Choices among these will depend, among other things, on issues related to data locality, human resources, and privacy and other regulations. Hybrid solutions (*e.g.* using on-demand cloud resources to supplement in-house deployments) are also frequent.

[10] Due to economies of scale, cloud computing providers have much lower operating costs than companies running their own IT infrastructure, which they can pass on to their customers.

[11] In 2009, Amazon introduced the Amazon Elastic MapReduce as a service to run Hadoop clusters on top of the Amazon S3 file system and Amazon Elastic Compute Cloud (EC2) (Amazon, 2009).

[12] In 2010, Borthakur (2010) claimed that Facebook had stored 21 petabytes (million gigabytes) of data using the largest Hadoop cluster in the world. One year later, Facebook announced that the data had grown by 42% to 30 petabytes (Yang, 2011).

[13] LinkedIn (2009) is using Hadoop together with Voldemort, another distributed data storage engine.

[14] IBM is offering its Hadoop solution through InfoSphere BigInsights. BigInsights augments Hadoop with a variety of features, including textual analysis tools that help identify entities such as people, addresses and telephone numbers (Dumbill, 2012b).

[15] Oracle provides its Big Data Appliance as a combination of open source and proprietary solutions for enterprises' big data requirements (Oracle, 2012). It includes, among others, the Oracle Big Data Connectors to allow customers to use Oracle's data warehouse and analytics technologies together with Hadoop, the Oracle R Connector to allow the use of Hadoop with R, an open-source environment for statistical analysis, and the Oracle NoSQL Database, which is based on Oracle Berkeley DB, a high-performance embedded database.

[16] From 2011, Microsoft started integrating Hadoop in Windows Azure, Microsoft's cloud computing platform, and one year later in Microsoft Server. It is providing Hadoop Connectors to integrate Hadoop with Microsoft's SQL Server and Parallel Data Warehouse (Microsoft, 2011).

[17] In 2012, SAP announced its roadmap to integrate Hadoop with its real-time data platform SAP HANA and SAP Sybase IQ (SAP, 2012).

[18] Specialised business-to-business companies include firms such as LexisNexis, which offers a complete background check of all possible business-related information about potential business partners. Regular data brokers such as Intelius and Locate Plus provide information solutions for consumers and small businesses using public records and publicly available information. Their services help people find each other, verify the identities of individuals they encounter, manage risk and ensure personal safety, to name a few. Finally localisation services such as LocatePeople.org, MelissaData.com, and 123people.com provide personal addresses of individuals for data marketers, or offer simple services to localise people, their telephone numbers, e-mail addresses, etc.

[19] See also Dumbill (2012a), for which "big data" is "data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it".

[20] See Watters (2012) for a comparison of Yahoo! and Google in terms of structured *vs.* unstructured data.

[21] See http://marketshare.hitslink.com/search-engine-market-share.aspx?qprid=4.

[22] This definition originated from the META Group (now part of Gartner) in 2001 (see Laney, 2001).

[23] According to Gartner (2012), the worldwide market for BI, analytic applications and performance management (PM) software grew by more than 16% in 2012 (from USD 12 million in 2011 to USD 16 million in 2012). The top five vendors (SAP, Oracle, SAS Institute, IBM, and Microsoft) account for close to three-quarters of the market.

[24] National statistics that provide occupational figures on data management and analytics professionals are a promising source for assessing data intensity not only by sector but also over time. This is only true if occupations related to data management and analytics can be identified in the occupation classification schemes.

[25] In 2011, financial activities, professional and business services, information, and public administration were the sectors mainly contributing to the increase in share of database administrators in the United States.

[26] According to data published by the World Information Technology and Services Alliance (WITSA), telecommunications (11.5%), financial services (6.6%), transport (5.1%), health care (4.1%) and

government (3.8%) are the five most ICT-intensive sectors. Using ICT intensity as a proxy for data intensity assumes that data-intensive industries have higher ICT expenditure than industries with low data intensity. However, this assumption can be easily challenged, since data analytics require less investment in ICTs today (because of cloud computing). In a historical perspective, this approach can still be useful.

27 OECD (2012d) work on "Understanding the Economics of Personal Data", which surveyed methodologies for measuring the monetary value, highlighted the context dependency of the monetary value of personal data.

28 In other cases, they could be tied to specific data sets (*e.g.* social networking or click-stream data with specific uses).

29 Countries include Austria, Germany, Denmark, Finland, France, Hungary, Italy, Korea, the Netherlands and Slovenia.

30 Adapted from Tucker (2010).

31 Web-bugs are 1x1-pixel pieces of code that allow advertisers to track customers remotely. These are also sometimes referred to as beacons, action tags, clear GIFs, web tags, or pixel tags (Gilbert, 2008). Web-bugs are different from cookies, because they are designed to be invisible to the user and are not stored on the user's computer. With web-bugs, a customer cannot know whether they are being tracked without inspecting a webpage's underlying html code.

