Chapter 3

How data now drive innovation

This chapter highlights the key drivers of data-driven innovation (DDI), today a widespread socio-economic phenomenon. It documents the key trends leading to the adoption of data and analytics across the economy, which are related to i) data generation and collection, ii) data processing and analysis, and iii) data-driven decision making. It also shows how the confluence of these trends is leading to the "industrialisation" of knowledge creation and a paradigm shift in decision making towards decision automation. The chapter then highlights the limitations of data-driven decision making, and concludes with a discussion of the key policy implications.

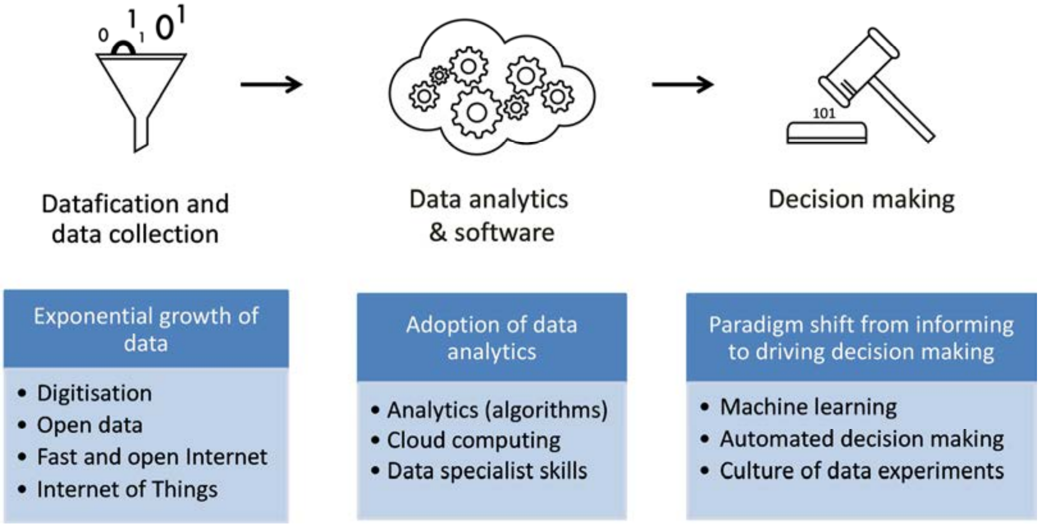
> *So how do we spot the future—and how might you? The seven rules that follow are not a bad place to start: […] 1. Look for cross-pollinators. […] 2. Surf the exponentials. […] 3. Favor the liberators. […] 4. Give points for audacity. […] 5. Bank on openness. […] 6. Demand deep design. […] 7. Spend time with time wasters. […]*. (Goetz, 2012)

The statistical data for Israel are supplied by and under the responsibility of the relevant Israeli authorities. The use of such data by the OECD is without prejudice to the status of the Golan Heights, East Jerusalem and Israeli settlements in the West Bank under the terms of international law.

Data are an increasingly significant resource that can drive value creation and foster new industries, processes and products – data-driven innovation, or DDI. The importance of data however, both economically and socially, is not new. Many activities have long revolved round the analysis and use of data. Before the digital revolution, data were already used for scientific discovery and for monitoring business activities such as through accounting. There is also evidence that they were systematically collected and used in early history – for instance, as a means to keep information about the members of a given population (i.e. census).[1] In business, furthermore, concepts such as "business intelligence" (Luhn, 1958)[2] and "data warehousing" (Keen, 1978; Sol, 1987) emerged in the 1960s and became popular in the late 1980s when computers were increasingly used as decision support systems (DSSs). The financial sector is a popular example of the longstanding use of DSSs for (e.g.) detecting fraud and assessing credit risks (Inmon and Kelley, 1992).

That said, a confluence of three major socio-economic and technological trends is making DDI a new source of growth today: i) the exponential growth in data generated and collected, ii) the pervasive power of data analytics, and iii) the emergence of a paradigm shift in knowledge creation and decision making. These three trends are developing along the data value cycle introduced in Chapter 1 of this volume (Figure 1.**7**). Their confluence along the data value cycle has enabled the exploitation of data in ways never before possible. These three major trends are discussed further below, with the focus on key enabling factors, as illustrated in Figure 3.1.

Figure 3.1. **DDI: The data value cycle and confluence of key trends and enabling factors**


0 1 1 0 1
101
Datafication and data collection
Data analytics & software
Decision making
Exponential growth of data
• Digitisation
• Open data
• Fast and open Internet
• Internet of Things
Adoption of data analytics
• Analytics (algorithms)
• Cloud computing
• Data specialist skills
Paradigm shift from informing to driving decision making
• Machine learning
• Automated decision making
• Culture of data experiments

Note: Data specialist skills, key to adopting data analytics, are discussed in Chapter 6.

Understanding the key trends and enabling factors of DDI is crucial for governments to assess their economies' readiness to take advantage of this new source of growth. Economies in which these trends and factors are more prevalent are expected to be in a better position to benefit from DDI, although that does not mean that all factors need to be fully developed in order to realise the benefits. The global nature of the data ecosystem allows countries to benefit from DDI through data and analytics-related goods and services produced elsewhere as discussed in Chapter 2 of this volume in more detail. However, it can be assumed that countries with enhanced capacities to both supply *and* use data and analytics will be in the best position to reap the fruits of DDI: A well-functioning supply side is a precondition for the development of a thriving data ecosystem, while a well-functioning demand side enables data-driven entrepreneurs to use data and analytics to innovate goods and services across the economy (see Chapter 2).

3.1. The exponential growth in data generated and collected

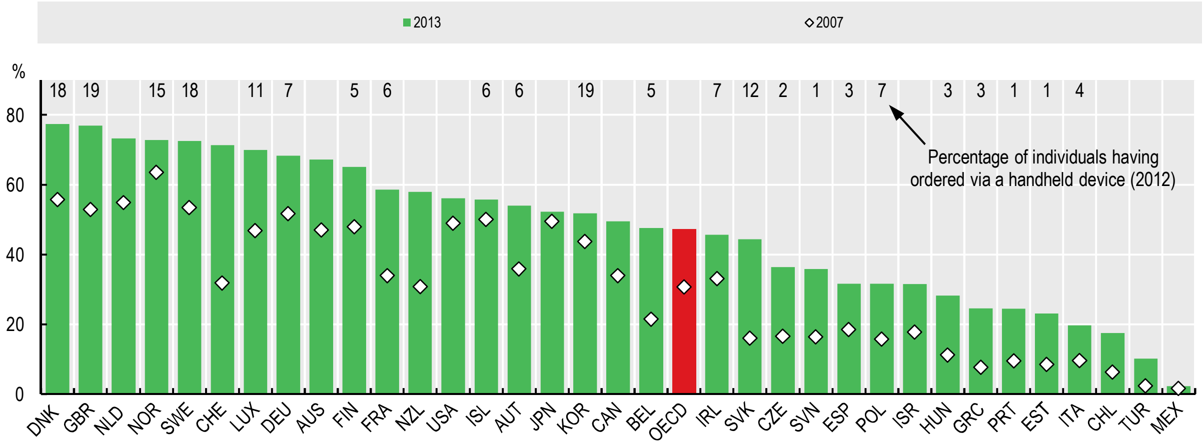
The first major trend driving DDI is the sheer growth in data volume. Measurement of the real total data generated, collected and stored is still speculative, but one source suggests that in 2010 alone, enterprises overall stored more than seven exabytes of new data on disk drives, while consumers stored more than six exabytes of new data (MGI, 2011). As a point of reference, one exabyte corresponds to one billion gigabytes and is equivalent, for example, to around 50 000 years of DVD-quality video. This growing storage has led to an estimated cumulative data volume of more than 1 000 exabytes in 2010, and some estimates suggest that that figure will multiply by a factor of 40 by the end of this decade, given the emerging Internet of Things (see section below) (IDC, 2012a).

The digitisation of nearly all media and the increasing migration of social and economic activities to the Internet (through Internet-based services such as social networks, e-commerce, e-health and e-government) have been two of the most important developments leading to the generation of unprecedented volumes of digital data across all sectors of the economy and in all areas of social life. In 2013, about half of the population in OECD countries, for example, had already purchased goods and services on line – thereby generating data that are increasingly used for personalised marketing, including product recommendation and personalisation (Figure 3.2).

In addition, the increasing deployment of connected devices through mobile and fixed networks captures an ever growing number of (offline) activities in the physical world. This process of transforming the world into processable and quantifiable data is sometimes referred to as “datafication”, a portmanteau for “data” and “quantification” (Hey, 2004; Bertolucci, 2013; Mayer-Schönberger and Cukier, 2013).[3] The datafication of offline activities is resulting in an additional tidal wave of data. In 2013, there were almost 7 billion mobile subscriptions worldwide, of which roughly 15% were smartphones capable of running mobile device applications (apps), which collect and transmit a wide array of sensor data (Cisco, 2013; ITU, 2014). As the number of smartphones continues to grow, as well as the number of apps installed on these devices, more data can be expected to be generated, even if all apps are not actively used. In 2013, smartphone users on average installed around 30 apps on their smartphones, most of which are collecting data related to (for instance) communication, locations and personal accounts (OECD, 2014).

Figure 3.2. **The diffusion of online purchases, 2013 and 2007**

Percentage of individuals ordering goods or services online



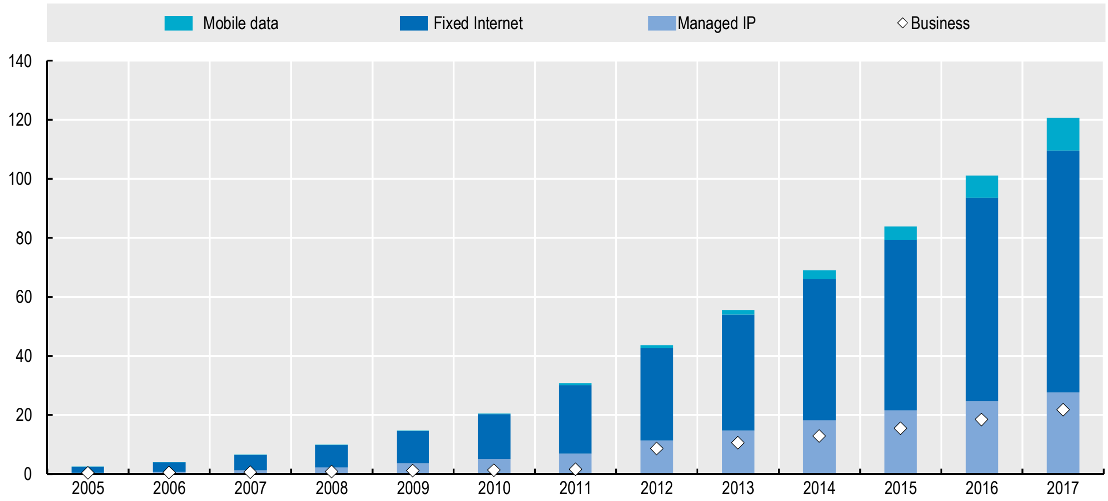
Note: For Australia, data refer to 2012/2013 (fiscal year ending in June 2013) instead of 2013. For 2007, data refer to 2006/2007 (fiscal year ending in June 2007), and to individuals aged 15 and over instead of 16-74 year-olds. For Canada, data refer to 2012 and relate to individuals who ordered goods or services over the Internet from any location (for personal or household use). For Chile, data refer to 2009 and 2012.For Japan, data refer to 2012 and to individuals aged 15-69 instead of 16-74 year-olds. For Israel, data refer to all individuals aged 20 and over who used the Internet for purchasing all types of goods or services. For Korea, the figure shows OECD estimates based on the Survey on the Internet Usage 2012. Data refer to the population aged 12 or more. In 2013, the share of individuals buying via handheld devices reached 35.5%. For New Zealand, data refer to 2006 and 2012 and relate to individuals who made a purchase through the Internet for personal use, which required an online payment. For Switzerland, data refer to 2005 instead of 2007. For the United States, data originate from May 2011 and September 2007 PEW Internet Surveys and cover individuals aged 18 or more.

Sources: OECD (2014), *Measuring the Digital Economy,* OECD Publishing, Paris, http://dx.doi.org/10.1787/888933148361, based on OECD ICT Database; Eurostat, Information Society Statistics and national sources, May 2014.

Apps have transformed smartphones into multi-purpose mobile devices that in 2013 have generated more than 1.5 exabytes (billions of gigabytes) of data every month worldwide. However, the growth in mobile data is not only driven by the use of smartphones or tablets, which are estimated to account for only half of total mobile traffic. An even faster-growing volume of data about (offline) activities in the physical world is being generated by what is called the Internet of Things (IoT): interconnected objects enabled by sensors and machine-to-machine communication (M2M). Overall, Cisco (2013) estimates that the amount of data traffic generated by all mobile devices will almost double every year, reaching more than 11 exabytes by 2017 (Figure 3.3). This "datafication" process will reach its tipping point once the volume of (fixed and mobile) M2M bypasses that of human data communication, signalling a new phase of DDI that today is only in its infancy even in the most advanced economies.

Figure 3.3. **Monthly global Internet Protocol (IP) data traffic, 2005-17**

In exabytes (billions of gigabytes)



Source: OECD based on Cisco (2013), "Cisco Visual Networking Index: Forecast and methodology", 2012-17, www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html (June 2014).

The following sections look at the key enabling factors for the exponential growth of data and the IoT. It should be pointed out that although the main enablers have been identified, further studies are needed to fully understand the social and economic effects and the policy implications of the IoT, which go far beyond the scope of this section. The key enabling factors for the exponential growth of data include:

- access to a fast and open Internet – enabling the free flow of data
- sensors and sensor networks – enabling the ubiquitous datafication of the physical world
- machine-to-machine communication – empowering data exchange in the Internet of Things.

