Briefing

September 2016

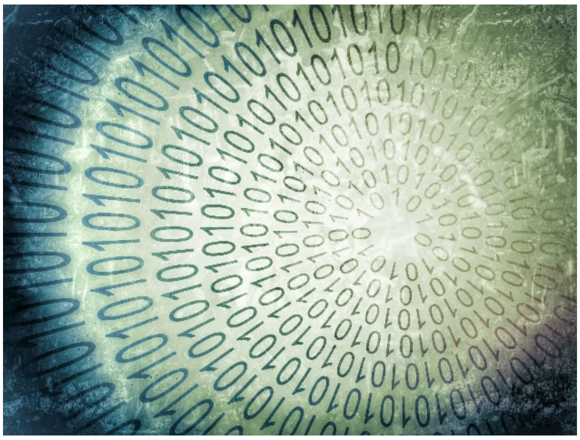
Big data and data analytics

The potential for innovation and growth

SUMMARY

Advances in information and communication technologies, the increasing use of electronic devices and networks, and the digitalisation of production processes mean that vast quantities of data are generated daily by economic and social activities. This 'big data' can be transmitted, collected, aggregated and analysed to provide insights into processes and human behaviours. Big data analytics have the potential to identify efficiencies that can be made in a wide range of sectors, and to lead to innovative new products and services, greater competitiveness and economic growth. Studies suggest that companies that adopt big data analytics can increase productivity by 5%-10% more than companies that do not, and that big data practices in Europe could add 1.9% to GDP between 2014 and 2020.

However big data analytics also pose a number of challenges for policy makers. Whilst protecting privacy and personal data has arguably received the most attention, other big-data-related issues are expected to appear on the European Union policy agenda. These include 'data ownership' principles that determine who shares in the rights associated with big data; data localisation requirements that may unjustifiably interfere with the 'free flow of data'; labour shortages of skilled data workers and data-aware managers; and the creation of a new digital divide that risks marginalising those who do not make extensive use of information and communication technologies.



In this briefing:

EPRS | European Parliamentary Research Service

Author: Ron Davies

Members' Research Service

PE 589.801

EN

Background

The increasing use of electronic devices and networks and the digitalisation of production processes mean that vast quantities of data are generated daily by economic and social activities. According to some estimates, the amount of data produced worldwide is doubling every two years; it is expected to increase from 4.4 zettabytes (or 4.4 trillion gigabytes) in 2013 to 44 zettabytes in 2020. This 'big data' may be drawn from interactions on the web, online commercial transactions, e-government records, social media, mobile phone records, mobile apps, and sensors in objects linked to the Internet of Things. At the same time, information and communication technologies (ICTs) have been advancing, notably in terms of decreases in storage costs, increases in network capacity, improved analytical tools, and the availability of high-performance, on-demand computing through the cloud. These advances have made it possible to store, transmit and process large amounts of data more cheaply, quickly and effectively than before.

The data in these large stores, either on their own or in combination with data collected from other sources, can then be processed to identify patterns and extract meaningful relationships. Advances in analytical techniques mean that an increasing number of types of data, even unstructured data such as natural language texts or videos, can now be analysed. The insights gained can then be used to design new products and services, improve production processes, optimise marketing or advertising, or produce better decision-making. Whilst big data and data analytics can thus stimulate innovation in the private and public sectors, they also pose challenges that policymakers must address both to protect citizens and to realise the considerable potential for social and economic development.

What are big data and data analytics?

Broadly speaking, *big data* refers to data sets so large and complex that they are difficult to process using traditional ICT applications. Most definitions focus on three 'V' words. They describe big data as being large (i.e. they have 'volume'); heterogeneous (i.e. they come from different sources in a 'variety' of different forms including unstructured data such as text and emails); and collected or analysed in near real time (i.e. they exhibit high 'velocity'). Some definitions also emphasise the need for data to be trustworthy (i.e. they must demonstrate 'veracity').

Data analytics refers to the techniques and processes that are applied to data, in particular big data, in order to reveal patterns and correlations. They are used to extract, from the raw data, information and knowledge that can be used in making decisions, improving productivity or developing innovations.