32 A cookie is simply a string of text stored by a user's web browser. Cookies allow firms to track customers' progress across browsing sessions. This can also be done using a user IP address, but cookies are generally more precise, especially when IP addresses are dynamic as in the case of many residential Internet services. Advertisers may also use a flash cookie as an alternative to a regular cookie. A flash cookie differs from a regular cookie in that it is saved as a Local Shared Object on an individual's computer, making it harder for users to delete using regular tools on their browser.

33 A/B Testing is a method used to test the effectiveness of strategies/future actions based on a sample that is split in two groups, an A-group and a B-group. While an existing strategy is applied to the (larger) A-group, another, slightly changed strategy is applied to the other group. The outcome of both strategies is measured to determine whether the change in strategy led to statistically relevant improvements. Google, for example, regularly redirects a small fraction of its users to pages with slightly modified interfaces or search results to (A/B) test their reactions. For more detail see Christian (2012).

34 For example, the online payment platform WePay designed its entire website through a testing process. For two months, users were randomly assigned a testing homepage, and at the end the homepage with the best outcome was selected (Christian, 2012).

35 This value does not include potential costs to consumers that may occur due to privacy violations, for example.

36 The public sector in the United States employed on average 1.6 database administrators per 1 000 employees in 2011.

37 Many of these potential benefits rely on personal data, obtained not only from third parties but also directly from individuals, for administering various programmes. Examples include various social service programmes, tax programmes or issuing licences. Some data are also commonly used to support hundreds of regulatory regimes ranging from voter registration and political campaign contribution disclosures to verification of employee identity and enforcement of the child support obligation. Other uses include maintaining vital records about major lifecycle events, such as birth, marriage, divorce, adoption and death; and operation of facilities such as toll roads and national parks.

[38] It is necessary to exercise caution when interpreting these results as the methodologies used for these estimates are not necessarily explicit.

[39] At a recent OECD meeting, government technology leaders underscored that such new data sources have great potential to complement existing evidence across all policy domains and to unleash productivity in economic sectors with traditionally restricted productivity gains, but in which governments have historically had a significant impact, *e.g.* health, energy, education and government administration itself (OECD, 2012f).

[40] Reasons for not reporting include intimidation of victims and witnesses, but also lack of trust in local authorities.

[41] Examples of the "open data" movement include: the United States www.data.gov; the United Kingdom: www.data.gov.uk; and Spain: Aporta Web portal www.proyectoaporta.es.

[42] For example, government data about the financial industry was previously available only through the US Securities and Exchange Commission (SEC) and the US Financial Industry Regulatory Authority (FINRA). However, BrightScope has made such information more usable, searchable and open to the public, and individuals can therefore make better informed financial decisions (Howard, 2012).

[43] See forthcoming OECD work on mobile applications.

[44] UN Globalpulse introduced the concept of "data philanthropy", whereby the private sector shares data to support more timely and targeted policy action, and to highlight the public interest in shared data. In this context two ideas are debated: *i)* the "data commons" where some data are shared publicly after adequate anonymisation and aggregation; and *ii)* the "digital smoke signals" where sensitive data are analysed by companies but results are shared with governments.

[45] For example, at the OECD-APEC (2012) workshop, Anticipating the Needs of the 21st Century Silver Ageing Economy, held 12-14 September 2012 in Tokyo, Japan, participants concluded that the multi-factorial nature of Alzheimer's disease (AD) will require sophisticated computational capabilities to analyse big streams of behavioural, genetic, environmental, epigenetic and clinical data to find patterns. In neurodegenerative research, many organisations are building big data repositories and contributing to the development of databases and global data-sharing networks. In the United States alone, the Alzheimer's Disease Neuroimaging Initiative and the Parkinson's Disease (PD) Progression Markers Initiative gather brain images and biological fluids from people with or at risk for AD and PD, respectively. The US National Alzheimer's Coordinating Center has amassed longitudinal records from more than 25 000 people, and recently started assessments for fronto-temporal dementia as well. Records from those who inherited an AD-linked gene are part of the Dominantly Inherited Alzheimer Network.

[46] Adopted from OECD (2012a).

[47] In 2008, for example, around of 8% of electricity generated worldwide was lost before it reached the consumer. This is estimated to correspond to over 600 million tonnes of CO_2 emissions (OECD, 2012a). In the case of water distribution networks, estimates suggest that globally more than 32 billion cubic meters of treated water are lost annually through leakage (Kingdom *et al*., 2006).