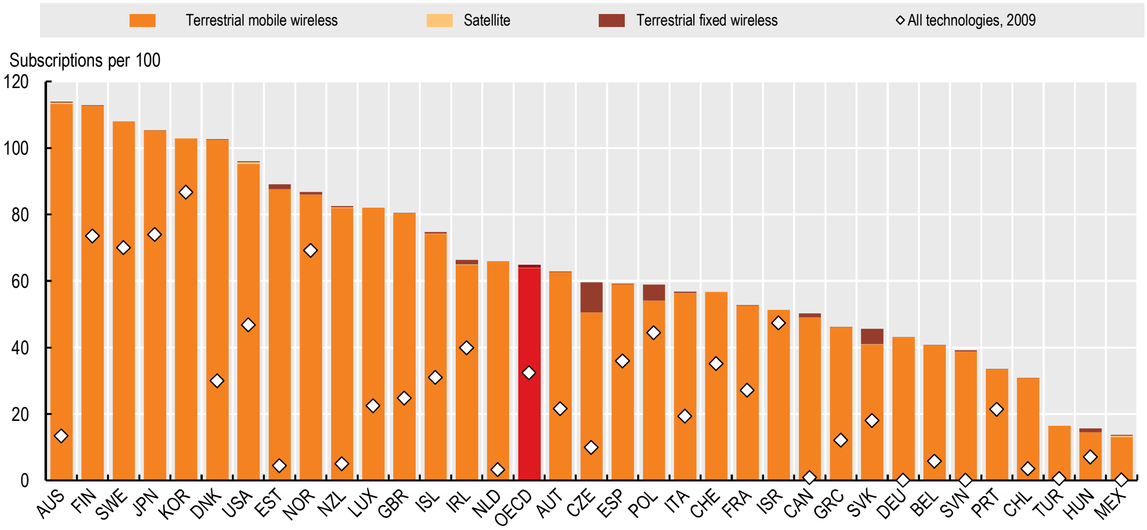
Access to a fast and open Internet – enabling the free flow of data

High-speed mobile broadband

The rapid diffusion of broadband is one of the most fundamental enablers of DDI. High-speed broadband is the underlying infrastructure for the exchange and free flow of data that are collected remotely through Internet applications and now increasingly through smart and interconnected devices. Where real-time applications are deployed, broadband networks enable timely data transmission (OECD, 2014a).[4] Mobile broadband in particular is essential, as mobile devices are becoming the leading means for data collection and dissemination. Moreover, high-speed mobile broadband is especially important to further improve connectivity in remote and less developed regions, where DDI could bring much needed (regional) growth and development (see Chapter 1). Within 10 years, between 2003 and 2013, fixed broadband penetration rates (subscribers per 100 inhabitants) in the OECD area have almost tripled, to reach around 30% of the OECD populations, but mobile broadband penetration rates have been more dynamic since surpassing fixed broadband penetration rates in 2008. Since then, mobile broadband penetration rates have more than doubled, currently reaching around 70% in the OECD area.

The lowering of mobile access prices is the prime factor behind the explosion of mobile subscriptions (OECD, 2014b). In Australia, Finland, Sweden, Japan, Korea, and Denmark mobile penetration rates exceeded 100% in 2013 (Figure 3.4). Australia, which edged into first place after a 13% surge in smartphone subscriptions in the first half of 2013 – as well as Estonia, New Zealand, the Netherlands, the Czech Republic, and Canada – have experienced a boost in mobile subscriptions since 2009. Penetration is still at 40% or less in Portugal, Greece, Chile, Turkey, Hungary and Mexico; however, considering progress to date and the universal diffusion of standard mobile subscriptions, mobile broadband could well catch up in lagging economies as well (OECD, 2014b). For countries, broadband constitutes a *necessary*, *although not sufficient*, infrastructure related condition for DDI. Other factors, such as the (local) availability of data-driven services – and the related question of how well countries' co-location and backhaul markets function –, and an open Internet that enables non-discriminatory access and the free flow of data, are also essential to ensure that DDI takes root within national borders.

Figure 3.4. **OECD wireless broadband penetration, by technology, December 2009 and June 2013**



Note: Standard mobile broadband subscriptions may include dedicated mobile data subscriptions when breakdowns are not available. Israel: data for June 2010 instead of 2009.

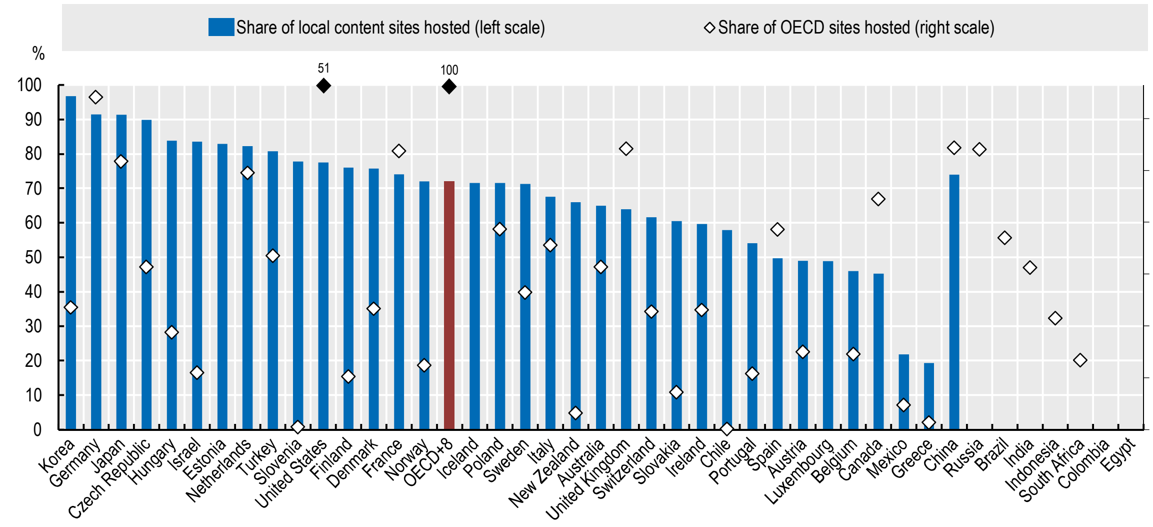
Source: OECD (2014), *Measuring the Digital Economy,* OECD Publishing, Paris, http://dx.doi.org/10.1787/888933148361, based on *OECD Broadband Portal*, www.oecd.org/sti/broadband/oecdbroadbandportal.htm, May 2014.

Co-location and backhaul markets

A recent OECD (2014c) study on "International Cables, Gateways, Backhaul and International Exchange Points" shows that the functioning of local markets for hosting and co-location has an effect on where digital local content (including data-driven services) is hosted. The study analyses the co-location of country code top-level domains (ccTLDs) such as ".fr" (France) and ".jp" (Japan) as identified in the Alexa One Million Domains (a list of the top million sites of the world).[5] The underlying assumption is that "if a larger portion of sites is hosted outside the country, it could indicate that the local market for hosting and co-location is not functioning efficiently" (OECD, 2014c).[6] The analysis of local content sites hosted within countries shows that countries above the OECD average (e.g. Korea, Germany, Japan, Czech Republic, and Hungary) tend to conform to expectations that local content is hosted primarily within the country. Countries such as Greece, Mexico, Canada, Belgium, Luxembourg, Austria, Spain and Portugal have the lowest proportion of their most popular local content sites hosted domestically.

As the OECD (2014c) suggests, "it seems possible … that the market for co-location in Greece is unfavourable and content providers have not chosen a domestic location to host traffic. […] The factors at work in Greece are likely to be similar for Mexico, combined with the proximity to the United States, which has a well-functioning co-location and backhaul market". How well the co-location and backhaul market in the United States functions is indicated by the total number of sites hosted in the United States, which accounts for almost 60% of all top sites hosted in the OECD area in 2013, or more than 50% of all top sites hosted in the OECD area plus Brazil, People's Republic of China (hereafter 'China'), Colombia, Egypt, India, Indonesia, Russia and South Africa altogether (Figure 3.5). Grouping the European and Asian countries into regions may give a better perspective. In 2013, the United States accounted for 42% of all top sites hosted, while Europe hosted 31% of the world's top sites and Asia 11% (Pingdom, 2013). The number of top sites hosted strongly correlates with the number of co-location data centres (see Figure 1.5 in Chapter 1). This suggests that these top countries will be the main destinations for the global data flows on which DDI relies. Further analysis of the data reveals that for mid-income countries, the percentage of local content sites (and data centres) domestically hosted is correlated with the reliability of the electricity supply of that country (OECD, 2014c). This underlines "the importance of considering local energy supply when developing initiatives to enhance local backhaul and data centre markets" (OECD, 2014c).[7]

Figure 3.5. **Local content sites hosted in country, 2013**



Note: Based on the analysis of 429 000 ccTLD of the top one million sites. The remaining sites including the generic top-level domains were omitted from the list, as there is no reliable public data as to where the domains are registered.

Sources: Based on Pingdom, 2013; and www.datacentermap.com, accessed 27 May 2014.

The open Internet

There is still no widely agreed on definition of the open Internet; further studies are needed to develop a better understanding of its characteristics, and its social and economic impact. As highlighted in Chapter 2 of this volume, the openness of the Internet is a condition not only for information and knowledge exchange, but also for global competition among data-driven service providers. Most importantly, it is a vital condition for the nurturing of data-driven services that use and combine content (including data) from more than one source (i.e. mashups[8]) (Leipzig and Li, 2011). Many of these data-driven mashups draw on the activities of firms that have made some of their innovative services available via application programming interfaces (APIs), many of which are open and for free. As a result, a data ecosystem has emerged that is distributed around the world (see Chapter 2).

Ushahidi, Inc., based in Nairobi, Kenya, is an illustration. This non-profit software company relies on the open Internet, as it provides free and open source software and services based on available APIs from Internet firms such as Google and Twitter. One of its first products, created in the aftermath of Kenya's disputed 2007 presidential election, is used to collect eyewitness reports of violence via email and text messages; the locations are visualised on Google Maps. Since then, Ushahidi's data-driven services have been used during crises around the world; for example, in the aftermath of the 2010 earthquake in Haiti and the 2010 earthquake in Chile, it was used to locate the wounded.

As discussed in Chapter 2, barriers to the open Internet, whether legitimate or not, can limit the effects of DDI. Some of these barriers may be technical, such as IP package filtering, or regulatory, such as "data localisation" requirements, and they may be the results of business practices and government policies. Some of these have a legal basis such as privacy and security (see Chapter 5) as well as the protection of trade secrets and copyright (see OECD, 2015b). However, barriers erected through technologies, business practices and/or regulation can have an adverse impact on DDI – for example, if they limit trade and competition.

There is a common interest among countries to find a consensus on how to maintain a vibrant and open Internet and to exchange views on better practices. The OECD's High-Level Meeting on the Internet Economy on 28-29 June 2011 addressed the openness of the Internet and how best to ensure the continued growth and innovation of the Internet and the digital economy. The resulting draft communiqué, which led to the OECD (2011a) *Council Recommendation on Principles for Internet Policy Making*, contains a number of basic principles aimed to help ensure that the Internet remains open and dynamic, that it "allows people to give voice to their democratic aspirations, and that any policy-making associated with it […] promote[s] openness and [is] grounded in respect for human rights and the rule of law". The first five principles, listed below, are particularly relevant for the use of data. This is not to say that other principles are less important to DDI overall:

1. promote and protect the global free flow of information
2. promote the open, distributed and interconnected nature of the Internet
3. promote investment and competition in high-speed networks and services
4. promote and enable the cross-border delivery of services
5. encourage multi-stakeholder cooperation in policy development processes.

Sensors and sensor networks: Enabling the ubiquitous datafication of the physical world

The ubiquity of sensors is already reflected in the widespread use of smartphones, which account for roughly 15% of the 7 billion mobile subscriptions worldwide. This vast reach has its origins in technology's shift two decades ago, from electro-mechanical constructions to sensors and actuators built in silicon, in much the same way chips are built. That made possible the mass production of sensors, which today are embedded in far more than smartphones. Over 30 million interconnected sensors are estimated to be deployed worldwide today in areas such as security, health care, the environment, transport systems and energy control systems, and their numbers are growing by around 30% a year (MGI, 2011). Almost every adult in the OECD area today carries a number of sensors with them on a daily basis.[9]

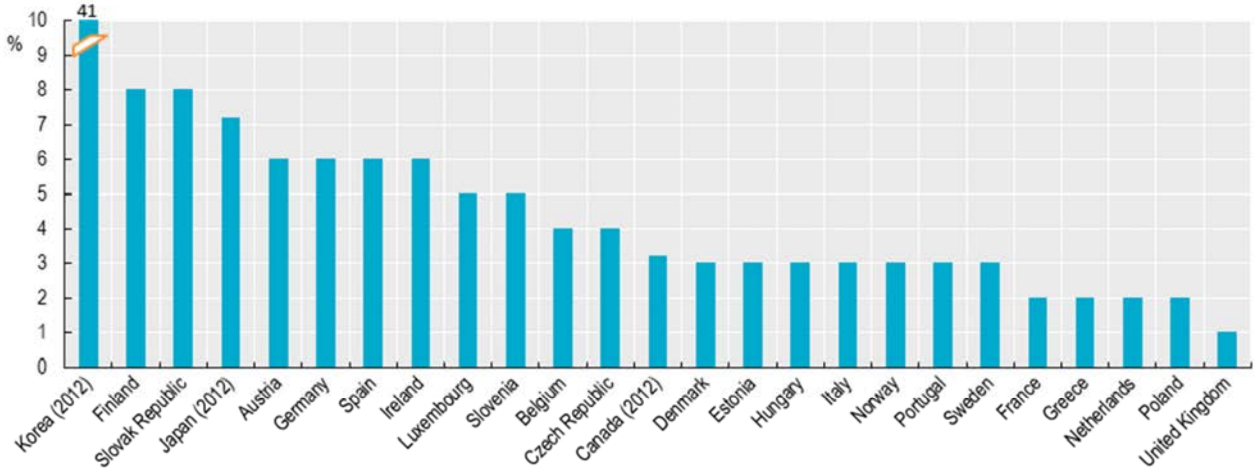
GPS sensors embedded in smartphones, for example, have enabled the generation of geo-locational data that are used in numerous apps and location-based services (mostly in real time), such as by online maps. In 2013, 68% of smartphone users in the OECD area have looked up directions or used a map on their smartphone, 18% more than in 2012; over 32% have searched for information about local businesses, and 14% (44% of those users) have actually visited the businesses afterwards (OECD, 2015a). Beyond its use in online maps and navigation systems, geo-locational data enable new services in areas such as shared mobility (see Chapter 9 of this volume) and multichannel retailing, to name but two.

Sensors have come down in price to such an extent that Apple's iPhone 5S now contains USD 3 worth of them, excluding camera (Hazard Owen, 2013). New sensors are being developed that can be used in novel ways, certain of which may seem unsettling to some. For example, Freescale, a semiconductor company, suggested a sensor that would be of use for gaming, fitness and health applications; at the same time however, it can measure emotions. This electrode-electrocardiogram and capacitive sensor can be integrated into a smart watch or fitness device. It can measure heart rate and sweat, and could thereby be used to measure physical activity. However, the heart rate and sweating are also involuntary indications of emotional states. The sensor would cost no more than USD 0.50 (Hazard Owen, 2013). More and more sensors are also becoming available for integration into different systems and devices. In industrial environments there are a great number of sensors for chemical and mechanical sensing, and more are currently being developed. This trend has been ongoing since the 1980s. A modern engine in a car cannot function without sensors; and it typically contains 50 sensors and sensor packages (Automotive Sensors Conference, 2015).

Looking across all sectors, however, sensor technologies remain underexploited. For instance, radio frequency identification (RFID) is still only adopted by a small set of businesses: less than 10% of all enterprises in OECD countries are using RFID technology. While in Korea 41% of enterprises with 10 or more employees have reported using RFID in 2012, the number is still only between 8% and 5% in Finland, the Slovak Republic, Japan, Austria, Germany, Spain and Ireland. In most other European countries the adoption rates are even lower (Figure 3.6).

Figure 3.6. **The diffusion of RFID in enterprises, 2011**

Percentages of enterprises employing 10 or more persons



Source: OECD (2014), *Measuring the Digital Economy,* OECD Publishing, Paris, http://dx.doi.org/10.1787/888933148361.