The rise of big data and data analytics

Even as the contemporary world becomes increasingly digital, and data volumes increase, opportunities for exploiting data will grow: the percentage of data that would be useful for analysis will rise from 22% to more than 35%. The worldwide market in big-data technology and services is expected to increase at a compound annual growth rate of about 23% between 2014 and 2019, and worldwide revenue for big data and business analytics has been forecast to increase more than 50% from almost US$122 billion in 2015 to more than US$187 billion in 2019. The largest sectors for big data include manufacturing, banking and insurance, government, professional services, telecommunications, health, transport and retail.[1]

Europe has been slow to adopt big data compared to the United States. More than half of worldwide revenue from big data is expected to come from the USA, and only one in twenty top big data companies is European. However, in Europe there is an active data market (including information and IT services, as well as companies that sell and buy data) that represented almost €55 billion in 2015, and that has been growing about 7% annually.[2] About 70% of that data market is concentrated in five Member States, namely Germany, France, Italy, Spain and the United Kingdom.

The impact of big data analytics

Nine out of ten global business leaders in a 2012 survey believed that data had become a 'new factor of production', as fundamental to business as physical assets, labour or capital. According to business guru Michael Porter, 'in many businesses, [data] is perhaps becoming the core asset'. About half of American business decision-makers surveyed in 2015 expected to gain a future competitive advantage for their companies through analysis of large quantities of data.

Although it is difficult to generalise, studies suggest that companies that embrace data-driven innovation have 5%-10% greater productivity growth than companies that do not. The UK's House of Commons Science and Technology Committee concluded that if more British companies analysed data, UK productivity could increase by 3%. According to one economic model, applying big-data analytics in the EU between 2014 and 2020 should result in a 1.9% increase in GDP. However the impact of big data is not restricted to the private sector. Full use of data analytics in Europe's 23 largest governments could reduce administrative costs by 15%-20% through greater efficiency, increased tax collection and fewer frauds and errors.

Examples of big data analytics are found in a wide range of sectors:[3]

- In **insurance**, predictive analytics software has been used to identify likely cases of public health insurance fraud before claims are paid, with the result that the programme saved three times what it cost in just its first year.
- In the **retail** sector, a UK supermarket chain accesses huge data flows through its consumer loyalty cards, processing more than 6 million transactions a day to provide tailored services to consumers.
- In the **transport** area, navigation companies collect billions of traffic measurement points daily; using this data to reduce congestion could result, by 2020, in worldwide savings of US$500 billion in time and fuel, and 380 megatons of CO_2 emissions.
- **Utilities** use data analytics to identify overall consumption patterns in order to forecast future demand and adjust prices and production capacities.
- Researchers analysing full **genetic data** from 35 000 people were able to identify a genetic variant related to schizophrenia that would have been difficult or impossible to pick out of smaller samples.
- Using big data analysis of terabytes of brain image data collected over thirty years by 30 institutions, a study into the progression of Alzheimer's **disease** was able to map five critical factors in 78 distinct regions of the brain, and identify the first sign of onset.
- Big data analytics can be used to complement the collection and analysis of official **statistics**: road traffic data can provide rapid indicators of the level of economic activity; internet-site prices can provide up-to-date information on inflation; and social media chatter can serve to estimate levels of consumer confidence on a weekly basis.

Despite many examples like these, a number of observers remain sceptical that big data and data analytics can deliver all these promised benefits. Low-quality data, biases in the data or errors in the analysis can all lead to incorrect or misguided conclusions. For example, Google abandoned the attempt to use search data to flag outbreaks of influenza when it discovered that searches for 'flu' from people who had read news stories (not just those suffering from symptoms) could result in significant overestimates of infection rates. Technical concerns include the potentially limited scope of data, problems of accessing the information, and poor interoperability. Some observers argue that data based on online transactions may not be a good model for offline 'real world' behaviour, and that rapid changes in the digital environment may mean that predictions based on past data are unreliable guides to future behaviour. Others believe that analytics will provide a sustainable competitive advantage to companies only if the data is inimitable, rare, valuable, exploitable and non-substitutable (conditions which big data may not, in many cases, fulfil).

Challenges for big data and data analytics

Big data raises a number of issues for public policy-makers. These include questions about how the value of data should be determined to establish economic statistics or calculate tax, or whether the lack of standardisation in big-data formats will create anti-competitive situations and reduce consumer value. However, sources generally agree that the main policy issues related to big data analytics are privacy and personal data protection, data ownership, barriers to the free flow of data, skills gaps in labour markets and an emerging new digital divide.