[48] This is not without any risks to security and privacy as smart meters can be subject to cyber attacks and even data collected legally can give insights into an individual's private life, such as whether he or she was at home at a given time and even an indication of what they were doing.

[49] See www.youtube.com/watch?v=JnBoCq6vPwA.

[50] TomTom reported intangible assets worth EUR 872 million at the end of 2011, or almost 50% of its total assets (or 70% of total if one exclude goodwill).

[51] In January 2012, for example, Orange signed an agreement with Mediamobile, a leading provider of traffic information services in Europe, to use FMD data for its traffic information service V-Trafic (see www.traffictechnologytoday.com/news.php?NewsID=36182)

[52] The purpose specification principle states that "the purposes for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfilment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose".

[53] In 2011 in the United Kingdom, for example, the government launched a voluntary programme, Midata, with industry with a view to providing consumers with increased access to their personal data in a portable, electronic format (BIS, 2012).

[54] Fornefeld (2009) notes that in Germany parallel systems of private and public weather stations have been developed following the failure of negotiations on commercial reuse of PSI.

[55] See http://opendefinition.org/.

[56] Operation Aurora targeted data and intellectual property repositories of high-technology companies such as Google (2010), Adobe Systems, Juniper Networks, and Rackspace. According to McAfee (2010), the primary goal of Operation Aurora was to gain access to and potentially modify intellectual property repositories in high-technology firms. The attack involved social engineering techniques, the exploitation of a zero-day vulnerability (of a web browser) and the use of distributed C&C botnet servers (Zetter, 2010). Operation Aurora was estimated to have affected more than 34 organisations, including Yahoo!, Northrop Grumman, Dow Chemical and Rand Corp. (Damballa, 2010).

[57] See, for example, www.linkedin.com/skills/skill/Data_Science for the most frequent locations of people with "data science" in their skill profile. However, the high frequency of the United States could be due to the fact that the term "data science" is biased towards the United States.

REFERENCES

Acquisti, A., L. John and G. Loewenstein (2011), "What is Privacy Worth?", mimeo, http://pages.stern.nyu.edu/~bakos/wise/papers/wise2009-6a1_paper.pdf.

Amazon (2009), "Amazon Elastic MapReduce Developer Guide API", 30 November, http://s3.amazonaws.com/awsdocs/ElasticMapReduce/latest/emr-dg.pdf.

Anderson, C. (2008), "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", Wired Magazine, 23 June, www.wired.com/science/discoveries/magazine/16-07/pb_theory/.

Askitas N. and KN. Zimmermann (2010), "Google econometrics and unemployment forecasting", Technical report, SSRN 899, 2010, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1465341.

Autonomy (2012), "How to Leverage Big Data to Monetize Customer Experiences", White Paper, www.marketingpower.com/ResourceLibrary/Documents/Whitepapers/Autonomy%20Whitepaper%20Final%202.28.2012.pdf.

BIAC (2011), "BIAC Thought Starter: A Strategic Vision for OECD Work on Science, Technology and Industry", 12 October.

BIS: United Kingdom Department for Business Innovation & Skills (2012), "Next steps making midata a reality", News, 22 August, www.gov.uk/government/news/next-steps-making-midata-a-reality.

Bollier, D. (2010), "The Promise and Peril of Big Data", The Aspen Institute, Washington, DC.

Borthakur, D. (2010), "Facebook has the world's largest Hadoop cluster!", Hadoopblog, 9 May, http://hadoopblog.blogspot.fr/2010/05/facebook-has-worlds-largest-hadoop.html.

Brian, C. (2012), "The A/B Test: Inside the Technology That's Changing the Rules of Business", Wired.com, www.wired.com/business/2012/04/ff_abtesting/.

Brynjolfsson, E., L. M. Hitt and H. H. Kim (2011), "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?", 22 April, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486.

Brynjolfsson, E., A. McAfee (2011), "Race Against The Machine: How the Digital Revolution is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and the Economy", 17 October.

Bullas, J. (2011), "50 Fascinating Facebook Facts And Figures", jeffbullas.com, 28 April, www.jeffbullas.com/2011/04/28/50-fascinating-facebook-facts-and-figures.

Burdon, M. (2010), "Privacy Invasive Geo-Mashups: Privacy 2.0 and the Limits of First Generation Information Privacy Laws", *University of Illinois Journal of Law, Technology and Policy*, Vol. 1, pp. 1-50.

BusinessWire (2012), "Internet Advertising Revenues Set First Quarter Record at $8.4 Billion", 11 June, www.businesswire.com/news/home/20120611005230/en/Internet-Advertising-Revenues-Set-Quarter-Record-8.4.