Some governments are increasingly deploying sensors as well, using them to measure everything from road conditions to trash collection. For example, the Netherlands government is deploying new fibre-based sensors to measure the stress dikes are undergoing, and integrating these into broadband fibre networks to neighbouring farms. Canada and Sweden use similar sensors to measure stress on the kilometre-long Île d'Orléans bridge near Quebec City and the Götaälvbron Bridge in Gothenburg. In the Swedish case it is a question of keeping the bridge safe and operational until a new bridge is built to replace it (Inaudi and del Grosso, 2008). Analysis of that data could boost smart transport and smart cities, which are discussed further in Chapter 9.

Machine-to-machine communication – Empowering data exchange in the Internet of Things

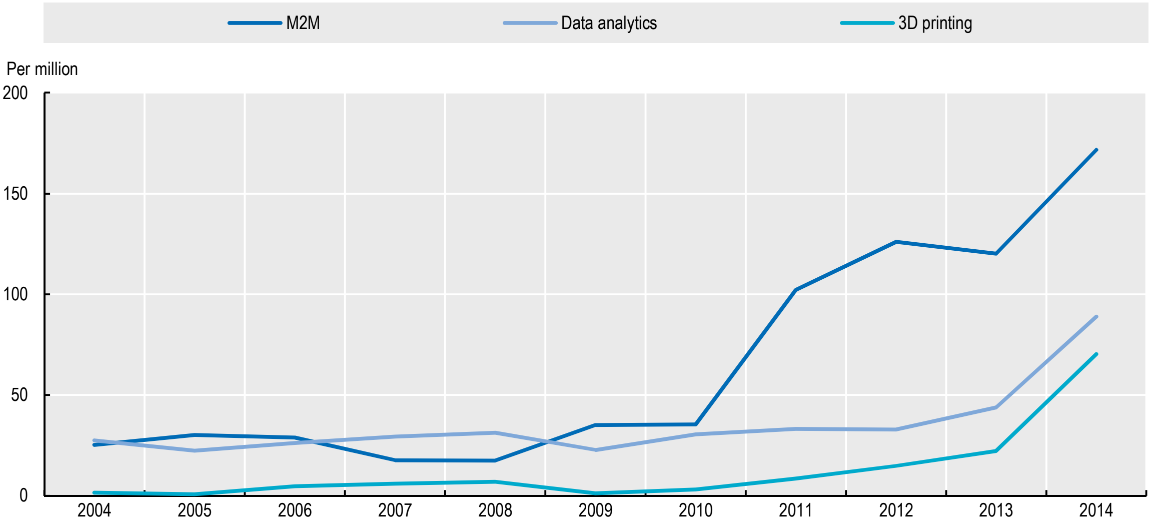
Today in OECD countries, an average family of four persons (including two teenagers) already has ten Internet-connected devices besides smartphones in and around the home, including tablets, printers and scanners, game consoles and increasingly smart TVs, smart meters, and Internet-connected cars. It is estimated that the average number of interconnected devices per household could reach 50 by 2022 (OECD, 2013a). In other words, the number of connected home devices in OECD countries would rise from over 1 billion today to 14 billion by 2022. This calculation only features OECD homes and does not take into account the growth in numbers of connected devices in industry, business, agriculture or public spaces, nor does it include devices in non-member economies. The McKinsey Global Institute (MGI, 2011) estimates that the number of connected smart devices 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, 2013). Ericsson (2010) estimates that by 2025, as many as 50 billion devices will be online. That equates to 6 devices for each of the 8.1 billion people in the world by that time.

All these devices will exchange data to communicate with each other, a process known as machine-to-machine communication (M2M) (OECD, 2012b). Keyword text searches on international patent filings with the World Intellectual Property Organization

(WIPO) under the Patent Cooperation Treaty (PCT) provide evidence that M2M is rapidly increasing in importance when it comes to inventive activity (Figure 3.7). Following a sharp increase in 2011, more than 150 patent applications per million PCT patent applications were related to M2M in 2014, compared to 20 PCT patent applications in 2008.

Figure 3.7. **Patents on M2M, data analytics and 3D printing technologies, 2004-14**

Per million PCT patent applications including selected text strings in abstracts or claims



Note: Patent abstracts and/or claims were searched for the following: (a) M2M: "machine to machine" or "M2M"; (b) Data analytics: "data mining" or "big data" or "data analytics"; (c) 3D printing: "3D printer" or "3D printing". 2014 is limited to data available before 31 May.

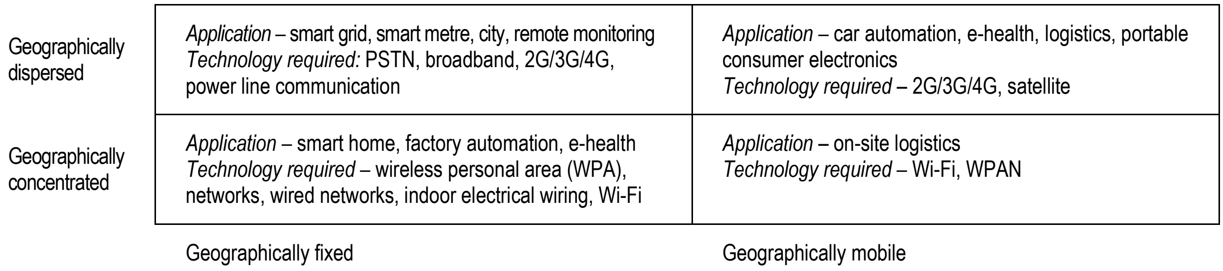
Source: OECD (2014), *Measuring the Digital Economy*, OECD Publishing, Paris, http://dx.doi.org/10.1787/888933148361, based on the OECD PATSTAT database.

Innovative products based on M2M include, for example, smart meters that collect and transmit real-time data on energy (OECD, 2012a), and Internet-connected automobiles that are now able to transmit real-time data on the state of the car's components and environment (OECD, 2012b) (Both applications are discussed in more detail in Chapter 9.) Many of these connected devices are based on sensor and actuator networks that sense and exchange data through wireless links "enabling interaction between people or computers and the surrounding environment" (Verdone et al., 2008, cited in OECD, 2009).[10] These sensors can be regarded as "the interface between the physical world and the world of electrical devices, such as computers" as they measure multiple physical properties. Examples include electronic sensors, biosensors, and chemical sensors (Wilson, 2008). The counterpart is represented by actuators that function the other way round, i.e. whose tasks consist in converting the electrical signal into a physical phenomenon (e.g. displays for quantities measures by sensors such as speedometers, temperature reading for thermostats, but also those that control the motion of a machine).

The use of sensor technology is not what is new here. What is changing is that the data are now not only used in the machine but also shared more widely and combined with other data. In early sensor systems, such as in vehicle engines, the data were measured, processed, acted upon and discarded. More recently however, more and more of the data generated are communicated and stored for further analysis. General Electric is one of the more visible companies promoting this development as an integral part of their vision of the Industrial Internet. Other industry initiatives such as those by Siemens on "networked manufacturing" highlight similar trends (*The Economist*, 2014). These initiatives see a future where machines are built with many different sensors that continuously collect and send data, which are then analysed and acted upon at a system-wide level.

The type of communication used can vary between wired and wireless, short or long range, low or high power, and low or high bandwidth (OECD, 2013a). A way to order the applications and technologies is to look at the geographic distribution and mobility that has to be supported by the M2M networks (Figure 3.8). An increase in mobility and dispersion comes at a cost to energy and bandwidth, meaning that the applications will likely need a bigger battery and can send fewer data than those devices that stay in one location.

Figure 3.8. **Machine-to-machine applications and technologies, by dispersion and mobility**



Geographically dispersed	*Application* – smart grid, smart metre, city, remote monitoring *Technology required:* PSTN, broadband, 2G/3G/4G, power line communication	*Application* – car automation, e-health, logistics, portable consumer electronics *Technology required* – 2G/3G/4G, satellite
Geographically concentrated	*Application* – smart home, factory automation, e-health *Technology required* – wireless personal area (WPA), networks, wired networks, indoor electrical wiring, Wi-Fi	*Application* – on-site logistics *Technology required* – Wi-Fi, WPAN
	Geographically fixed	Geographically mobile

Source: OECD (2012), "Machine-to-Machine communications: Connecting billions of devices", *OECD Digital Economy Papers*, No. 192, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k9gsh2gp043-en.

Given the enormous amount of devices that will come on line in the coming years (50 billion devices by 2025), one important question is whether networks will be able to support all these devices. Network interactions initiated by humans are of a more intermittent character, with pauses between interactions. However, when people interact over networks, they expect interactions within less than 0.2 seconds, which limits the amount of data that can be sent to the user. On a 100 Mbit per second connection, this effectively reduces the amount of data that can be exchanged to 1.25 megabytes or less. M2M is different from traditional applications in that it is more upload focused and less "bursty". A smart meter may send many times more measurement data than it receives in control data over its lifetime, and it does so in a continuous stream. The same holds true with any other type of sensor. The data rates achieved very much depend on the data that are collected and the sampling rate.

Actual data rates also depend on the type of processing done on these data. In the case of an automobile, the data may be processed on board and as a result reduced in size to facilitate easier uploads of whatever data are relevant. As a consequence, however, an automobile would need adequate on-board processing power and sufficient energy supply. In other application cases, the data must first be uploaded because there is no local processing power available, or they need to be combined with other data before they becomes useful. The time sensitivity of data is another concern. If there are real-time feedback loops (e.g. smart meters), the data should be sent uncompressed; absent such loops, "lossless compression" can save bandwidth. In the case of real-time data, the data will need to be streamed.

Furthermore, the increasing deployment of interconnected devices will require governments to address the issue of migration to a new system of Internet addresses (IPv6). The current IPv4 addresses are essentially exhausted, and mechanisms for connecting the next billion devices are urgently needed. IPv6 is a relatively new addressing system that offers the possibility of almost unlimited address space, but adoption has been slow. M2M also raises regulatory challenges related to opening the access to mobile wholesale markets to firms not providing public telecommunication services; there are also numbering and frequency policy issues (see Box 3.1).

Box 3.1. **M2M and regulatory barriers to data-driven mobile applications**

Machine-to-machine communication (M2M) is an enabler of DDI in many industrial applications and services, including logistics, manufacturing, and even health care. However, a major barrier for the M2M-enabled mobile applications (and users) is the lack of competition once a mobile network provider has been chosen. The problem is the SIM (subscriber identity module) 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 between USD 100 and 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 wants to introduce. However, at present it cannot do so in most countries 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 not only to become independent of the mobile network but also to create competition. This would yield billions in savings on mobile connectivity and revenue from new services.

Source: OECD (2012), "Machine-to-Machine communications: Connecting billions of devices", *OECD Digital Economy Papers*, No. 192, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k9gsh2gp043-en.

3.2. The pervasive power of data analytics

Data analytics is the second major development favouring DDI. The large volume of data generated by the Internet, including the IoT, has no value if no information can be extracted from the data. Data analytics refers to a set of techniques and tools that are used to extract information from data. These techniques and tools extract information from data by revealing the context in which the data is embedded and its organisation and structure. They help reveal the "signal from the noise" and with that, the data's "manifold hidden relations (patterns), e.g. correlations among facts, interactions among entities, relations among concepts" (Merelli and Rasetti, 2013; see also Cleveland, 1982 and Zins, 2007). There are a number of terms that are used (as synonyms) to refer to data analytics, some of which may include aspects that go beyond traditional data analysis.

- *Data mining* refers to a set of techniques used to extract information patterns from data sets. It is often said to go beyond data analytics as it combines data analytic methods such as statistics and machine learning with data management technologies (e.g. SQL [structured query language] databases, distributed data management with tools such as Hadoop), and data pre-processing methods (data cleaning). The key aspect here, however, is the discovery of information patterns. Data mining is thus often used as a synonym for another term used more frequently in the past: *knowledge discovery*.

- *Profiling* refers to the use of data analytics for the construction of profiles and the classification of entities in specific profiles, both based on the attributes of these entities. The term is often used in cases where the profiled entities are individuals from which personal identifying information (PII) have been collected; credit scoring, price discrimination and targeted advertisement are typical examples of activities involving profiling. But the term can also be used where non-personal-related entities are being profiled (e.g. malware activities).

- *Business intelligence (BI)* was a term coined by Luhn (1958), who defines it by combining two Webster Dictionary definitions: that of i) intelligence: "the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal", and ii) business: "a collection of activities carried on for whatever purpose, be it science, technology, commerce, industry, law, government, defense, et cetera." Today, BI refers to tools and techniques used to process data that have been previously stored in a database or data warehouse. The objective in BI is the creation of standard reports on a periodic basis, or the display of real-time business-related information on "management dashboards" highlighting key operation metrics for business management. BI typically focuses on databases for business reporting and monitoring. However, the boundaries between BI and data mining are blurring, as BI software vendors are increasingly offering products and services covering BI as well as data mining.

- *Machine or statistical learning* is a subfield in computer science, and more specifically in artificial intelligence. It is concerned with the design, development and use of algorithms that allow computers to "learn" – that is, to perform certain tasks while improving performance with every empirical data set it analyses. Machine learning involves activities such as pattern classification, cluster analysis, and regression (Mitchell et al., 1986, Duda, Hart and Stork, 2000; Russel and Norvig, 2009; Hastie, Tibshirani and Friedman, 2011; James et al., 2013).

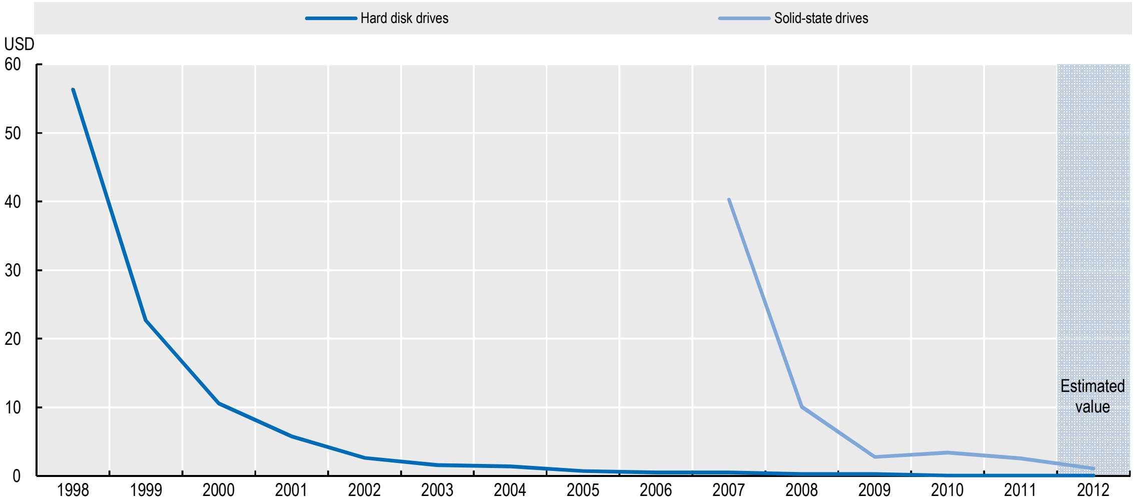
- *Visual analytics* refers to techniques and tools used for data visualisation. They are used for gaining insights at a glance, including through interactive data exploration, and for communicating these insights to others (Unwin, Theus and Hofmann, 2006; Janert, 2010).