Privacy and personal data protection

The 2016 General Data Protection Regulation (GDPR), which will apply from 2018, has reinforced EU data protection standards, considered by the European Commission to be the highest in the world. Data subjects must agree to data collection and consent to the purpose for which the data is used. Whilst restrictions on the flow of personal data within the EU are banned, there must be adequate protection measures in place before personal data can move outside Europe's borders.

Open data

'Open data' describes data which is public, accessible at no or low cost and which can be reused or redistributed freely. The category overlaps with that of big data: not all open datasets are big (e.g. data for a small research study), and not all big data is open (e.g. online transactions of millions of clients of a commercial online platform). However large quantities of open data can be analysed in the same way as big data sources, or used to enrich other big data (even private data) to provide new insights.

Much open data comes from government bodies, where the main cost of collecting the data has already been assumed, as well as from publicly funded scientific research. While releasing open data forgoes possible revenues from selling data in a data market, the potential benefits to society are commonly considered to be more important. The EU's Public Sector Information (PSI) Directive of 2003, as amended in 2013, sets out minimum rules in regard to open data.

A large proportion of big data is not personal (e.g. weather information, satellite imaging, operational machine data); however some big data may include elements that link directly to a person (e.g. name, address, card or phone numbers) and hence be considered personal data. Even when this data has been aggregated and 'pseudonymised'[4] to remove explicit identifiers, analytical techniques applied to very large data sets make it technically possible to 're-identify' a person a large percentage of the time.[5] The danger is that the use of this personal big data can lead to surveillance, unwanted disclosure of private information and discriminatory profiling. Maintaining privacy depends on appropriate security measures to protect the big data sources and the commitment of the data curators (including third parties to whom data may be transferred) to ensure misuse does not happen. It should be noted, however, that some have argued that legal measures to ensure privacy cannot be relied upon to prevent post-collection misuse.

The European Commission argues that the high standards set in the GDPR act as a competitive advantage for European businesses, in as much as they foster trust on the part of citizens who are consequently more willing to share data. Other observers, however, believe that European companies are losing out in the application of big data to American firms because of stricter regulation, and that strong privacy rights for citizens in terms of personal data may in fact prevent the realisation of the potential benefits from big data, as costs will outweigh efficiency gains.[6] Restrictions on the use of personal data – to a specific purpose, context or for a limited period – could be problematic if they are interpreted strictly, since the benefits of big data may come from unanticipated uses long after the original purpose has been served, and 'public interest' exceptions are limited. Other observers have noted that notions of privacy change from one generation to another: achieving the right trade-offs between privacy/security and innovation/ convenience may require a better understanding of privacy in the digital world.[7]

Data ownership

'Ownership' is considered by a number of observers to be an overly simple or impractical concept in relation to big data. Big data does not have one owner, but typically comes with a complex set of rights and privileges associated with different stakeholders. For example, large quantities of technical data are generated in a 'smart car' but what rights to that data accrue to the owner of the car, the driver, the dealer that sold the car or the automobile manufacturer? Providing guidelines and model arrangements could increase legal certainty for businesses, simplify the conclusion of contracts, and facilitate the use of big data to improve outcomes. On the other hand, complex regulations that hinder companies from buying, selling or exchanging data could have a negative effect on applying big data to solve real world problems.

Ownership and rights also need to be seen in the framework of innovation and competition policies (e.g. abuse of dominant position and risks of consumer lock-ins) and against the background of data flows in the single market. The Belgian secretary of state for privacy recently argued that governments should be free to sell anonymised patient data to pharmaceutical companies, as long as patients can see what has been done with their data and they end up benefiting from the resulting reduced costs and innovative medicines.