Carrière-Swallow, Y. and F. Labbé (2010), "Nowcasting with Google Trends in an Emerging Market", *Central Bank of Chile Working Papers*, No. 588, July.

Cate, F.H. (2008), "Government Data Mining: The Need for a Legal Framework", *Harvard Civil Rights-Civil Liberties Law Review*, Vol. 43.

Cavoukian, A. and J. Jonas (2012), "Privacy by Design in the Age of Big Data", www.privacybydesign.ca.

Cebr (2012), "Data equity: Unlocking the value of big data", Report for SAS, April.

Chang, F., J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A. Fikes and R.E. Gruber (2006), "Bigtable: A Distributed Storage System for Structured Data", Google, appeared in: Seventh Symposium on Operating System Design and Implementation (OSDI'06), November, http://research.google.com/archive/bigtable.html.

Choi, H. and H. Varian (2009), "Predicting the Present with Google Trends", Discussion paper, Google, 10 April, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1659302.

Christian, B. (2012), "The A/B Test: Inside the Technology That's Changing the Rules of Business", *Wired*, 25 April, www.wired.com/business/2012/04/ff_abtesting.

Cisco (2012), "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2011–2016", White Paper, www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf.

Cukier, K. (2010), "META: The Rise and Governance of Information About Information", 2010 Global Leaders of Information Policy Conference, January, Singapore.

Damballa (2010), "The Command Structure of the Aurora Botnet: History, Patterns and Findings", Damballa , 3 March, www.damballa.com/downloads/r_pubs/Aurora_Botnet_Command_Structure.pdf.

D'Amuri, F. and J. Marcucci (2010), "Google it! Forecasting the US unemployment rate with a Google job search index", SSRN, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1594132.

Dean J. and S. Ghemawat (2004), "MapReduce: Simplified Data Processing on Large Clusters", in Sixth Symposium on Operating System Design and Implementation (OSDI'04), December, San Francisco, CA, http://research.google.com/archive/mapreduce.html.

DeCarlo, M. (2012), "Gartner: SSDs will reach mainstream prices in 2012", TechSpot, 11 May, available at www.techspot.com/news/43752-gartner-ssds-will-reach-mainstream-prices-in-2012.html.

DeCarlo, M. (2013), "Kaspersky uncovers five-year cyber espionage campaign, Red October", Techspot, 14 January, www.techspot.com/news/51332-kaspersky-uncovers-five-year-cyber-espionage-campaign-red-october.html.

Dumbill, E. (2010), "The SMAQ stack for big data: Storage, MapReduce and Query are ushering in data-driven products and services", O'Reilly Radar, 22 September, http://radar.oreilly.com/2010/09/the-smaq-stack-for-big-data.html.

Dumbill, E. (2012a), "What is big data? An introduction to the big data landscape", O'Reilly Radar, 11 January, http://radar.oreilly.com/2012/01/what-is-big-data.html.

Dumbill, E. (2012b), "Big data market survey: Hadoop solutions", O'Reilly Radar, 19 January, http://radar.oreilly.com/2012/01/big-data-ecosystem.html.

EC: European Commission (2006), Measuring European Public Sector Information Resources (MEPSIR), "Final report of study on exploitation of public sector information – benchmarking of EU framework conditions", Executive summary and Final report Part 1 and Part 2, Brussels.

EC: European Commission (2010), "Riding the Wave: How Europe can gain from the rising tide of scientific data", Final report by the High-level Expert Group on Scientific, October, http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf.

Esposito, R. (2012), "'Astonishing' Cyber Espionage Threat from Foreign Governments: British Spy Chief", ABC News, 25 June, http://abcnews.go.com/Blotter/astonishing-cyberespionage-threat-foreign-governments-british-spy-chief/story?id=16645690.

Fornefeld, M. (2009), "The Value to Industry of PSI: The Business Sector Perspective", Chapter 4, in: OECD, NRC (2009), *The Socioeconomic Effects of Public Sector Information on Digital Networks: Toward a Better Understanding of Different Access and Reuse Policies*, Workshop Summary, available at: www.nap.edu/openbook.php?record_id=12687&page=10.

Gartner (2011), "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data", Press release, www.gartner.com/it/page.jsp?id=1731916.

Gartner (2012), "Gartner Says Worldwide Business Intelligence, Analytics and Performance Management Software Market Surpassed the $12 Billion Mark in 2011", Press release, www.gartner.com/it/page.jsp?id=1971516.