Data analytics is now becoming more affordable for start-ups and small and medium-sized enterprises (SMEs), and their adoption will intensify as the volume of data continues to grow. The growing interest in data analytics is also reflected in the number of scientific articles related to the topic. Within the past ten years (2004-2014) that number has, on average, grown by 9% each year (see Figure 1.9 in Chapter 1).

The adoption of data analytics has been greatly facilitated by the declining cost of data storage and processing. In the past, collection, storage and processing were expensive and for the most part available only to large corporations, governments and universities. With ICTs becoming increasingly powerful, ubiquitous and inexpensive, the exploitation of data is now becoming accessible to a wider population (OECD, 2013a). For example, storage costs in the past discouraged keeping data that were no longer, or unlikely to be, needed (OECD, 2011b). But storage costs have decreased to the point where data can be kept for long periods of time if not indefinitely. This is illustrated 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 3.9). With new generation storage technologies such as solid-state drives (SSDs), the decline in costs per gigabyte is even more rapid.

Figure 3.9. **Average data storage cost for consumers, 1998-2012**

In USD per gigabyte



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.

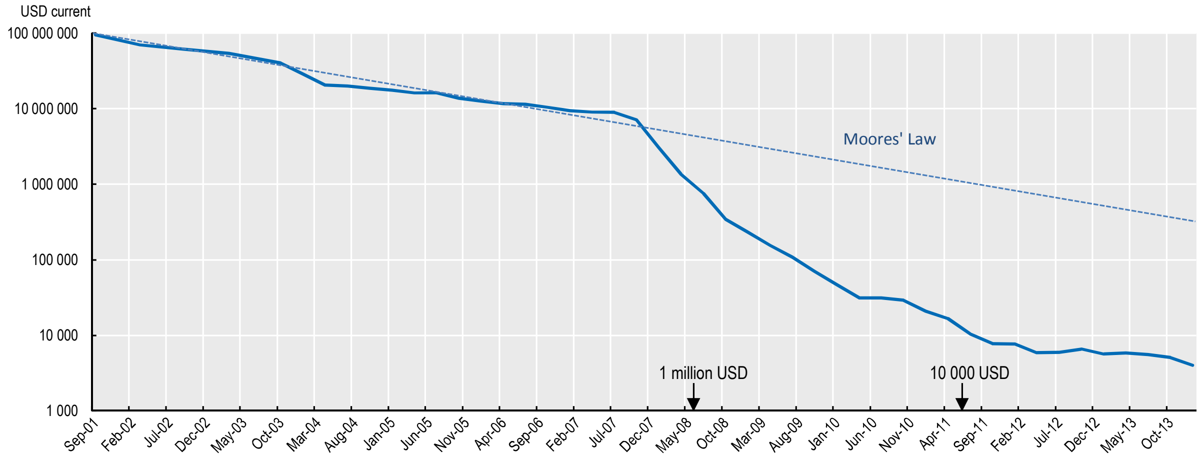
Source: Based on Royal Pingdom blog, December 2011.

The decline in data storage and processing costs is very much a reflection of Moore's Law, which holds that processing power doubles about every 18 months, relative to cost or size. However, as the evolution of the cost of DNA gene sequencing shows, other trends besides Moore's Law have largely contributed to the decreasing cost. The sequencing cost per genome has dropped at higher rates than Moore's Law would predict, from USD 100 million in 2001 to less than USD 6 000 in 2013 (Figure 3.10). Among the factors that have led to the dramatic cost reduction in data storage and processing, the following ones discussed further below should be highlighted:

1. improvements in algorithms and heuristic methods
2. the availability of open source software (OSS), covering the full range of solutions needed for data collection, storage, processing and analytics
3. the availability of computing power at massive scale thanks to cloud computing.

Figure 3.10. **Cost of genome sequencing, 2001-14**

Cost per genome in USD, logarithmic scale



Source: OECD (2014), *Measuring the Digital Economy*, OECD Publishing, Paris, http://dx.doi.org/10.1787/888933148361, based on the National Human Genome Research Institute (NHGRI) Genome Sequencing Program (GSP) www.genome.gov/sequencingcosts/.

Algorithms, heuristic methods and data processing techniques

Significant progress has been made in the development of algorithms and heuristic methods to process and analyse large data sets. It comes as no surprise that Internet firms, in particular providers of web search engines, have been at the forefront of the development and use of techniques and technologies for processing and analysing large volumes of data. They were among the first to confront the problem of handling big streams of mainly unstructured data stored on the web in their daily business operation. Google in particular inspired the development of a series of technologies after it presented MapReduce, a programming framework for processing large data sets in a distributed fashion, and BigTable, a distributed storage system for structured data, in a paper by Dean and Ghemawat (2004) and Chang et al. (2006) respectively. Examples include Hadoop and CouchDB – both open source solutions (under the Apache License) – which have become the engine behind many of today's big data processing platforms (see Chapter 2, Box 2.3).

Some of the progress in data analytics is captured by patents. This includes, for example, the software method patent for a "system and method for efficient large-scale data processing" (US 7650331 B1) that covers the principle of *MapReduce* and that was awarded to Google by the United States Patent and Trademark Office (USPTO) in 2010. Looking at patent applications overall (Figure 3.7), one can observe the growth in the number of patent applications related to data analytics, in particular for "machine learning, data mining or biostatistics, e.g. pattern finding, knowledge discovery, rule extraction, correlation, clustering or classification" (IPC G06F 19/24). However, it is important to highlight that the numbers of data analytics patents can be misleading, for several reasons. Most importantly, numbers of patent applications and patents in data processing in general do not fully reflect ongoing innovation and therefore should be interpreted with caution, in particular when undertaking cross-country comparisons. This is because innovation in data processing is to a large extent embodied in software, for which the application and granting of a patent may vary significantly between countries.

Furthermore, much of the innovation in this field involves open source software (OSS), which is provided with free software licences such as the MIT License,[11] the BSD License,[12] the Apache License[13] and the GNU general public license (GPL v2 or v3).[14] While some of these free software licences provide an express granting of patent rights from contributors to users (e.g. Apache), others may include some form of patent "retaliation" clauses, which stipulate that some rights granted by the licence (e.g. redistribution) may be terminated if patents relating to the licensed software are enforced (e.g. the Apple Public Source License).[15]

The use of patents and copyright has raised a number of concerns in the data analytic community. For example, some have expressed concerns that the patent US 7650331 B1 on MapReduce awarded to Google could put at risk companies that rely entirely on the open source implementations of MapReduce, such as Hadoop and CouchDB (Paul, 2010; Metz, 2010a; 2010b). While such a concern may be justified given that Hadoop is widely used today, including by large companies such as IBM, Oracle and others as well as by Google, expectations are that Google "obtained the patent for 'defensive' purposes" (Paul, 2010).[16] By granting a licence to Apache Hadoop under the Apache Contributor License Agreement (CLA), Google has officially eased fears of legal action against the Hadoop and CouchDB projects (Metz, 2010b). In the area of copyrights, issues are related more to copyright protected data sources, which under some conditions may restrict the effective use of data analytics (Box 3.2).

Box 3.2. **Copyrights and data analytics**

Data analytics is leading to an "automation" of knowledge creation, with text mining constituting a key enabling technology (Lok, 2010). Based on early work by Swanson (1986), scientists are now further exploring the use of data analytics for automated hypothesis generation, and some have proposed analytical frameworks for standardising this scientific approach. Abedi et al. (2012), for example, have developed a hypothesis generation framework (HGF) to identify "crisp semantic associations" among entities of interest. Conceptual biology, as another example, has emerged as a complement to empirical biology, and is characterised by the use of text mining for hypothesis discovery and testing. This involves "partially automated methods for finding evidence in the literature to support hypothetical relationships" (Bekhuis, 2006). Thanks to these types of methods, insights are possible that otherwise would have been difficult to discover. One example is the discovery of adverse effects of drugs (Gurulingappa et al., 2013; Davis et al. 2013).

The potential for productivity gains in the creation of scientific knowledge are thus huge. However, questions have emerged about whether current copyright regimes are appropriately calibrated with regard to "automatic" scientific knowledge creation. According to the JISC (2012) analysis of the value and benefits of text mining, "the barriers limiting uptake of text mining appeared sufficiently significant to restrict seriously current and future text mining in UKFHE [UK further and higher education], irrespective of the degree of potential economic and innovation gains for society." Copyright has been identified as one these barriers, which has led to debates between the scientific community and the publishers of scientific journals (see OECD, 2015b).

Open source analytics

OSS applications that cover the full range of solutions needed for data processing and analysis (including visualisation) have contributed significantly to making data analytics accessible to a wider population. Many data processing and analytic tools that are now spreading across the economy as enablers of new data-driven goods and services were initially developed by Internet firms. Hadoop, the open source implementation of

Google's MapReduce, was already mentioned above. Another well-known example is R, a GPL-licensed open source environment for statistical analysis, which is increasingly used as an alternative to commercial packages such as SPSS and SAS. Today R is also an important part of the product portfolio of many traditional providers of commercial database and enterprise servers such as IBM,[17] Oracle,[18] Microsoft[19] and SAP,[20] which have started integrating R together with Hadoop into their product lines.

Measured by scholarly publications in Google Scholar, Muenchen (2014) estimates the popularity of statistical software including R to have grown significantly over past ten years, the assumption being that "the more popular a software package is, the more likely it will appear in scholarly publications as a topic and as a method of analysis". Muenchen's analysis of the number of articles for the most popular six statistics software from 1995 through 2012 suggests that the most popular statistics software (SPSS, SAS) is declining in popularity, while R is becoming more and more popular.[21] A survey undertaken by the data mining website KDnuggets (2013) confirms the trend that a large number of data analysts are using open source or free software for data analysis.[22]

Cloud computing: Providing super computing power as a utility

Cloud computing has played a significant role in increasing the capacity to store and analyse data. 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, 2014d). Super computing power and data analytics are complementary resources needed to make sense of "big data", as analysis of large volumes of data requires huge computational resources – especially if the analysis needs to be performed in real time.

Cloud computing can be classified into three different service models according to the resources it provides: infrastructure as a service (IaaS), platform as a service (PaaS) and software as a service (SaaS):[23]

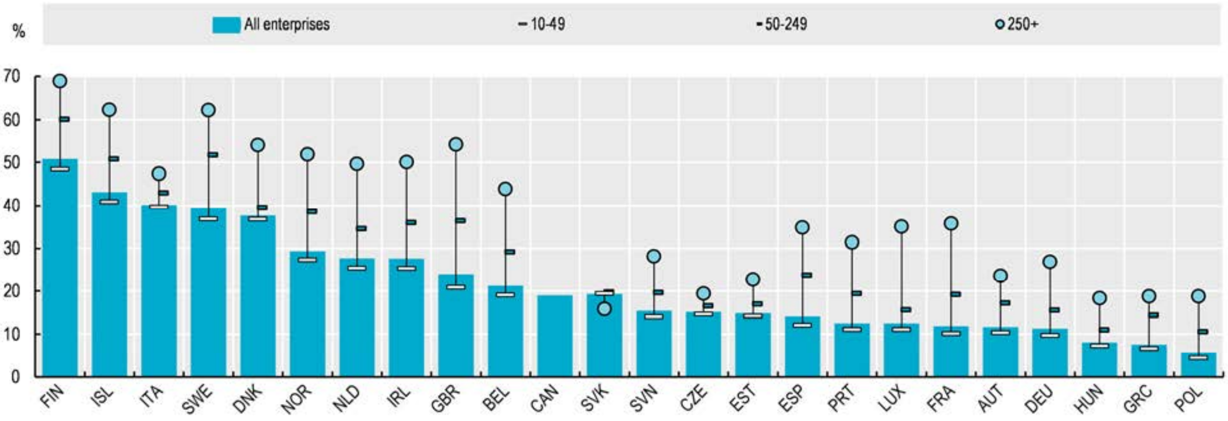
- IaaS provides users with managed and scalable raw resources such as storage and computing resources
- PaaS provides computational resources (full software stack) via a platform on which applications and services can be developed and hosted
- SaaS offers applications running on a cloud infrastructure.

The benefits of cloud computing services can be summarised as efficiency, flexibility, and innovation. Cloud computing reduces computing costs through demand aggregation, system consolidation, and improved asset utilisation. In addition, it provides near-instantaneous increases and reductions in capacity in a pay-as-you-go model, which enables service users to act more responsively to customers' needs and demand without much initial investment in IT infrastructure (Kundra, 2011). All these factors lower the entry barriers of cloud-using markets for start-ups and SMEs, and consequently make the markets more competitive and more innovative.[24] Cloud computing also allows data-driven entrepreneurs to focus on creating and marketing innovative data-driven products without much concern about scaling computing and networking to fit demand.[25] A number of consulting companies have forecast tremendous growth in the public cloud computing market, particularly in the field of SaaS, in the next decade (Ried, 2011).

Surveys by the cloud computing technology provider VMware (2011) confirm that i) increasing business agility and ii) decreasing ICT investment costs are the main motivations for business adoption of cloud computing. Fifty-seven per cent of all respondents point to accelerating the execution of projects and improvement of the customer's experience as the most frequent reasons for cloud computing adoption, followed by rapid adaption to market opportunities (56%) and the ability to scale cost (55%). Recent figures on the adoption of cloud computing reveal significant cross-country variations however (Figure 3.11). In countries such as Finland, Israel, Italy, Sweden and Denmark, almost half of all businesses are already using cloud computing services. There is also large variation by business size, with larger enterprises (250 or more employees) more likely to use cloud computing. In the United Kingdom, for instance, 21% of all smaller enterprises (10 to 49 employees) are using cloud computing services, compared to 54% of all larger enterprises.

Figure 3.11. **Enterprises using cloud computing services by *employment size class*, 2014**

As a percentage of enterprises in each employment size class



Note: Data for Canada refer to the use of "software as a service", a subcategory of cloud computing services.

Source: Based on Eurostat, Information Society Statistics, and Statistics Canada, January 2015.

The use of cloud computing brings other possible benefits that could greatly facilitate the introduction of DDI. Some cloud computing platforms come with standardised interfaces that make it easier to bring several services together, or to interconnect with another smart service that is operating on the same or another cloud platform. As a result, it becomes possible to integrate these services and thereby develop new innovative services. As manufacturers or operators of the smart device will not provide all services, a new market may emerge for companies that would offer to integrate the data from various devices into one package. A home management service may bring together data from sensors and actuators for lights, energy, temperature or movement with other types of sensors and devices, and so provide an integrated overview of all home services. Based on the data collected from many homes and other sources such as weather forecasting, the system may be able to optimise energy consumption. Such cloud-based services could be effective on a very wide scale, with large numbers of customers and large data sets involved.

Despite its widely recognised benefits, there remain significant issues limiting adoption of cloud computing. Privacy and security are among the two most pressing issues, which are discussed further in Chapter 5 of this volume. Another major challenge is the lack of appropriate standards and the potential for vendor lock-in due to the use of proprietary solutions (OECD, 2014d). According to recent surveys among potential users of cloud computing, a lack of standards and the lack of widespread adoption of existing standards are seen as two of the biggest challenges. The lack of open standards is a key problem mainly in the area of PaaS. In this service model, application programming interfaces (APIs) are generally proprietary. Applications developed for one platform typically cannot be easily migrated to another cloud host. While data or infrastructure components that enable cloud computing (e.g. virtual machines) can currently be ported from selected providers to other providers, the process requires an interim step of manually moving the data, software, and components to a non-cloud platform and/or conversion from one proprietary format to another. As a consequence, once an organisation has chosen a PaaS cloud provider, it is – at least at the current stage – locked in (see Chapter 2).