Data localisation and barriers to the free flow of data

Data localisation refers to requirements to store and process data within a set of geographic boundaries. Governments may adopt data localisation measures to safeguard data privacy or to promote local digital industry.[8] The General Data Protection Regulation prohibits data localisation measures for personal data inside the borders of the EU, but bans the export of personal data outside the EU unless it is protected by comparable measures abroad. However data localisation can apply to non-personal data, including both inherently non-personal data such as climate information, and data that has been aggregated or pseudonymised. For example, plans for a consolidated German federal IT infrastructure (the *Bundescloud*) indicate that any protected information, such as business secrets, must be exclusively processed in Germany. Data localisation can also result from regulations on the handling of data that act as barriers to the free flow of data, such as requirements to obtain the consent of data subjects, the rights of subjects to review data for accuracy, and legal obligations to notify security breaches. Indeed some stakeholders have seen data localisation not as a policy yielding concrete results but rather as a way of increasing pressure on other nations to increase their own data protection measures.[9]

Nevertheless, data localisation policies are generally considered to run counter to the design of the internet, to decrease information security by concentrating storage in one place, and to increase costs for local businesses due to loss of scale, while only offering a limited incentive to local development. They also run counter to liberal trade measures, since data constitutes a significant part of international trade in goods and services. The total effect of regulatory barriers to the free flow of data in the EU is considered by some economists to have a negative impact on growth of between 0.4% and 1.1% of GDP.

The labour market and skills gaps

IDC estimates that in the EU, there are 6 million data workers who collect, store, manage or analyse data as their primary activity, representing about 3% of the workforce.[10] In IDC's baseline forecast for the future, by 2020 there will be some 7.1 million data workers, a cumulative 4.2% increase over 2015. Nevertheless, there is a significant mismatch between the demand and supply of data workers as almost 400 000 jobs go unfilled.

To support the right mix of skills and competence for data scientists, the OECD believes that further investments in both education and workplace training are needed in areas such as mathematics, statistics, computer science and data-intensive scientific fields such as experimental physics, molecular biology and bioinformatics. Other observers emphasise that, to make best use of big data, diverse skills are needed, not just for data scientists, but also for managers who can combine quantitative training with an understanding of human behaviour and social systems in order to apply the insights that big data analytics can provide. In addition to the 2016 New Skills Agenda and its Digital Skills and Jobs Coalition which target digital skills, the EU is helping to fund a European Data Science Academy to support data science training and education.

A 'data divide'

Some observers also believe that we are witnessing the beginning of a new digital divide. Citizens may not regularly use ICTs (either directly or indirectly through the use of 'smart' devices which they may find expensive) to communicate with others, purchase products, monitor their health or manage their homes. These people will not have data about their communications, purchasing patterns, health or behaviours collected and analysed. As a result of this 'data poverty', groups of citizens may increasingly find their voices and values under-represented when it comes to designing new products or services or to developing government policies. To achieve a balance that reflects all of society and that bridges this 'data divide', governments need to ensure that data collection extends to hard-to-reach communities, that educational programmes promote digital literacy, and that technology support is considered in community funding.

European research

The EU has earmarked more than €500 million from the Horizon 2020 research framework programme for big data research over a five year period (2016-2020). These funds will be leveraged by investments of private firms (including small and medium-sized enterprises) participating in a contractual public-private partnership (cPPP) to reach a total of approximately €2 billion in research funds. Five research calls specifically dealing with big data were launched for the 2016-2017 period. These focus on fostering the exchange and integration of data across sectors, producing pilot actions in data-intensive sectors, improving fundamental data technologies and architectures, and improving the education of data professionals. In addition, the Picasso project, set up in 2016 to encourage cooperation between the EU and the USA in ICTs, has created a Big Data Expert Group that will analyse the status of big data in Europe, the USA and around the world and identify opportunities for future collaboration between the EU and the USA.

EU policy

In October 2013**, the European Council** described big data as 'an important enabling technology for productivity and better services', and called for the right framework to create a single market for big data, including a network of national digital coordinators from the Member States.[11]

In 2014 the **European Commission** set out its initial approach towards a data-driven economy, including measures to monitor the European data market. In its 2015 communication on the Digital Single Market, the European Commission highlighted the potential of big data to maximise growth and stimulate competitiveness. While stressing investment in ICT infrastructure and technologies in order to maximise the growth potential of the digital economy, the Commission admitted that restrictions on data location, doubts about conditions for data re-use, and barriers to data portability contribute to market fragmentation.

In 2014, the **Article 29 Data Protection Working Party**, an EU advisory body, expressed some scepticism concerning the real value of big data, while asserting that EU data protection principles were still applicable. In 2015, the **European Data Protection Supervisor**, an independent EU authority, recognised the tension between innovation and protecting rights but stressed the need for user control, transparency and accountability in big data use.