GE (2007), "What is the real potential of the Smart Grid?", GE Energy Presentation, Autovation 2007, The AMRA International Symposium, 30 September-3 October, www.ge-energy.com/prod_serv/plants_td/en/downloads/real_potential_grid.pdf.

Gentile, B. (2011), "The New Factors of Production and the Rise of Data-Driven Applications", *Forbes*, 31 October, www.forbes.com/sites/ciocentral/2011/10/31/the-new-factors-of-production-and-the-rise-of-data-driven-applications/.

GeSI (2008), "SMART 2020: Enabling the low carbon economy in the information age", 23 June, www.gesi.org/LinkClick.aspx?fileticket=tbp5WRTHUoY%3d&tabid=60.

Gilbert, F. (2008), "Beacons, Bugs, and Pixel Tags: Do You Comply with the FTC Behavioral Marketing Principles and Foreign Law Requirements?", *Journal of Internet Law*, May.

Ginsberg, J., M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski and L. Brilliant (2009), "Detecting influenza epidemics using search engine query data". *Nature*, p. 1012-14 http://research.google.com/archive/papers/detecting-influenza-epidemics.pdf.

Goodin, D. (2011), "Sony says data for 25 million more customers stolen", *The Register*, 03 May, www.theregister.co.uk/2011/05/03/sony_hack_exposes_more_customers.

Google (2010), "A new approach to China", Google Official Blog, 13 January, http://googleblog.blogspot.fr/2010/01/new-approach-to-china.html.

Groenfeldt (2012), "Morgan Stanley Takes On Big Data With Hadoop", Forbes, 30 May, www.forbes.com/sites/tomgroenfeldt/2012/05/30/morgan-stanley-takes-on-big-data-with-hadoop/.

Hachman, M. (2012), "Facebook Now Totals 901 Million Users, Profits Slip", PC Magazine, 23 April, available at: www.pcmag.com/article2/0,2817,2403410,00.asp.

Harris, D. (2011), "Hadoop Kills Zombies Too! Is There Anything It Can't Solve?", Gigaom, 18 April, http://gigaom.com/cloud/hadoop-kills-zombies-too-is-there-anything-it-cant-solve/.

Hildebrandt, M. and B.J Koops (2010), "The Challenges of Ambient Law and Legal Protection in the Profiling Era", *The Modern Law Review*, Vol. 73, No. 3, pp. 428-460.

Hill, K. (2012), "How Target Figured Out a Teen Girl Was Pregnant Before Her Father Did", *Forbes*, 16 February, www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/.

Houghton, J., B. Rasmussen and P. Sheehan (2010), "Economic and Social Returns on Investment in Open Archiving Publicly Funded Research Outputs", Report to SPARC, July, Centre for Strategic Economic Studies, Victoria University, available at: www.arl.org/sparc/bm~doc/vufrpaa.pdf.

Howard, A. (2012), "Data for the Public Good", 17 February, O'Reilly Media.

IDC (2012), "The Digital Universe in 2020: Big Data, Bigger Digi tal Shadows, and Biggest Growth in the Far East", December, available at: www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf.

IEEE (2009), "Smart Grid Communications Preliminary Proposal to IEEE P802.15 Working Group for Wireless Personal Area Networks", 1 March.

ITU (2012), "Key ICT indicators for developed and developing countries and the world (totals and penetration rates)", www.itu.int/ITU-D/ict/statistics/at_glance/KeyTelecom.html , last accessed 7 December 2012.

Jones, S. (2012), "Why 'Big Data' is the fourth factor of production", *Financial Times*, 27 December, www.ft.com/intl/cms/s/0/5086d700-504a-11e2-9b66-00144feab49a.html.

Kingdom, B., R. Liemberger and P. Marin (2006), "The Challenge of Reducing Non-Revenue Water (NRW) in Developing Countries: How the Private Sector Can Help", *Water Supply and Sanitation Sector Board Discussion Paper Series*, Paper No. 8, December.

Kroll (2012), *Global Fraud Report: Economist Intelligence Unit Survey Results*, Annual Edition 2011/12, October, www.krolladvisory.com/media/pdfs/KRL_FraudReport2011-12_US.pdf.

Laney, D. (2001), "3D Data Management: Controlling Data Volume, Velocity, and Variety", META Group, 6 February, http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

LinkedIn (2009), "Building a terabyte-scale data cycle at LinkedIn with Hadoop and Project Voldemort", SNA Project Blog, LinkedIn's Search Network and Analytics team, 16 June, http://project-voldemort.com/blog/2009/06/building-a-1-tb-data-cycle-at-linkedin-with-hadoop-and-project-voldemort/.