3.3. From informing to driving decision-making

The exponential growth in the data generated and collected, combined with the pervasive power of data analytics, has led to a paradigm shift in the ways knowledge is created and – in particular – decisions are made. These two moments, namely when data are transformed into knowledge (gaining insights) and then used for decision making (taking action), are when the social and economic value of data is mainly reaped. The decision-making phase seems to be the most important one for businesses. According to a survey by the Economist Intelligence Unit (2012), for example, almost 60% of business leaders use big data for decision support and almost 30% for decision automation. This is echoed in estimates by Brynjolfsson, Hitt and Kim (2011), which suggest 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 (see Chapter 1). This section highlights:

1. How value is created when knowledge is extracted from data.
2. How that knowledge is then used for data-driven decision making. Here, two major trends are highlighted: i) decision making is increasingly based on real-time experiments, and ii) it is automated.

Gaining insights: From data to information to knowledge

To understand the value creation process through data analytics, it helps to see data, information and knowledge as different but interrelated concepts. Information is often conveyed through data, while knowledge is typically gained through the assimilation of information. The boundaries between data, information and knowledge may seem extremely fuzzy sometimes, which explains why these concepts are often used as synonyms in media and literature (see Hess and Ostrom, 2007; Daniel Bell, cited in Cleveland, 1982).[26] However, separating these concepts is important to better understand data-driven value creation. A clearer distinction can also help explain certain paradoxes – for example, why one can have a lot of data, but not be able to extract value from them when not equipped with the appropriate analytic capacities (OECD, 2013b; Ubaldi, 2013). Similarly, one can have a lot of information, but not be able to gain knowledge from it – a phenomenon nowadays better known as "information overload" (see Speier et al., 2007)[27] and which Nobel prize-winning economist Herbert Simon described with the words: "a wealth of information creates a poverty of attention" (Shapiro and Varian, 1999). This section discusses the three main functions through which data analytics today is used to gain insights: i) extracting information from unstructured data; ii) real-time monitoring; and iii) inference and prediction. It is interesting to note here that the first two functions are related to two of the three Vs which many see as the key characteristics of big data: variety and velocity (see Glossary). The first V (volume) refers to the exponential growth in data generated and collected, already discussed in the previous section.

Extracting information from unstructured data

Data analytics today has attracted a lot of attention due to its capacity to analyse in particular unstructured data – that is, data that lack a predefined data model (i.e. an abstract representation of "real world" objects and phenomena) (see Hoberman, 2010).[28] Data are considered structured if they are based on such a predefined model. These data models are needed for data processing and can be explicit, as in the case of a SQL database where the data model is reflected in the structure of the database's tables and their inter-linkages. The data model can also be implicit, as in the case of structured web content – or of web logs, where the underlying (implicit) model can be made explicit at relatively low cost. As they do not have an explicit but implicit model, these types of data are often referred to as semi-structured data. Semi-structured data can also refer to data without an explicit data model but to which are attached semantic elements such as tags that highlight the structure within the data. In contrast, with unstructured data, model can only be extracted at significant cost. Typical examples include text-heavy data sets such as text documents and emails, as well as multimedia content such as videos, images and audio streams.

Unstructured data are by far the most frequent type of data, and thus provide the greatest potential for data analytics today. According to a survey of data management professionals by Russom (2007), less than half of the total data stored in businesses is structured. The remaining data are either unstructured (31%) or semi-structured (21%). The author admits, however, that the real share of unstructured (including semi-structured) data could be much higher, as only data management professionals dealing mostly with structured data and rarely with unstructured data were surveyed. Older estimates suggest that the share of unstructured data could be as high as 80% to 85% (see Shilakes and Tylman, 1998).[29] A recent study by IDC (2012b) estimates that not even 5% of the "digital universe" is tagged, and thus can be considered structured or semi-structured data.

However, the difference between structured, semi-structured and unstructured data is becoming less important in the long run, since with growing computing capacities data analytics is increasingly able to automatically *extract* the information embedded in unstructured data. In the past, extracting that information was labour-intensive. The potential of data analytics for automating the processing of unstructured data sets can be illustrated via 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 that automatically crawled through "unstructured" web content, using links to extract even more information about the relevance of the content.[30] 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.[31] (See Watters, 2012 for a comparison of Yahoo! and Google in terms of structured vs. unstructured data.)

A series of technologies have further increased the capacity of data analytics to *process* unstructured data. Optical character recognition (OCR), for example, can transform images of text into machine-encoded text, which then can be interpreted by software, such as for example when indexed for search services such as via Google Books. Natural language processing (NLP), another example, can then be used for tagging or for extracting relevant communication patterns and even emotional patterns. Twitter, for example, has been discussed as a potential (unstructured) data source for analysing and even predicting the "emotional roller coaster" and its impact on the ups and downs of stock markets (Grossman, 2010; *MIT Technology Review*, 2010). Other examples include applications based on face recognition, which – powered by machine-learning algorithms – are able to recognise individuals from images and even video streams. Facebook, for example, is known for using face recognition algorithms to automatically identify and tag its users out of user-provided images (Andrade, Martin and Monteleone, 2013).

Real-time monitoring and tracking

The speed at which data are collected, processed and analysed is often also highlighted as one of the key benefits of data analytics today. The collection and analysis of data in (near to) real time has empowered organisations to base decisions on "close-to-market" evidence. For businesses, this means reduction of time to market and first- or early-mover advantages. For governments, it can mean real-time evidence-based policy making (Reimsbach-Kounatze, 2015).[32] For example, policy 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, 2013). 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). Data analytics is 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, computes and assigns reputation scores to IP addresses in real time to identify traffic patterns from bot-infected machines (Harris, 2011).

Inference and prediction: the new power of machine learning

Data analytics enables the "discovery" of information even if there was no prior record of such information. Such information can be derived in particular, as indicated earlier, by "mining" available data for patterns and correlations. As the volume and variety of available data sets increases, so does the ability to derive further information from these data, notably when they are linked. In particular, personal information can be "inferred" from several pieces of seemingly anonymous or non-personal data (see Chapter 5 of this volume). As the need for data analytics becomes more focused on real-time insights rather than historical and periodical information, the market demands for data analytics change as well, leading to higher demand for advanced specialised data analytic services. In addition, it is becoming increasingly important not only to generate the best actionable output, but also to present it in such a way that it is aligned with the business process that it strives to support, in order to establish competitive differentiation (Dumbill, 2011). For the next couple of years it is expected that most of the value added of data will come from advanced analytical techniques, in particular predictive analytics, simulations, scenario development and advanced data visualisations, many of which are based on advanced use of machine learning (Russom, 2011).

Machine or statistical learning, as mentioned above, is based on the use of algorithms that allow computers to "learn" from data. Having analysed similar situations, computers can apply this analysis to infer and predict a present and a future situation. To make this work, machine learning uses many techniques that are also used in data analytics – for example, a large patient data set can help determine correlations for illnesses. Although machine learning involves such techniques, it is sometimes viewed as different from data analytics, which often attempts to describe the current situation and to find new and unknown correlations in the data. But the distinction is blurring, as machine learning relies on common techniques such as statistical and regression analysis of data to determine future actions in new situations, while data analytics increasingly relies on (unsupervised) learning algorithms for inference and prediction – for instance, via cluster analysis (Hastie, Tibshirani and Friedman, 2011; James et al., 2013). For an historical perspective of machine learning, see Box 3.3.

Web services are notably an area where machine learning is very important. Many of the modern tools and techniques that have become available were developed for web-based services. Search engines are large-scale users of machine-learning technologies, which is not surprising given their relation to translation and speech recognition. Related to this field are the recommendation engines that power services such as Amazon, Deezer, Spotify and Netflix. These services use machine learning to predict the goods that best fit a user's taste. In order to determine this, they use data on the ratings given by the users, for instance to music, as well as information on how they used the service – for example, skipping a song or stopping a movie halfway through and not returning to it. These algorithms are essential to the success of the service: research has shown that consumers will not make a decision when faced with too many options. Machine learning reduces the stress associated with choice (*The Economist*, 2010). Netflix went as far as organising a contest where it awarded winners USD 1 million for the best predictive algorithm. Netflix (2012) tests the algorithms by performing A|B tests,[33] where different algorithms are pitched against each other and their success is measured.

Box 3.3. **Machine learning: An historical perspective**

Translation and speech recognition was one of the first areas where artificial intelligence (AI) was applied. The traditional approach was to describe all the rules related to a language in the software such as the grammar, but also the meaning of words in context. The complexity came from teaching the computer rules to determine the difference among the meanings of one word, such as for instance right as correct, right as a direction, and other meanings of the word. The academic work was mostly performed by linguists, who benefited from an ever better knowledge of language and its rules as a result. However, the computer systems failed to be practical. The alternative approach, statistical analysis of data to derive probabilities, had been discussed in the late 1950s and 1960s. This approach, however, found opposition from noted academics such as Noam Chomsky, who wrote: "we are forced to conclude that grammar is autonomous and independent of meaning and that probabilistic models give no insight into the basic problems of syntactic structure" (quoted in Young, 2010). As a result, the statistical approach was not given full academic attention for some decades.

Box 3.3. **Machine learning: An historical perspective (cont.)**

In 1976, a seminal paper titled "Continuous speech recognition by statistical methods" was published by Frederick Jelinek of IBM. Jelinek (1997) approached the problem of speech recognition not as a linguistic problem, but as a mathematical problem on a par with signals analysis in fixed and wireless networks. This was the start of the resurgence of statistical methods. What made his approach unique was that it relied on statistical analysis of speech and language and not on complex rule models of language. This required the training of the system with many examples of the language. From so-called n-grams (trigrams), combinations of generally three words that were commonly together were derived, and statistics were used to best fit the matching words or pronunciation. The results were significantly better systems compared to earlier rule-based systems, despite the fact that the system was not "knowing" why the result was better.

Today the statistical approach is the basis of speech recognition and translation, such as Siri of Apple and Google Now, and online translation tools offered by Google, Microsoft and Yahoo. Systems are trained by feeding them large corpora of texts, such as subtitled television programmes and the official translations into the official languages of the Canadian Parliament, European Union and United Nations, but also by web pages and scanned books. The results, though not perfect, are often usable. The application continuously adds to the systems by scanning more and more data and by analysing user-provided corrections to texts.

But machine learning is not used solely by Internet firms. In health care, for example, data collected on patients are recorded by imaging and other sensors. Data on the environment in both the health care facilities and the patients' environment can be of relevance as well. Researchers are therefore looking into machine-learning algorithms to better detect conditions and at the same time cut back the number of false positives and negatives. In Chinese Taipei, researchers report that a system using machine learning delivered better results in avoiding false positives while determining three metabolic diseases in newborns. The system was trained on the data of close to 350 000 newborns, which had been collected and tested in prior years (Chen, 2013).

Machine learning is used in industrial applications as well.[34] One of the earlier examples was use in steel mills, where rollers of steel had to apply a controlled force on a hot piece of steel to achieve a particular thickness (Tresp, 2010). The traditional model used an approach based on analytical formulas. However, many effects were non-linear and therefore difficult to model and predict. Using machine learning, the error rate was significantly reduced.

Human decision making: Towards a business culture of data-driven experiments

The ubiquity of data generation and collection has enabled organisations to base their decision-making process on data even more than in the past. Two major trends deserve to be highlighted here: i) human decision making is increasingly based on rapid data-driven experiments; and ii) crowdsourcing –"the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community" (Merriam-Webster, 2014) – has been made further affordable thanks to the increased capacity to extract information from unstructured data from the Internet, and to share data with other analysts.

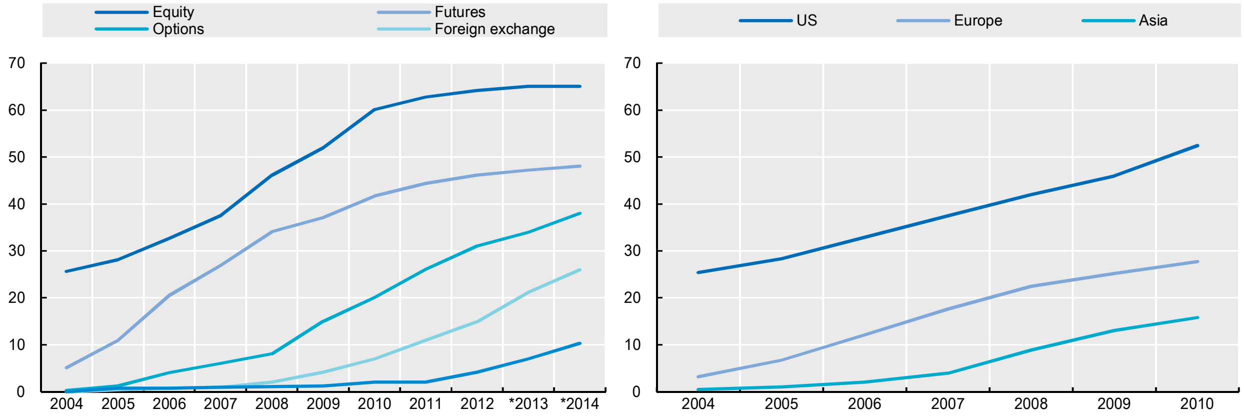
In business, for example, an increasing number of companies are crowdsourcing and analysing data as diverse as online, social media and sensor data to improve the design and quality of their products early in the design phase. They are also analysing these data sources to identify product-related problems to swiftly recall the products if necessary. The rapid analysis of these data sources enables firms to explore different options during product (re-)design and to reduce their opportunity costs and their investment risks. The online payment platform WePay, for instance, designs its web services based on A|B testing. For two months, users are randomly assigned a testing site. The outcome is then measured to determine whether the change in design led to statistically relevant improvements (Christian, 2012). Another example is John Deere, the agriculture equipment manufacturer, which provides farmers with a wide range of agricultural data that enable them to optimise agricultural production by experimenting with the selection of crops, and where and when to plant and plough the crops (Big Data Startups, 2013).

The use of data analytics in decision-making processes described above points to a shift in the way decisions are made in data-driven organisations. Decision makers do not necessarily need to understand the phenomenon before they act on it. In other words: first comes the analytical fact, then the action, and last, if at all, the understanding. For example, a company such as Wal-Mart Stores may change the product placement in its stores based on correlations without the need to know *why* the change will have a positive impact on its revenue. As Anderson (2008) explains: “Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity.” Anderson has even gone as far as to challenge the usefulness of models in an age of massive data sets, arguing that with large enough data sets, machines can detect complex patterns and relationships that are invisible to researchers. The data deluge, he concludes, makes the scientific method obsolete, because correlation is enough (Anderson, 2008; Bollier, 2010). This has opened the door to increasing numbers of applications for decision automation (autonomous machines and systems), while raising “key questions about the constitution of knowledge, the processes of research, how we should engage with information, and the nature and the categorisation of reality” (Boyd and Crawford, 2011; see Chapter 7 of this volume).

