

Big Data: An Overview



'Big data' is data on a scale or of a complexity that makes it challenging to use. This POST note examines definitions of big data, how it is managed, used and regulated, and the public concerns that it generates. It sets the scene for a series of briefings exploring how data are increasingly used across a range of sectors.

Background

Unprecedented quantities of data are being collected, stored, copied and analysed from a variety of sources. In 2013 there were an estimated 4.4 trillion gigabytes (GB) of data globally,¹ equivalent to approximately 120 DVD movies for every person on the planet (1 GB is around 1 billion pieces of data). The total amount of global data is predicted to grow by about 40% year on year for the next decade. This rapid increase in the availability and complexity of data has led to the term 'big data', although this has no universally-agreed definition (Box 1).²

Broadly speaking, big data describes data with characteristics that make data collection, processing, analysis or interpretation a challenge; often requiring the use of innovative techniques. The process of seeking insights by examining big data is referred to as big data analytics.³ Much has been made of the opportunities that big data presents, although it remains unclear to what extent these will be realised. Some argue that it is not a new concept, but a frontier that moves with improving data storage and computer processing.⁴

Drivers of Big Data

Rapid growth in the acquisition, production and use of data has been attributed to a range of technological, societal and economic factors. Technological factors include the creation

Overview

- Increasing quantities of data are being collected and analysed, producing new insights into how people think and act, and how systems behave.
- This often requires innovative processing and analysis known as 'big data analytics'.
- These are being applied in sectors such as business, crime, security, healthcare, transport, the utilities and research.
- Big data analytics have the potential to personalise products and services, improve efficiency and create new jobs.
- Concerns include data security breaches, privacy infringements, discrimination, and job losses due to increased automation.
- A new EU Data Protection Regulation is currently being debated.

Box 1. What is Big Data?

Key characteristics of the scale and complexity of 'big data' include:²

- High **volume** – large quantities of data that can range from terabytes (1000 GB) up to petabytes (1 million GB) in size. High volume data sets, such as the Google Search index which is used to recommend web pages, are far too large for a conventional spreadsheet or home computer and often contain many variables.
- High **velocity** – data that must be collected and analysed rapidly to be useful, often at the same rate as they are collected (in 'real-time'). Examples include applications that analyse stock-market data or military surveillance footage.
- High **variety** – data consisting of many different forms, often brought together from varying sources. For example, medical records containing everything from GP records to X-ray images.

Other big data characteristics have been suggested including, **veracity** (if data may be inaccurate or unreliable) and **variability** (data with qualities that change, for example over time). Big data can also involve using newly-collected data or finding new ways of using existing data, for example by applying novel analysis methods, by linking together data sets to give a fuller picture, or by asking previously un-thought of questions.

of new data sources, such as smart phones, and increasing capacity to store and analyse data (Box 2). Among the key societal factors driving big data is the wide-spread adoption of new forms of communication through social media (such as Facebook, YouTube and Twitter), which are the subject of a current Select Committee inquiry⁵ and POSTnote 460.

Box 2. Technological Factors Driving the Growth of Big Data

New sources of data are being created through:

- digitization of existing processes and services, for example online banking, email and medical records
- automatic generation of data, such as web server logs that record web page requests
- reduction in the cost and size of sensors found in aeroplanes, buildings and the environment
- production of new gadgets that collect and transmit data, for example GPS location information from mobile phones and capacity updates from 'smart' waste bins (POSTnote 423).

Enhanced computing capabilities driving big data include:

- improved data storage at higher densities, for lower cost
- greater computing power for faster and more complex calculations
- cloud computing (remote access to shared computing resources via a device connected to a network), facilitating cheaper access to data storage, computation, software and other services
- recent advances in statistical and computational techniques, which can be used to analyse and extract meaning from big data
- development of new tools such as Apache Hadoop (which enables large data sets to be processed across clusters of computers) and extension of existing software, such as Microsoft Excel.

The opening-up of non-personal data in the public sector is another driver. This ranges from greater sharing of resources, such as software and research data in academia (POSTnotes 397 and 414), to increased use of data collected by government bodies to improve accountability and generate economic benefits.⁶

Estimates suggest that between 2012-17, use of big data could contribute £216 billion to the UK economy via business creation, efficiency and innovation, and generate 58,000 new jobs.⁷ Benefits in the public sector could include: data-sharing across departments to reduce repeated data-entry; collection and analysis of more detailed management data to improve performance; automatic identification of potential errors and fraud in tax credit claims; and applying information about users to tailor and target services and products.⁴

Big Data Funding and Research

The UK Government has identified big data as one of its 'eight great technologies', with the potential to drive future UK economic growth. Since 2012, it has announced capital investments of £189 million to support the use of big data and improve the UK's data infrastructure (Box 3), alongside funding for project grants and skills training. The European Commission recently outlined ways to promote a European data-driven economy,⁸ and is also supporting big data research through Horizon 2020 and the Seventh Framework Programme for Research and Innovation.

Management of Big Data

Making use of any kind of data requires data collection, processing and analysis, followed by interpretation of results. Specialised skills are needed to address the challenges involved, which may be exacerbated by the scale, complexity or speed of big data.