Lohr, S. (2009), "For Today's Graduate, Just One Word: Statistics", *The New York Times*, 5 August, www.nytimes.com/2009/08/06/technology/06stats.html.

Loukides, M. (2010), "What is data science? The future belongs to the companies and people that turn data into products", O'Reilly Radar, 2 June, http://radar.oreilly.com/2010/06/what-is-data-science.html.

Mandel, M. (2012), "Beyond Goods and Services: The (Unmeasured) Rise of the Data-Driven Economy", Progressive Policy Institute, October, www.progressivepolicy.org/wp-content/uploads/2012/10/10.2012-Mandel_Beyon d-Goods-and-Services_The-Unmeasured-Rise-of-the-Data-Driven-Economy.pdf.

Masters, B. and J. Menn (2010), "Data theft overtakes physical losses", *Financial Times*, 18 October, www.ft.com/intl/cms/s/2/3c0c9998-da1a-11df-bdd7-00144feabdc0.html.

McAfee (2010), "Protecting Your Critical Assets: Lessons Learned from 'Operation Aurora'", White Paper, www.wired.com/images_blogs/threatlevel/2010/03/operationaurora_wp_0310_fnl.pdf.

McAfee (2011), "Revealed: Operation Shady RAT", 8 August, www.mcafee.com/us/resources/white-papers/wp-operation-shady-rat.pdf.

McGuire, T., J. Manyika and M. Chui (2012), "Why Big Data is the New Competitive Advantage", *Ivey Business Journal*, July/August, www.iveybusinessjournal.com/topics/strategy/why-big-data-is-the-new-competitive-advantage.

McKinsey (2010), "Consumers driving the digital uptake", Technical report, McKinsey and Company and IAB Europe, September.

MGI: McKinsey Global Institute (2011), "Big data: The next frontier for innovation, competition and productivity", McKinsey & Company, June, www.mckinsey.com/~/media/McKinsey/dotcom/Insights%20and%20pubs/MGI/Research/Technology%20and%20Innovation/Big%20Data/MGI_big_data_full_report.ashx.

Microsoft (2011), "Microsoft Expands Data Platform With SQL Server 2012, New Investments for Managing Any Data, Any Size, Anywhere", Microsoft News Center, 12 October, www.microsoft.com/en-us/news/press/2011/oct11/10-12PASS1PR.aspx.

Muenchen, R. (2012), "The Popularity of Data Analysis Software", r4stats.com, http://r4stats.com/articles/popularity/, last accessed: 28 August 2012.

Muthukkaruppan, K. (2010), "The Underlying Technology of Messages", Notes, Facebook, 15 November, www.facebook.com/notes/facebook-engineering/the-underlying-technology-of-messages/454991608919.

Narayanan A. and V. Shmatikov (2010), "Privacy and Security Myths and Fallacies of 'Personally Identifiable Information'", *Communications of the ACM*, Vol. 53 (6).

NCIX: National Counterintelligence Executive (United States) (2012), *Foreign Spies Stealing US Economic Secrets in Cyberspace: Report to Congress on Foreign Economic Collection and Industrial Espionage, 2009-2011*, NCIX, October, www.ncix.gov/publications/reports/fecie_all/Foreign_Economic_Collection_2011.pdf.

OECD (1980), *Recommendation of the Council on Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data*, 23 September, OECD, Paris.

OECD (2002), *OECD Guidelines for the Security of Information Systems and Networks: Towards a Culture of Security*, OECD Publishing. doi: 10.1787/9789264059177-en-fr

OECD (2007), *OECD Principles and Guidelines for Access to Research Data from Public Funding*, OECD Publishing. doi: 10.1787/9789264034020-en-fr

OECD (2008), *OECD Recommendation for Enhanced Access and More Effective Use of Public Sector Information*, 16 June, C(2008)36, available at: www.oecd.org/internet/ieconomy/40826024.pdf.

OECD (2009), "Smart Sensor Networks: Technologies and Applications for Green Growth", *OECD Digital Economy Papers*, No. 167, OECD Publishing. doi: 10.1787/5kml6x0m5vkh-en

OECD (2009b), "Network Developments in Support of Innovation and User Needs", *OECD Digital Economy Papers*, No. 164, OECD Publishing. doi: 10.1787/5kml8rfvtbf6-en

OECD (2011a), *OECD Skills Strategy: Towards an OECD Skills Strategy*, September, OECD, Paris, www.oecd.org/edu/47769000.pdf.

OECD (2011b), "Terms of Reference for Ensuring the Continued Relevance of the OECD Framework for Privacy and Transborder Flows of Personal Data", DSTI/ICCP/REG(2011)4/FINAL, www.oecd.org/sti/interneteconomy/48975226.pdf.