Autonomous machines and machine decision making

Data-driven decision making does not stop with the human decision maker. In fact, one of the largest impacts of DDI on (labour) productivity can be expected to come from decision automation, due to “smart” applications that are “able to learn from previous situations and to communicate the results of these situations to other devices and users” (OECD, 2013b). These applications are powered by machine-learning algorithms that are getting more and more powerful. They can perform an increasing number of tasks that required human intervention in the past. Google’s driverless car is an illustration of the potential of smart applications. It is based on the collection of data from all the sensors connected to the car (including video cameras and radar systems), and combines it with data from Google Maps and Google Street View (for data on landmarks, traffic signs and lights). Another example is algorithmic trading systems (ATS) that can autonomously decide what stock to trade, when to trade it, and at what price. In the United States, ATS are estimated to account for more than half of all trades today (Figure 3.12).

Figure 3.12. **Algorithmic trading as a share of total trading**



Note: 2013-14 based on estimates.

Source: Based on *The Economist*, 2012.

Autonomous machines are seen as having great potential in logistics, manufacturing and agriculture. In manufacturing, robots have traditionally been used mostly where their speed, precision, dexterity and ability to work in hazardous conditions are valued. This is radically changing because of sensors, machine learning and cloud computing. Some modern factories, such as the Philips shaver factory in Drachten in the Netherlands, are almost fully robotic (Markoff, 2012). It employs only one-tenth of the workforce employed in its factory in China that makes the same shavers (see Chapter 6 for further discussion on the skills and employment implications of autonomous machines and machine decision making).

The limits of data-driven decision making

The use of data and analytics does not come without limitations, which given the current "big data" hype are even more important to acknowledge. There are considerable risks that the underlying data and analytic algorithms could lead to unexpected false results. The risks are higher where decision making is automated – as illustrated by the case of the Knight Capital Group, which lost USD 440 million in 2012, most of it in less than an hour, because it's ATS behaved unexpectedly (Mehta, 2012). Users should be aware of these limitations; otherwise they may (unintentionally) cause social and economic harm (costs), to themselves as well as to third parties. The risk of social and economic costs to third parties (including individuals) raises important questions related to the attribution of responsibility for inappropriate decisions.

That risk also raises the question of the extent to which the risk-based approach to security and privacy discussed in Chapter 5 allows taking into account all potential (negative) externalities. At times the incentives for the data and analytic user (i.e. the data controller) to minimise the risks to third parties may indeed be low. This is typically the case where the third parties will bear the main share of the social and economic costs of the data controller's action. Accordingly, there should be a careful examination both of the appropriateness of fully automated decision making, and of the need for human intervention in areas where the potential harm of such decisions may be significant (e.g. harm to the life and well-being of individuals, denial of financial or social rights).

Thought must also be given to increasing the transparency of the processes and algorithms underlying these automated decisions (i.e. algorithmic transparency),[35] while preserving proprietary intellectual property rights (IPRs) including in particular trade secrets, which some businesses would consider the "secret sauce" of their business operations (see OECD, 2015b).

The following types of errors are discussed further below: i) data errors; ii) errors that come with inappropriate use of data and analytics; and iii) errors caused by unexpected changes in the environment from which data are collected (i.e. the data environment). The latter issue is particularly relevant for decision automation.

Poor-quality data

The information that can be extracted from data depends on the quality of the data. Poor-quality data will therefore almost always lead to poor results ("garbage in, garbage out"). Therefore, data cleaning (or scrubbing) is often emphasised as an important step before the data can be analysed. And this often involves significant costs, as it can account for 50% to 80% of a data analyst's time together with the actual data collection (Lohr, 2014). As highlighted in Chapter 4, information is context dependent, and as a result data quality will typically depend on the intended use of the data: data that are of good quality for certain applications can thus be of poor quality for other applications (Lohr, 2014). The OECD (2011c) *Quality Framework and Guidelines for OECD Statistical Activities* therefore defines data quality as "fitness for use" in terms of user needs: "If data is accurate, they cannot be said to be of good quality if they are produced too late to be useful, or cannot be easily accessed, or appear to conflict with other data". The OECD (2013c) *Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data* (OECD Privacy Guidelines) also provides a number of criteria for data quality in the context of privacy protection. The Recommendation states that "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".

Inappropriate use of data and analytics

As highlighted above, some have suggested that with big data, decision makers could base their actions only on analytical facts without the need to understand the phenomenon on which they are acting. As correlation would be enough with big data, scientific methods and theories would be less important. While it is true that analytics can be effective in detecting correlations in "big data", especially those that would not be visible with smaller-sized volumes of data, it is also widely accepted among practitioners that data analysis itself relies on rigorous scientific methods in order to produce appropriate results.

The rigour starts with how the quality of the data is assessed and assured. But even if data are of good quality, data analytics can still lead to wrong results if the data used are irrelevant and do not fit the business or scientific questions they are supposed to answer (see section above). Experts recognise that it is often too tempting to think that with big data one has sufficient information to answer almost every question and to neglect data biases that could lead to false conclusions, because correlations can often appear statistically significant even if there is no causal relationship. Marcus and Davis (2014) give the illustration of big data analysis revealing a strong correlation of the United States murder rate with the market share of Internet Explorer from 2006 to 2011. Obviously, any causal relationship between the two variables is spurious.

The risk of inappropriate use of data and analytics underlines the need for high skills in data analysis, and challenges the belief that everyone and every organisation today is in a position to apply data analytics appropriately (see Chapter 6 on the skills implications of DDI). As O'Neil (2013a) argues, the simplicity of applying machine-learning algorithms today thanks to software improvements makes it easy for non-experts to believe in software-generated answers that may not correspond to reality. Furthermore, the need for understanding causal relationships means that sufficient domain-specific knowledge is necessary to apply data and analytics effectively. Obviously, the availability of high skills in data analysis and the rigorous use of data and analytics do not prevent data and analytics from being wrongly used intentionally for economic, political or other advantages. Literature is full of cases where (e.g.) sophisticated econometric models have been used to lie with data. O'Neil (2013b) discusses examples.

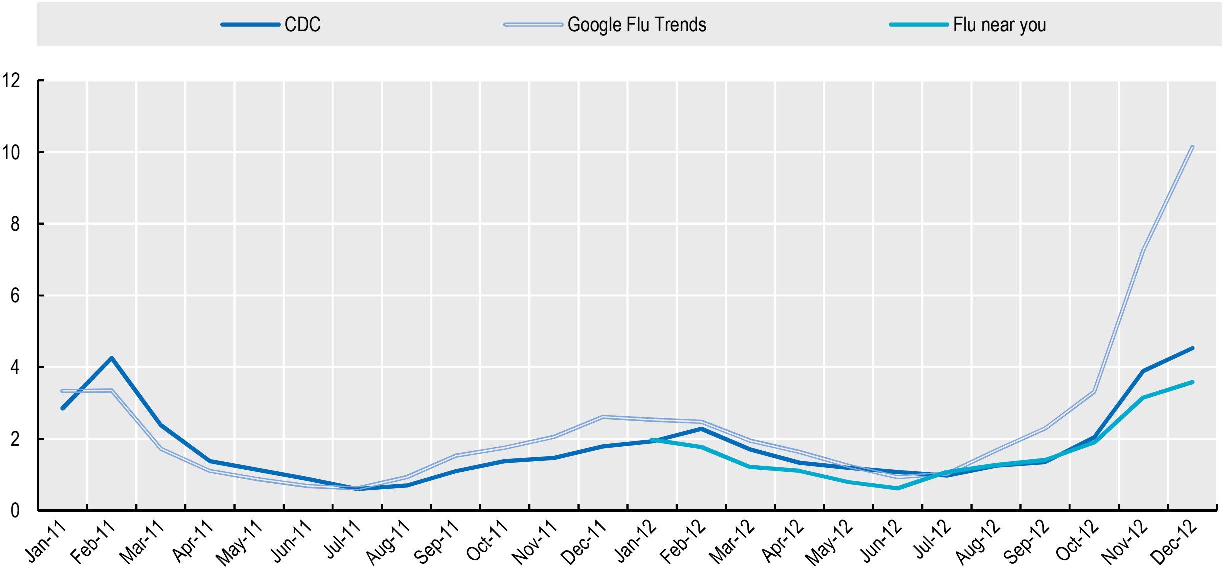
The changing data environment

Even when the data and the analytics are perfectly used initially, this does not mean that they will always deliver the right results. Data analytics, in particular when used for decision automation, can sometimes be easily "gamed" once the factors affecting the underlying algorithms have been understood – for example, through reverse engineering. Marcus and Davis (2014) for example present the case where academic essay evaluation analytics that relied on measures like sentence length and word sophistication to determine typical scores given by human graders, were gamed by students who suddenly started "writing long sentences and using obscure words, rather than learning how to actually formulate and write clear, coherent text". More popular examples (with business implications) are techniques known as "Google bombing" and "spamdexing", where users are adjusting Internet content, links and sites to artificially elevate website search placement in search engines (Segal, 2011; Marcus and Davis, 2014).

Data analytics does not need to be intentionally gamed to lead to wrong results. Often it is simply not sufficiently robust to address unexpected changes in the data environment. This is because data analytics users (including the developers of autonomous systems) cannot envision all eventualities that could affect the functioning of their analytic algorithms and software, in particular when they are used in a dynamic environment. In other words, data analytics is not perfect and some environments are more challenging than others. The case of the Knight Capital Group, which lost USD 440 million in financial markets in 2012 due to some unexpected behaviour from its trading algorithm, was already mentioned above. A more recent example is Google Flu Trends, which is based on Google Insights for Search and provides statistics on the regional and time-based popularity of specific keywords that correlate with flu infections.[36] Google Flu Trends has been used by researchers and citizens as a means to accurately estimate flu infection trends, and this at faster rates than the statistics provided by the Centers for Disease Control and Prevention (CDC). However, in January 2013, Google Flu Trends drastically overestimated flu infection rates in the United States (Figure 3.13). Experts assessed that this was due to "widespread media coverage of [that] year's severe US flu season", which triggered an additional wave of flu-related searches by people unaffected by flu (Butler, 2013).

Figure 3.13. **Fever estimations in the United States, January 2011-December 2012**

Estimated % of population with influenza-like illness, monthly average



Source: Based on Butler, 2013.

These incidents, intentioned or not, are caused by the dynamic nature of the data environment. The assumptions underlying many data analytics applications may change over time, either because users suddenly change their behaviour in unexpected ways as presented above (essay evaluation analytics) or because new behavioural patterns emerge out of the complexity of the data environment (algorithmic trading). As Lazer et al. (2014) further explain, one major cause of the failures (such as in the case of Google Flu Trends) may have been that the Internet constantly changes, and as a result the Google search engine itself constantly changes. Patterns in the data collected are therefore hardly robust over time.

3.4. Key findings and policy conclusions

This chapter has highlighted the key enablers of data-driven innovation, the understanding of which is crucial for governments to assess the degree of readiness of their economies to take advantage of DDI. Economies in which these enablers are more prevalent are expected to be in a better position to reap the benefits of DDI. This does not mean that all factors need to be fully developed in order to realise those benefits. As shown in Chapter 2 of this volume, the global nature of the data ecosystem allows countries to profit from DDI through data- and analytics-related goods and services produced elsewhere. However, it can be assumed that countries with enhanced capacities to supply *and* use data and analytics will be in the best position.

A fast and open Internet (including the Internet of Things) is the most fundamental condition for DDI. In particular:

1. Mobile broadband enables mobile devices (many of which are smart devices enabled by M2M and sensors) to be used for DDI, including in remote and less developed areas where DDI could bring much needed (regional) growth (e.g. DDI

in agriculture). However, while in Finland, Australia, Japan, Sweden, Denmark and Korea mobile penetration rates exceeded 100%, they are still at 40% or less in Portugal, Greece, Chile, Turkey, Hungary and Mexico.

2. The functioning of co-location and backhaul markets is key for the local deployment of data-driven services. Analysis of the share of the most popular local content sites hosted domestically suggests that the local market for hosting and co-location is not functioning efficiently in countries with a low proportion of their most popular local content sites hosted domestically. Underlying reasons may differ vastly from country to country and may deserve for follow-up studies.

3. There are regulatory barriers preventing effective deployment of some M2M-based mobile applications. In particular, large-scale M2M users such as car manufacturers who need to control their own devices with their own SIM cards cannot do so in many countries, as it would make a car manufacturer the equivalent of a mobile operator. Removing regulatory barriers to entry in the mobile market would allow the million-device customer to become independent of the mobile network and to further competition.

4. Barriers to the open Internet, whether legitimate or not, can limit the effects of DDI. Some of these barriers may be technical, such as IP package filtering, or regulatory, such as "data localisation" requirements, and they may be the results of business practices and government policies. Some of these have a legal basis such as privacy and security (see Chapter 5) as well as the protection of trade secrets and copyright (see OECD, 2015b). However, these barriers can have an adverse impact on DDI – for example, if they limit trade and competition (see Chapter 2). Governments looking to promote DDI in their countries should take the OECD (2011b) *Council Recommendation on Principles for Internet Policy Making* further into consideration as well as ongoing OECD work to develop a better understanding of the characteristics, and the social and economic impact of the open Internet.

Data analytics and super computing power are complementary resources needed for the use of "big data". Access to these resources is therefore critical for realising the potential of DDI. However, there are two important issues:

1. Lack of interoperability and the risk of vendor lock-in are two major concerns potential cloud computing users have that may warrant policy makers' attention. The lack of open standards is mainly a huge problem in the area of PaaS. Initiatives are under way to address this issue, covering the full spectrum from infrastructure standards – such as virtualisation formats and open APIs for management to standards for web applications and services and data linkage, but also privacy, security and identity management.

2. Access to and effective use of data analytics can be affected by IPR, in two ways. First, data analytics (including its algorithms) can be protected by software patents or copyright, which under some conditions can limit access and the range of applications. Second, in the special case of text mining, the use of data analytics can be restricted due to copyright, even where scientists may have legal access to scientific publications. While the first issue may not always pose a serious problem to the data analytics community, the latter is still subject to controversial debates between the scientific community and the publishers of scientific journals.

Finally, analysis of value creation mechanisms shows that:

1. Data analytics leads to new ways of decision making, in particular through low cost and rapid experiments (often based on correlations and A|B testing), as well as through use of autonomous machines and systems (based on machine-learning algorithms) that are able to learn from previous situations and to (autonomously) improve decision making.

2. However, there are serious risks that the use of data and analytics may lead to inappropriate results. Strong skills are thus needed in data analysis and domain-specific knowledge, as discussed in Chapter 6. This challenges the "democratisation" of data analytics, according to which everyone and every organisation can use data and analytics appropriately. The risks are elevated when analytics are used for decision automation in dynamic environments, in which case the environments need to be properly understood as well. Likewise, careful examination of the appropriateness of fully automated decision making and of the need for algorithmic transparency and human intervention is critical in areas where the potential harm of such decisions may be significant (e.g. harm to the life and well-being of individuals, denial of financial or social rights).

Notes

1 It is estimated that the Babylonian census, introduced in 1800 BC, was the first practice of systematically counting and recording people and commodities for taxation and other purposes. See www.wolframalpha.com/docs/timeline.

2 Luhn (1958) introduces the concept of business intelligence, citing the following Webster's Dictionary definition of intelligence: "the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal". He further defines business as "a collection of activities carried on for whatever purpose, be it science, technology, commerce, industry, law, government, defense, et cetera."