The **European Parliament**, in a 2015 resolution in response to the Commission's 2014 communication, affirmed that the creation of a data-driven economy lay at the heart of the single market and welcomed a proposal on the free flow of data. In considering the Digital Single Market in 2016, Parliament called on the Commission to carry out a review on big data, and reiterated its call for a free flow of data initiative that would clarify rules on use and ownership of data, data localisation and data portability.

Stakeholder positions

EDRi, a digital rights advocacy group, argues that big data which can be combined with other data sources has fundamentally changed the nature of personal data. **BEUC**, the European consumer association, is anxious that issues of data flows and data localisation between the EU and third countries like the USA are discussed outside the context of trade negotiations, so as to avoid any risk of weakening consumers' rights to privacy. On the other hand, **BusinessEurope**, which represents European business federations, believes that a 'free flow of data' initiative is needed to prevent obligations to store data in a specific country hindering business operations, and that Europe needs to invest in developing skills such as data analysis. **DigitalEurope**, an association representing the digital technology industry, believes that big data presents enterprises with opportunities to improve performance, develop new business models and improve products and services. However they hold that the GDPR did not provide a harmonised approach to data but instead introduced new restrictions, rather than incentives, for big data applications.

Next steps

The European Commission has indicated that it plans to tackle unjustified restrictions on the free movement of data within the EU, including issues of data ownership, interoperability, usability and access. Although some observers have difficulty in imagining how all these different aspects can be dealt with in a general legal text, an initiative from the Commission is expected in November 2016.

Main references

Bauer, M. et al. The costs of data localisation: friendly fire on economic recovery, European Centre for International Political Economy (ECIPE), 2014.

Bonneau, V., *et al*., Big data: economic business and policy challenges, Special issue of the Digiworld economic Journal 97, 2015.

Executive Office of the President, Big data: seizing opportunities, preserving values, The White House, Washington, 2014.

OECD, Exploring data-driven innovation as a new source of growth, Digital economy papers, No 222, 2013.

OECD, Data-driven innovation: big data for growth and well-being, 2015.

European Parliament, DG IPOL, Policy Department C, study Big data and smart devices and their impact on privacy, 2015. PE 536.455.

Science and Technology Committee, House of Commons (UK), The big data dilemma, 2016.

Endnotes

[1] For further information on the economic impact of big data, see EPRS infographic.

[2] For comparison purposes, the EU market in all ICTs represents about €720 billion annually.

[3] Where sources are not linked, the examples are from OECD, Exploring data-driven innovation as a new source of growth, OECD digital economy papers, No 222, 2013.

[4] Pseudonymisation is processing such that personal data cannot be associated with a specific, identifiable person.

[5] Researchers using 3 months of credit card transactions for a million users, or 15 months of mobile phone records for 1.5 million callers, showed that three or four data points were enough to uniquely identify users about 95% of the time. See Ian Levy, quoted in De Prato, G.; J.-P. Simon (ed.), The next wave: 'big data'?, Digiworld economic journal No 97 (2015), pp. 15-39.

[6] Ciriani, S., The economic impact of the European reform of data protection, Digiworld economic journal 97, pp. 41-58.

[7] De Prato, G.; J-P. Simon, The next wave: 'big data'? Digiworld economic journal 97, 2015, pp. 15-39.

[8] Arguably, they may also do so to facilitate surveillance of their citizens. Russian requirements that citizens' personal data be processed on national territory has been condemned for being economically counter-effective and for facilitating infringement of freedom and civil rights.

[9] Gloria Gonzalez Fuster, Vrije Universiteit Brussel, as reported in the discussions on Safe Harbor 2.0 during EDRi's Privacy Camp, 26 January 2016, See *EDRi co-hosts the Privacy Camp,* published 4 February 2016 on the EDRi blog.

[10] European data market, D8 – 2nd interim report: executive summary, IDC, 2016. Available on the Datalandscape site.

[11] France, for example, has nominated an *administrateur général des données* or national 'chief data officer'.

Disclaimer and Copyright

eprs@ep.europa.eu
http://www.eprs.ep.parl.union.eu (intranet)
http://www.europarl.europa.eu/thinktank (internet)
http://epthinktank.eu (blog)