OECD (2011a), "Measuring the Economics of 'Big Data'", OECD internal working document.

OECD (2012), "ICT Applications for the Smart Grid: Opportunities and Policy Implications", *OECD Digital Economy Papers*, No. 190, OECD Publishing. doi: 10.1787/5k9h2q8v9bln-en.

OECD (2012), "Machine-to-Machine Communications: Connecting Billions of Devices", *OECD Digital Economy Papers*, No. 192, OECD Publishing. doi: 10.1787/5k9gsh2gp043-en.

OECD (2012c), "Cloud Computing: The Concept, Impacts and the Role of Government Policy", internal working document.

OECD (2012d), "Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value", *OECD Digital Economy Papers*, No. 220, OECD Publishing. http://dx.doi.org/10.1787/5k486qtxldmq-en.

OECD (2012e), "OECD E-Government Project: The Role of New Technologies for Strategic and Agile Public Governance", 20 March, internal working document.

OECD (2012f), "Summary report of the E-Leaders Meeting 2012: New ICT Solutions for Public Sector Agility", 26-27 March, Mexico City, available at: http://www.oecd.org/governance/eleaders/50251374.pdf.

OECD (2012g), "Big Data and Statistics: Understanding the Proliferation of Data and Implications for Official Statistics and Statistical Agencies", internal working document

OECD (2012h), "Joint-Consultation of the OECD Health Care Quality Indicator Expert Group and the Working Party on Information Security and Privacy", internal working document.

OECD (2012), "ICT Skills and Employment: New Competences and Jobs for a Greener and Smarter Economy", *OECD Digital Economy Papers*, No. 198, OECD Publishing. doi: 10.1787/5k994f3prlr5-en.

OECD (2012), *Better Skills, Better Jobs, Better Lives: A Strategic Approach to Skills Policies*, OECD Publishing. doi: 10.1787/9789264177338-en.

OECD (2013), "Building Blocks for Smart Networks", *OECD Digital Economy Papers*, No. 215, OECD Publishing. doi: 10.1787/5k4dkhvnzv35-en

OECD-APEC (2012), Workshop on Anticipating the Special Needs of the 21st Century Silver Economy: From Smart Technologies to Services Innovation: From Smart Technologies to Services Innovation, Main Conclusions And Options For Future Work, internal working document.

OECD-NSF (2011), "OECD-NSF Workshop: Building a Smarter Health and Wellness Future", Summary of Key Messages, 15-16 February, available at: http://www.oecd.org/internet/ieconomy/48915787.pdf.

Ohm, P. (2010), "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization", *UCLA Law Review*, Vol. 57, p. 1701-1777.

Oracle (2012), "Oracle: Big Data for the Enterprise", White Paper, January, www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf.

OSTP (2010), "Blue Ribbon Task Force on Sustainable Digital Preservation and Access, Sustainable Economics for a Digital Planet: Ensuring Long Term Access to Digital Information", February, http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.

Pingdom, R. (2011), "Would you pay $7,260 for a 3 TB drive? Charting HDD and SSD prices over time", 19 December, http://royal.pingdom.com/2011/12/19/would-you-pay-7260-for-a-3-tb-drive-charting-hdd-and-ssd-prices-over-time.

Polgreen, P. M., Y. Chen, D. M. Pennock and F.D. Nelson (2009), "Using internet searches for influenza surveillance", *Clinical Infectious Diseases*, Vol. 47, p.1443-1448, www.ncbi.nlm.nih.gov/pubmed/18954267.

Reuters (2011), "UPDATE 2-Sony breach could cost card lenders $300 mln", 28 April, www.reuters.com/article/2011/04/28/sony-creditcards-cost-idUSN2826485220110428.

Rogers, S. (2012), "What is a data scientist?", Guardian Datablog, 2 March, www.guardian.co.uk/news/datablog/2012/mar/02/data-scientist.

Rudder, C. (2010), "The 4 Big Myths of Profile Pictures", 20 January, http://blog.okcupid.com/index.php/the-4-big-myths-of-profile-pictures/.

Russom, P. (2007), "Bi Search and Text Analytics: New Additions to the BI Technology Stack", *tdwi Best Practices Report*, tdwi, Second Quarter 2007.

SAP (2012), "SAP Unveils Unified Strategy for Real-Time Data Management to Grow Database Market Leadership", Press release, 10 April, www.sap.com/canada/about/press/press.epx?pressid=18621.

Severs, H. (2013), "The Greatest Transfer of Wealth in History: how significant is the cyber-espionage threat?", theriskyshift.com, Essays, 17 January, http://theriskyshift.com/2013/01/cyber-espionage-the-greatest-transfer-of-wealth-in-history/.