3 As Mayer-Schönberger and Cukier (2013) explain: "To datafy a phenomenon is to put it in a quantified format so it can be tabulated and analyzed".

4 It has been argued however that in some cases, additional measures guaranteeing the delivery of time-sensitive data may be needed (e.g. quality of service).

5 For that study's analysis, "the generic top-level domains were omitted from the list, as there is no reliable public data as to where the domains are registered. Out of the one million top sites, 946 700 were scanned, 474 000 were generic top-level domains, 40 000 had no identifiable host country, and 3 700 had no identifiable domain, just an IP-address. The remaining 429 000 domains were analysed and their hosting country identified. For each country the percentage of domains hosted in the country were [sic] identified". See also Royal Pingdom blog, available at http://royal.pingdom.com/2012/06/27/tiny-percentage-of-world-top-1-million-sites-hosted-africa/.

6 As discussed in the study, there are caveats that need to be highlighted – for example, the high share of generic top-level domains hosted in the United States for historical reasons. The ccTLD .us is also a valid top-level domain in that country, but it is very lightly used. … There are some further caveats with the data. In some cases there may be a national and an international site for the content. For example, it might be the case that a newspaper has a site hosted in the country, for all web requests coming from the country and an international site located close to where the countries [sic] diaspora lives. The local site will likely not show up as the query was run from Sweden. Similarly, some of the largest sites in the world use content delivery networks (CDNs) to distribute their data. These sites show as hosted outside the country, though for visitors in country, they may be local".

7 This comes as no surprise considering the importance of reliable energy supply for the operation of data centres (Reimsbach-Kounatze, 2009).

8 Mashups or mash-ups are web applications that use and combine content from different sources, including but not limited to web documents such as web pages and multimedia content; data such as cartographic and geographic data; application converters; and communication and visualisation tools.

9 Today a standard smartphone, for example, contains the following sensors besides microphones and video sensors: (i) accelerometer – measures magnitude and direction of acceleration, (ii) global positioning system (GPS) – measures location based on the position of satellites, (iii) gyroscope – measures orientation of a device, (iv) barometer – measures air pressure, which is also used to measure vertical movement, and (v) magnetometer (compass) – measures device orientation.

10 These sensors can be regarded as "the interface between the physical world and the world of electrical devices, such as computers" as they measure multiple physical properties. Examples include electronic sensors, biosensors, and chemical sensors (see Wilson, J. (2008), Sensor Technology Handbook, Newnes/Elsevier, Oxford). The counterpart is represented by actuators that function the other way round, i.e. whose tasks consist in converting the electrical signal into a physical phenomenon (e.g. displays for quantities measures by sensors such as speedometers, temperature reading for thermostats, but also those that control the motion of a machine).

11 "The MIT License is a permissive license that is short and to the point. It lets people do anything they want with your code as long as they provide attribution back to you and don't hold you liable. jQuery and Rails use the MIT License." (See http://choosealicense.com/.)

12 The BSD License is "a permissive license that comes in two variants, the BSD 2-Clause and BSD 3-Clause. Both have very minute differences to the MIT license." (See http://choosealicense.com/licenses/.)

13 "The Apache License is a permissive license similar to the MIT License, but also provides an express grant of patent rights from contributors to users. Apache, SVN, and NuGet use the Apache License." (See http://choosealicense.com/.)

14 "The GPL (V2 or V3) is a copyleft license that requires anyone who distributes your code or a derivative work to make the source available under the same terms. V3 is similar to V2, but further restricts use in hardware that forbids software alterations. Linux, Git, and WordPress use the GPL." (See http://choosealicense.com/.)

15 A well-known example is R, a GPL-licensed open source environment for statistical analysis, which is increasingly used as an alternative to commercial packages such as SPSS and SAS (see section below). Another example is the library scikit-learn, which provides a set of data analytics and machine-learning algorithms for the programming language Python, and is provided under the BSD License. It was developed during a Google Summer of Code project as a third party extension to a separately developed Python project, SciPy, a BSD-licensed open source ecosystem for scientific and technical computing.

16 As Paul explains: "Many companies in technical fields attempt to collect as many broad patents as they can so that they will have ammunition with which to retaliate when they are faced with patent infringement lawsuits." For more on IP strategies see OECD (2015b).

17 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, 2012a).

18 Oracle provides its Big Data Appliance as a combination of open source and proprietary solutions for enterprises' big data requirements. It includes, among others, the Oracle Big Data Connectors that allows customers to use Oracle's data warehouse

and analytics technologies together with Hadoop; the Oracle R Connector, which allows the use of Hadoop with R; and the Oracle NoSQL Database, which is based on Oracle Berkeley DB, a high-performance embedded database.

19 In 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).

20 In 2012, SAP announced its roadmap to integrate Hadoop with its real-time data platform SAP HANA and SAP Sybase IQ.

21 Surveys on the use of data analytics software are also confirming these results. A survey by KDnuggets, for example, suggests that RapidAnalytics (free edition), R, Excel, Weka/Pentaho, and Python were the top five data analytics tools used in 2013. Although all except Excel are free or open source tools, the authors of the survey conclude that commercial and free/open source software are used almost equally among the surveyed data analysts.

22 Four of the top five packages used were open source, including RapidMiner (free edition), R, Weka/Pentaho, and the combination of Python tools numpy, scipy and panda.

23 Sometimes, clouds are also classified into private, public, and hybrid clouds according to their ownership and control of management of the clouds.

24 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.

25 Big data solutions are typically provided in three forms: software-only, as a software-hardware appliance, or cloud-based (Dumbill, 2012b). 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.

26 According to Hess and Ostrom (2007), "knowledge […] refers to all intelligible ideas, information, and data in whatever form in which it is expressed or obtained". Daniel Bell defines information as "data processing in the broadest sense" and knowledge as "an organized set of statements of facts or ideas […] communicated to other".

27 As Speier et al. explain: "Information overload occurs when the amount of input to a system exceeds its processing capacity. Decision makers have fairly limited cognitive processing capacity. Consequently, when information overload occurs, it is likely that a reduction in decision quality will occur."

28 According to Hoberman, "a data model is a wayfinding tool for both business and IT professionals, which uses a set of symbols and text to precisely explain a subset of real information to improve communication within the organization and thereby lead to a more flexible and stable application environment."

29 In health care, for example, health records and medical images are the dominant type of data, and they are sometimes stored as unstructured data. Estimates suggest that in the United States alone, 2.5 petabytes are stored away each year from mammograms.

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

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

32 Real-time data can also be a source for real-time evidence for policy making. 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. It is not only based on five times what the US government collects, but it is also cheaper, and is has a periodicity of days as opposed to months.

33 A|B testing is typically based on a sample that is split into 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.

34 Now that sensor data are becoming more widely available in industrial applications, companies such as Siemens and General Electric are increasingly promoting machine-learning applications.

35 At the fourth meeting of the OECD Global Forum on the Knowledge Economy (GFKE) on “Data-driven Innovation for a Resilient Society”, held 2-3 October 2014 in Tokyo (www.gfke2014.jp/), EPIC President, Marc Rotenberg, highlighted the need for “algorithmic transparency”, which would make data processes that impact individuals public (see Annex of Chapter 1 of this volume on the highlights of the GFKE).

36 Google Trends now also include surveillance for a second disease, dengue.

References

Abedi, V. et al. (2012), "An automated framework for hypotheses generation using literature", *BioData Mining*, Vol. 5, No. 1.

Anderson, C. (2008), "The end of theory: The data deluge makes the scientific method obsolete", *Wired*, 23 June, www.wired.com/science/discoveries/magazine/16-07/pb_theory/, accessed 05 May 2015.

Andrade, N.N.G, A. Martin and S. Monteleone (2013), "'All the better to see you with, my dear': Facial recognition and privacy in online social networks", *IEEE Security & Privacy*, Vol. 11, No. 3, pp. 21-28, May/June.

Automotive Sensors Conference (2015), Conference website, www.automotivesensors2015.com, accessed 21 October 2014.

Bekhuis, T. (2006), "Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy", *Biomed Digit Library*, http://dx.doi.org/10.1186/1742-5581-3-2 (accessed 13 August 2014).

Bertolucci, J. (2013), "IBM, universities team up to build data scientists", *InformationWeek*, 15 January, www.informationweek.com/big-data/big-data-analytics/ibm-universities-team-up-to-build-data-scientists/.

Big Data Startups (2013), "Walmart is making big data part of its DNA", www.bigdata-startups.com/BigData-startup/walmart-making-big-data-part-dna/, accessed 22 August 2014.

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

Boyd, D. and K. Crawford (2011), "Six Provocations for Big Data", A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, 21 September, http://ssrn.com/abstract=1926431 or http://dx.doi.org/10.2139/ssrn.1926431.

Brynjolfsson, E., L.M. Hitt and H.H. Kim (2011), "Strength in numbers: How does data-driven decision making affect firm performance?", 22 April, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486.

Butler, D. (2013), "When Google got flu wrong", *Nature*, 13 February, www.nature.com/news/when-google-got-flu-wrong-1.12413.

Carrière-Swallow, Y. and F. Labbé (2013), "Nowcasting with Google trends in an emerging market", *Journal of Forecasting,* Vol. 32, No. 4, pp. 289-98.

Chang, F. et al. (2006), "Bigtable: A distributed storage system for structured data", Google Research Publications, appeared in the proceedings of the Seventh Symposium on Operating System Design and Implementation (OSDI'06), November, http://research.google.com/archive/bigtable.html.

Chen, W.-H. et al. (2013), "Web-based newborn screening system for metabolic diseases: Machine learning versus clinicians", Journal of Medical Internet Research, www.jmir.org/2013/5/e98/.

Choi, H. and H. Varian (2009), "Predicting the present with Google trends", *Social Science Electronic Publishing*, 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 (2013), "Cisco Visual Networking Index: Forecast and methodology", 2012-2017, www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html, accessed 12 November 2013.

Cleveland, H. (1982), "Information as a resource", *The Futurist*, December, http://hbswk.hbs.edu/pdf/20000905cleveland.pdf, accessed 12 September 2013.

Davis, A.P. et al. (2013), "A CTD-Pfizer collaboration: Manual curation of 88 000 scientific articles text mined for drug-disease and drug-phenotype interactions", *Database*, http://dx.doi.org/10.1093/database/bat080 published online 28 November.

Dean, J. and S. Ghemawat (2004), "MapReduce: Simplified data processing on large clusters", Sixth Symposium on Operating System Design and Implementation (OSDI'04), San Francisco, December, http://research.google.com/archive/mapreduce.html.

Duda, R., P.E. Hart and D.G. Stork (2000), *Pattern Classification*, Second Edition, Wiley-Interscience, 9 November.

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

Dumbill, E. (2012b), "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. (2011), "Five big data predictions for 2012", *O'Reilly Radar*, http://strata.oreilly.com/2011/12/5-big-data-predictions-2012.html.

Economist Intelligence Unit (2012), "The deciding factor: Big data & decision making", Economist Intelligence Unit commissioned by Capgemini, 4 June, http://www.capgemini.com/insights-and-resources/by-publication/the-deciding-factor-big-data-decision-making.

Ericsson (2010), "CEO to shareholders: 50 billion connections 2020", Ericsson press release, 13 April, www.ericsson.com/thecompany/press/releases/2010/04/1403231, accessed 12 January 2014.

Goetz, T. (2012), "How to spot the future", *Wired*, 24 April, www.wired.com/2012/04/ff_spotfuture/.

Grossman, L. (2010), "Twitter can predict the stock market", *Wired*, 19 October, www.wired.com/wiredscience/2010/10/twitter-crystal-ball/.

Gurulingappa, H. et al. (2013), "Automatic detection of adverse events to predict drug label changes using text and data mining techniques", *Pharmacoepidemiology and Drug Safety*, Vol. 22, No. 11, November, pp. 1189-94.

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/.

Hastie, T., R. Tibshirani and J. Friedman (2011), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, Springer, New York.

Hazard Owen, L. (2013), "Add $0.50 worth of sensors to your iPhone 5s and it'll be able to track your emotions", *Gigaom,* https://gigaom.com/2013/10/17/add-0-50-worth-of-sensors-to-your-iphone-5s-and-itll-be-able-to-track-your-emotions, accessed 21 October 2014.

Hess, C. and E. Ostrom (2007), *Understanding Knowledge as a Commons: From Theory to Practice,* MIT Press, Cambridge, Mass.

Hey, J. (2004), "The data, information, knowledge, wisdom chain: The metaphorical link", working paper, December, www.dataschemata.com/uploads/7/4/8/7/7487334/dikwchain.pdf, accessed 12 March 2013.

Hoberman, S. (2010), "How do you justify data modeling?", *Erwin Expert Blog*, 20 April, http://erwin.com/community/expert-blogs/how-do-you-justify-data-modeling.

Inmon, W.H. and C. Kelly (1992), *Rdb/VMS: Developing the Data Warehouse*, QED Publishing.

IDC (International Data Corporation) (2012a), "The Digital Universe in 2020: Big data, bigger digital shadows, and biggest growth in the Far East", EMC Digital Universe project, IDC *iView*, December, www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf, accessed 15 September 2013.

IDC (2012b), "Worldwide big data technology and services 2012-2015 forecast", IDC Market Analysis, March, www.idc.com/research/viewtoc.jsp?containerId=233485.

Inaudi, D. and A. del Grosso (2011), "Fiber optic sensors for structural control", www.roctest-group.com/sites/default/files/bibliography/pdf/c196.pdf, accessed 5 September 2011.

ITU (2014), "The world in 2014: ICT facts and figures", International Telecommunication Union, April, www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2014-e.pdf.

Janert, P. (2010), *Data Analysis with Open Source Tools: A Hands-On Guide for Programmers and Data Scientists*, O'Reilly Media.

James, G. et al. (2013), *An Introduction to Statistical Learning with Applications in R*, Springer, New York.

Jelinek, F. (1997), *Statistical Methods for Speech Recognition*, MIT Press.

JISC (2012), "The value and benefits of text mining", JISC**,** www.jisc.ac.uk/sites/default/files/value-text-mining.pdf, accessed 14 June 2014.

Keen, P.G.W. (1978), *Decision Support Systems: An Organizational Perspective*, Addison-Wesley, Reading, Mass.

Lazer, D. et al. (2014), "The parable of Google flu: Traps in big data analysis", *Science*, Vol. 343, No. 14, March, http://scholar.harvard.edu/files/gking/files/0314policyforumff.pdf.

KDnuggets (2013), "What analytics, big data, data mining, data science software you used in the past 12 months for a real project?", KDnuggets, May, www.kdnuggets.com/polls/2013/analytics-big-data-mining-data-science-software.html.

Kundra, V. (2011), "Federal cloud computing strategy", US Chief Information Officers Council, www.cio.gov/documents/federal-cloud-computing-strategy.pdf, accessed 7 October 2013.

Leipzig, J. and X. Li (2011), *Data Mashups in R: A Case Study in Real-World Data Analysis*, O'Reilly Media.

Lohr, S. (2014), "For big-data scientists, 'janitor work' is key hurdle to insights", *New York Times*, 17 August, www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html.