Seybold, P. (2011), "Update on PlayStation Network and Qriocity", PlayStation.Blog, 26 April, http://blog.us.platstation.com/2011/04/26/update-on-playstation-network-and- qriocity.

Shapiro, C. and H. Varian (1998), *Information Rules: A Strategic Guide to the Network Economy*, 1st Edition, 19 November, *Harvard Business Review Press*, Boston, MA.

**Shilakes, C. and J. Tylman (1998), "Enterprise Information Portals: Move Over Yahoo!; the Enterprise Information Portal Is on Its Way", Merrill Lynch, 16 November.

Spiekermann, S., J. Grossklags and B. Berendt (2001), "E-privacy in 2nd Generation E-Commerce: Privacy Preferences Versus Actual Behavior", *Proceedings of the ACM Conference on Electronic Commerce*, pp. 38-47.

Stiglitz, J., P. Orszag and J. Orszag (2000), "Role of Government in a Digital Age", Computer and Communications Industry Association, October, www.ccianet.org/CCIA/files/ccLibraryFiles/Filename/000000000086/govtcomp_report.pdf.

Suhoy, T. (2009), "Query indices and a 2008 downturn: Israeli data", Technical Report, Bank of Israel, 2009, www.bankisrael.gov.il/deptdata/mehkar/papers/dp0906e.pdf.

Surowiecki, J. (2011), "A Billion Prices Now", *The New Yorker*, 30 May, p. 28.

Tene, O. and J. Polonetsky (2012), "Privacy in the Age of Big Data: A Time for Big Decisions", *Stanford Law Review Online*, Vol. 64, Symposium Issue, pp. 63-69.

The Economist (2010), "Data, data everywhere" 27 February.

Tucker, C. (2010), "The Economic Value of Online Customer Data", Background Paper #1, Joint WPISP-WPIE Roundtable "The Economics of Personal Data and Privacy: 30 Years after the OECD Privacy Guidelines", 1 December, www.oecd.org/sti/interneteconomy/theeconomicsofpersonaldataandprivacy30yearsaftertheoecdprivacyguidelines.htm#Background_Reports.

Uhlir, P. F. (2009), "The Socioeconomic Effects of Public Sector Information on Digital Networks: Towards Better Understanding of Different Access and Reuse Policies", Workshop Summary, National Academy of Sciences, Washington DC.

United Nation [UN] Global Pulse (2012), "Big Data for Development: Opportunities & Challenges", Global Pulse White Paper, May, available at: www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf.

Valdivia A. and S. Monge-Corella (2010), "Diseases tracked by using Google Trends, Spain", *Emerging Infectious Diseases*, January, www.cdc.gov/EID/content/16/1/168.htm.

Vickery, G. (2012), "Review of Recent Studies on PSI Re-Use and Related Market Developments", *Information Economics*, May, Paris, available at: http://ec.europa.eu/information_society/policy/psi/docs/pdfs/report/final_version_study_psi.docx.

Villars, R, C. Olofson and M. Eastwood (2012), "Big Data: What It Is and Why You Should Care", White Paper, IDC.

Voreacos, D. (2009), "Hacker Agrees to Plead Guilty in Second Computer Data Theft", Bloomberg, 9 December, www.bloomberg.com/apps/news?pid=newsarchive&sid=aE0.8o_7QcGc.

Watters, A. (2012), "Embracing the chaos of data", O'Reilly Radar, 31 January, http://radar.oreilly.com/2012/01/unstructured-data-chaos.html.

WSJ (2010), "What They Know: A Glossary", *Wall Street Journal*, 31 July, http://online.wsj.com/article/SB10001424052748703999304575399492916963232.html.

Yang, P. (2010), "Moving an Elephant: Large Scale Hadoop Data Migration at Facebook", Facebook, 27 July, www.facebook.com/notes/paul-yang/moving-an-elephant-large-scale-hadoop-data-migration-at-facebook/10150246275318920.

Zetter, K. (2009), "TJX Hacker Charged With Heartland, Hannaford Breaches", Wired.com, 17 August, www.wired.com/threatlevel/2009/08/tjx-hacker-charged-with-heartland.

Zetter, K. (2010), "'Google' Hackers Had Ability to Alter Source Code", Wired.com, 3 March, www.wired.com/threatlevel/2010/03/source-code-hacks/.

Zinow, R. (2012), "Big Data, Mobile and Cloud combined – the new paradigm shift?", SAP, Presentation at the ECD Conference, 23 May, http://ecd-conference.de/wp-content/blogs.dir/46/files/2011/03/Zinow_SAP_1545_2.pdf.