Lok, C. (2010), "Literature mining: Speed reading", *Nature*, Vol. 463, 27 January, pp. 416-18, www.nature.com/news/2010/100127/full/463416a.html.

Luhn, H. (1958), "A business intelligence system", *IBM Journal of Research and Development*, Vol. 2, No. 4, p. 314, http://domino.watson.ibm.com/tchjr/journalindex.nsf/c469af92ea9eceac85256bd50048567c/fc097c29158e395f85256bfa00683d4c!OpenDocument.

Marcus, G. and E. Davis (2014), "Eight (no, nine!) problems with big data", *New York Times*, 6 April, www.nytimes.com/2014/04/07/opinion/eight-no-nine-problems-with-big-data.html.

Markoff, J. (2012), "Skilled Work, Without the Worker", The iEconomy, Part 6: Artificial Competence, *New York Times*, 18 August.

Mayer-Schönberger, V. and K. Cukier (2013), *Big Data: A Revolution That Will Transform How We Live, Work and Think*, John Murray, London.

Mehta, N. (2012), "Knight $440 million loss sealed by rules on cancelling trades", Bloomberg, 14 August, www.bloomberg.com/news/2012-08-14/knight-440-million-loss-sealed-by-new-rules-on-canceling-trades.html.

Merelli, E. and M. Rasetti (2013), "Non locality, topology, formal languages: New global tools to handle large data sets", International Conference on Computational Science, ICCS 2013, *Procedia Computer Science,* No. 18, pp. 90-99, http://dx.doi.org/10.1016/j.procs.2013.05.172.

Merriam-Webster (2014), "Crowdsourcing", *Merriam-Webster.com,* Merriam-Webster, www.merriam-webster.com/dictionary/crowdsourcing, accessed 24 September 2014.

Metz, C. (2010a), "Google's MapReduce patent - No threat to stuffed elephants", *The Register*, 22 February, http://www.theregister.co.uk/2010/02/22/google_mapreduce_patent.

Metz, C. (2010b), "Google blesses Hadoop with MapReduce patent license", *The Register*, 27 April, http://www.theregister.co.uk/2010/04/27/google_licenses_mapreduce_patent_to_hadoop.

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, accessed 24 May 2015.

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.

MIT Technology Review (2010), "Twitter mood predicts the stock market", 18 October, www.technologyreview.com/view/421251/twitter-mood-predicts-the-stock-market/.

Muenchen, R. (2012), "The popularity of data analysis software", *r4stats.com*, http://r4stats.com/articles/popularity/, accessed 21 October 2014.

Netflix (2012), "Netflix recommendations: Beyond the 5 stars", Netflix, June, http://techblog.netflix.com/2012/06/netflix-recommendations-beyond-5-stars.html, accessed 7 June 2014.

O'Connor, S. (2013), "Amazon unpacked", *FT Magazine*, 8 February, www.ft.com/intl/cms/s/2/ed6a985c-70bd-11e2-85d0-00144feab49a.html#slide0, accessed 24 March 2015.

OECD (2015a), *Digital Economy Outlook 2015*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264232440-en.

OECD (2015b), *Inquiries into Intellectual Property's Economic Impact*, OECD, forthcoming.

OECD (2014a), "Connected televisions: Convergence and emerging business models", *OECD Digital Economy Papers*, No. 231, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jzb36wjqkvg-en.

OECD (2014b), *Measuring the Digital Economy: A New Perspective*, OECD Publishing, Paris.

OECD (2014c), "International cables, gateways, backhaul and international exchange points", *OECD Digital Economy Papers*, No. 232, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jz8m9jf3wkl-en.

OECD (2014d), "Cloud computing: The concept, impacts and the role of government policy", *OECD Digital Economy Papers*, No. 240, OECD Publishing, Paris, http://dx.doi.org/10.1787/5jxzf4lcc7f5-en.

OCDE (2013a), "Building blocks for smart networks", OECD Digital Economy Papers, No. 215, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k4dkhvnzv35-en.

OECD (2013b), "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 (2013c), Recommendation of the Council concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data, 11 July, C(2013)79, www.oecd.org/sti/ieconomy/2013-oecd-privacy-guidelines.pdf.

OECD (2012a), "ICT applications for the smart grid: Opportunities and policy implications", *OECD Digital Economy Papers*, No. 190, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k9h2q8v9bln-en.

OECD (2012b), "Machine-to-machine communications: Connecting billions of devices", *OECD Digital Economy Papers*, No. 192, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k9gsh2gp043-en.

OECD (2011a), Recommendation of the Council on Principles for Internet Policy Making, 13 December, C(2011)154, www.oecd.org/sti/ieconomy/49258588.pdf, accessed 19 May 2015.

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 (2011c), Quality Framework and Guidelines for OECD Statistical Activities, 17 January, www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=std/qfs%282011%291&doclanguage=en, accessed 6 January 2015.

OECD (2009), "Smart sensor networks: Technologies and applications for green growth", *OECD Digital Economy Papers*, No. 167, OECD Publishing, Paris, http://dx.doi.org/10.1787/5kml6x0m5vkh-en.

O'Neil, C. (2013a), "K-nearest neighbors: Dangerously simple", *Mathbabe*, 4 April, http://mathbabe.org/2013/04/04/k-nearest-neighbors-dangerously-simple/.

O'Neil, C. (2013b), "We don't need more complicated models, we need to stop lying with our models", *Mathbabe,* 3 April, http://mathbabe.org/2013/04/03/we-dont-need-more-complicated-models-we-need-to-stop-lying-with-our-models (accessed 7 June 2014).

Paul, R. (2010), "Google's MapReduce patent: What does it mean for Hadoop?", *Arstechnica.com*, 20 January, http://arstechnica.com/information-technology/2010/01/googles-mapreduce-patent-what-does-it-mean-for-hadoop, accessed 10 May 2014.

Pingdom (2013), "The top 100 web hosting countries", 14 March, http://royal.pingdom.com/2013/03/14/web-hosting-countries-2013, accessed 20 April 2014.

Reimsbach-Kounatze, C. (2015), "The proliferation of data and implications for official statistics and statistical agencies: A preliminary analysis", *OECD Digital Economy Working Papers*, http://dx.doi.org/10.1787/5js7t9wqzvg8-en.

Reimsbach-Kounatze, C. (2009), "Towards green ICT strategies: Assessing policies and programmes on ICT and the environment", *OECD Digital Economy Papers*, No. 155, OECD Publishing, Paris, http://dx.doi.org/10.1787/222431651031.

Ried, S. (2011), "Sizing the cloud", *Forrester*, 21 April, http://blogs.forrester.com/stefan_ried/11-04-21-sizing_the_cloud (accessed 21 April 2013).

Russell, S. and P. Norvig (2009), *Artificial Intelligence: A Modern Approach*, 3rd edition, Prentice Hall.

Russom, P. (2011), *Big Data Analytics*, TDWI Best Practices Report, The Data Warehousing Institute, p. 24, http://tdwi.org/research/2011/09/best-practices-report-q4-big-data-analytics.aspx.

Russom, P. (2007), "BI search and text analytics: New additions to the BI technology stack", TDWI Best Practices Report, The Data Warehousing Institute, Second Quarter 2007.

Scheuer, Mark (2014), “Continuous EEG monitoring in the intensive care unit”, *Epilepsia*, Vol. 43 (Suppl. 3), Blackwell Publishing, Inc., pp. 114-27 and 200,

Schubert, L., K. Jefferey and B. Neidecker-Lutz (2010), “The future of cloud computing: Opportunities for European cloud computing beyond 2010”, Public Version 1.0, Expert Group Report, European Commission, http://cordis.europa.eu/fp7/ict/ssai/docs/cloud-report-final.pdf.

Segal, D. (2011), “The Dirty Little Secrets of Search” *New York Times*, 12 February, www.nytimes.com/2011/02/13/business/13search.html.

Shapiro, C. and H.R. Varian (1999), *Information Rules: A Strategic Guide to the Network Economy*, Harvard Business Press, Boston, Mass.

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

Smith, B. W. (2014), “Automated Vehicles Are Probably Legal in the United States”, 1 Tex. A&M L. Rev. 411 (2014), http://ssrn.com/abstract=2303904.

Sol, H. (1987), “Expert systems and artificial intelligence in decision support systems”, proceedings of the Second Mini Euroconference, Lunteren, Netherlands, 17-20 November, Springer.

Speier, C., J. Valacich, and I. Vessey (1999), “The influence of task interruption on individual decision making: An information overload perspective”. *Decision Sciences* 30, 2, 7 June, pp 337-360, DOI: 10.1111/j.1540-5915.1999.tb01613.x.

Stewart-Smith, H. (2012), “Foxconn chairman compares his workforce to ‘animals’”, *ZDnet*, 20 January, www.zdnet.com/blog/asia/foxconn-chairman-compares-his-workforce-to-animals/776.

Surowiecki, J. (2011), “A Billion Prices Now”, *The New Yorker*, 30 May.

Swanson D.R. (1986), “Undiscovered Public Knowledge”, *Library Quarterly*, Vol. 56, pp. 103-18.

The Economist (2014), “Networked manufacturing: The digital future”, March, www.economistinsights.com/sites/default/files/EIU%20-%20Siemens%20-%20Networked%20manufacturing%20The%20digital%20future%20WEB.pdf.

The Economist (2012), “High-frequency trading: The fast and the furious”, 25 February, www.economist.com/node/21547988.

The Economist (2010), “The tyranny of choice: You choose”, 18 December, www.economist.com/node/17723028.

Tresp, V. (2010), *On the Growing Impact of Machine Learning in Industry*, Siemens, www.sics.se/~aho/tor/Volker_Tresp_ToR-101125.pdf.

Ubaldi, B. (2013), “Open government data: Towards empirical analysis of open government data initiatives”, *OECD Working Papers on Public Governance*, No. 22, OECD Publishing, Paris, http://dx.doi.org/10.1787/5k46bj4f03s7-en.

Unwin, A., M. Theus and H. Hofmann (2006), “Graphics of large datasets: Visualising a million”, *Statistics and Computing* Series, Springer, Singapore.

Verdone, R. et al. (2008), *Wireless Sensor and Actuator Networks*, Academic Press/Elsevier, London.

VMware (2011), “Business agility and the true economics of cloud computing”, business white paper, www.vmware.com/files/pdf/accelerate/VMware_Business_Agility_and_the_True_Economics_of_Cloud_Computing_White_Paper.pdf, accessed 15 February 2015.

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

Wilson, J. (2008), *Sensor Technology Handbook*, Newnes/Elsevier, Oxford.

Young, S. (2010), “Obituary of Frederick Jelinek 1932-2010: The pioneer of speech recognition technology”, *SLTC Newsletter*, November, www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2r010-11/jelinek/.

Zins, C. (2007), “Conceptual approaches for defining data, information, and knowledge”, *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 4, pp. 479-93, 1 February, www.success.co.il/is/zins_definitions_dik.pdf.

Further reading

Anderson, C. (2012), "The man who makes the future: Wired icon Marc Andreessen", *Wired*, 24 April, www.wired.com/2012/04/ff_andreessen/5/, accessed 05 May 2015.

Amazon (2009), "Amazon Elastic MapReduce Developer Guide API", 30 November, http://s3.amazonaws.com/awsdocs/ElasticMapReduce/latest/emr-dg.pdf (accessed 21 April 2014).

Bakhshi, H. and J. Mateos-Garcia (2012), *Rise of the Datavores: How UK Businesses Analyse and Use Online Data*, Nesta, London.

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, accessed 14 July 2014.

Esmeijer, J. et al. (2013), "Thriving and surviving in a data-driven society", TNO report, 24 September, http://publications.tno.nl/publication/34610048/xcv74S/TNO-2013-R11427.pdf, accessed 24 September 2013.

Hachman, M. (2012), "Facebook now totals 901 million users, profits slip", *PC Magazine*, 23 April, www.pcmag.com/article2/0,2817,2403410,00.asp.

ISO (International Organization for Standardization) (2009), ISO/IEC Standards 15408-1, 2, 3:2009 – "Information technology – Security techniques – Evaluation criteria for IT security", http://isotc.iso.org/livelink/livelink/fetch/2000/2489/Ittf_Home/PubliclyAvailableStandards.htm.

Kan, M. (2013), "Foxconn to speed up 'robot army' deployment", *PCWorld*, 26 June, www.pcworld.com/article/2043026/foxconn-to-speed-up-robot-army-deployment-20000-robots-already-in-its-factories.html.

Kommerskollegium (2014), "No transfer, no trade – The importance of cross-border data transfers for companies based in Sweden", January, www.kommers.se/Documents/dokumentarkiv/publikationer/2014/No_Transfer_No_Trade_webb.pdf, accessed July 2014.

McGuire, T., J. Manyika and M. Chui (2012), "Why big data is the new competitive advantage", *Ivey Business Journal*, July/August, http://iveybusinessjournal.com/topics/strategy/why-big-data-is-the-new-competitive-advantage#.VCJ7lPnoQjM, accessed January 2013.

Metha, N. (2012), "Knight $440 million loss sealed by rules on canceling trades", *Bloomberg*, 14 August, www.bloomberg.com/news/2012-08-14/knight-440-million-loss-sealed-by-new-rules-on-canceling-trades.html.

Mivule, K. (2013), “Utilizing noise addition for data privacy: An overview”, Proceedings of the International Conference on Information and Knowledge Engineering (IKE 2012), pp. 65-71, http://arxiv.org/pdf/1309.3958.pdf, accessed 25 March 2015.

Muthukkaruppan, K. (2010), “The underlying technology of messages”, *Notes, Facebook*, 15 November, www.facebook.com/notes/facebook-engineering/the-underlying-technology-of-messages/454991608919, accessed 24 March 2015.

Narayanan, A. and V. Shmatikov (2007), “How to break anonymity of the Netflix prize dataset”, 22 November, http://arxiv.org/abs/cs/0610105v2.

OECD (2013d), *The OECD Privacy Framework*, OECD Publishing, Paris, www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf.

OECD (1997), Recommendation of the Council concerning Guidelines for Cryptography Policy, [C(97)62/FINAL], 27 March, OECD Publishing, Paris, www.oecd.org/internet/ieconomy/guidelinesforcryptographypolicy.htm.

Ohm, P. (2009), “The rise and fall of invasive ISP surveillance”, *University of Illinois Law Review,* 1417.

Pfitzmann, A. and M. Hansen (2010), “A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management”, v0.34, TU Dresden, 10 August, http://dud.inf.tu-dresden.de/Anon_Terminology.shtml, accessed 24 March 2015.

Rao, L. (2011), “Index and Khosla lead $11M round in Kaggle, a platform for data modeling competitions”, *TechCrunch*, 2 November, http://techcrunch.com/2011/11/02/index-and-khosla-lead-11m-round-in-kaggle-a-platform-for-data-modeling-competitions/.

Warden, P. (2011), “Why you can’t really anonymize your data”, O’Reilly Strata, 17 May, http://strata.oreilly.com/2011/05/anonymize-data-limits.html.

From:

Data-Driven Innovation

Big Data for Growth and Well-Being

Access the complete publication at:

https://doi.org/10.1787/9789264229358-en

Please cite this chapter as:

OECD (2015), “How data now drive innovation”, in *Data-Driven Innovation: Big Data for Growth and Well-Being*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/9789264229358-7-en