Box 3. Publicly-Funded UK Big Data Research

Big data capital investments via UK Research Councils include:⁹

- **Arts and Humanities Research Council** – projects to make academic data available to the public
- **Biotechnology and Biological Sciences Research Council** – e-infrastructure and training for bioscience researchers
- **Economic and Social Research Council** – providing social science researchers with access to big data from business and government, and extending the value of large-scale surveys
- **Engineering and Physical Sciences Research Council** – supporting research in data science and big data analytics and providing infrastructure to store, manage and process data
- **Medical Research Council** – bioinformatics research bringing together science, engineering and maths to process biological data, for applications that include understanding human disease
- **Natural Environment Research Council** – developing the infrastructure needed for open access to big data; providing computers capable of running complex environmental models; and capturing real-time data from sensors
- **Science and Technology Facilities Council** – creating the Square Kilometre Array radio telescope to address fundamental questions about the universe.

Collection

Big data can be acquired in myriad formats from a vast, and increasing, number of sources. These include images, sound recordings, user click streams that measure internet activity, and data generated by computer simulations (such as those used in weather forecasting). Key to managing data collection are metadata, which are data about data. An email, for example, automatically generates metadata containing the addresses of the sender and recipient, and the date and time it was sent, to aid the manipulation and storage of email archives. Producing metadata for big data sets can be challenging, and may not capture all the nuances of the data.

Processing

Data may undergo numerous processes to improve quality and usability before analysis, including:

- **extraction** – pulling out required information from the initial data and expressing it in a structured form
- **cleansing** – detecting and then correcting or removing corrupt or inaccurate records
- **standardization** – formatting data to aid interoperability
- **linkage** – connecting records from different sources.

These processes can be more difficult when applied to big data. For example: it may contain multiple data formats that are difficult to extract; require rapid real-time processing to enable the user to react to a changing situation; or involve the linkage of different databases, which requires data formats that are compatible with each other.

Analysis

Analytics are used to gain insight from data. They typically involve applying an algorithm (a sequence of calculations) to data to find patterns, which can then be used to make predictions or forecasts. Big data analytics encompass various inter-related techniques, including the following examples.

- **Data mining** identifies patterns by sifting through data. It can be applied to user click streams to understand how customers use web pages to inform web page design.
- **Machine learning** describes systems that learn from data. For example, a system that compares documents in two different languages can infer translation rules; human correction of any errors in the rules can result in the system learning how to improve the software.
- **Simulation** can be used to model the behaviour of complex systems. For example, building a trading simulation can help to assess the effectiveness of measures to reduce insider trading.

Interpretation

For the results of analysis to be useful, they need to be interpreted and communicated. Interpreting big data needs to take context into account, such as how the data were collected, their quality and any assumptions made.

Interpretation requires care for several reasons:

- Despite being large, a data set may still contain biases and anomalies, or exclude behaviour not captured by the data.
- There may be limitations to the usefulness of big data analytics, which can identify correlations (consistent patterns between variables) but not necessarily cause. Correlations can be extremely useful for making predictions or measuring previously unseen behaviour, if they occur reliably. However, they may also be misleading. For example, knowing that lots of people in an area are searching online for information on flu might be useful for targeting sales of flu remedies, but may not be a reliable predictor of a new flu epidemic.
- Techniques can be reductionist and not appropriate for all contexts. For example, some researchers have argued that big data is disconnected from social context and unable to capture the subjective experiences of individuals.¹⁰ Opinions also differ on how well modelling can be used to predict dynamic social systems.^{11,12}

Communicating the results of big data can be difficult. New tools are being developed to facilitate this. For instance, UK companies are using visualization software such as Spotfire and Qlikview to help employees interpret data.

Infrastructure and People

Managing and making sense of big data requires a combination of specialist skills and knowledge – such as mathematical, statistical and computer programming techniques – as well as more general communication skills and field-specific knowledge. It may be difficult to find all of these in one person, so multi-disciplinary team-working is a common approach. Indeed, organisations are reporting a shortage of people with the relevant skills. For example, a survey of UK companies implementing big data analytics found that 57% had difficulty filling roles in 2012,¹³ and forecasts suggest that demand for big data staff will grow by between 13% and 23% per annum from 2012-17.¹⁴ Competition for skilled workers from the private sector may inhibit Government implementation of data analytics.¹⁵

While big data has the potential to create new jobs, it (along with other technologies such as mobile robotics) may also reduce the need for some job roles. For example, a study has estimated that, over the next 20 years or so, it may be technologically possible to automate nearly 50% of all US jobs. Transport, logistics, administrative support, services, sales and construction were highlighted as most at risk.¹⁶

Applications of Big Data

There has been a long history of research projects that push the limits of data management and analysis. In the early 2000s, researchers working on the Human Genome Project overcame huge computational and information management challenges to produce one of the world's first gigabyte databases. Today, researchers analysing data from particle physics experiments at CERN's Large Hadron Collider sift through 15 petabytes (approximately 16 million GB) of data every year. When the Square Kilometre Array radio telescope becomes operational in the mid-2020s, it is anticipated to generate many petabytes of data per second, driving substantial developments in computing and data processing techniques.

Big data is being used across a wide range of fields, including in business (POSTnote 469), crime and security (POSTnote 470), smart metering (POSTnote 471), transport (POSTnote 472), biomedical research (POSTnotes 473 and 474) and citizen science (autumn 2014). As new applications develop, it may also play an increasing role in other fields. For example, the Economic and Social Research Council (Box 3) is investing £34 million in the Administrative Data Research Network, which aims to enable safe access to government data for accredited researchers undertaking approved projects. In the educational sector, data about how students interact with on-line teaching materials is being used to understand how individuals learn and to personalise courses accordingly. Big data approaches are also being used in construction to model the likely impact of planned buildings, and in public policy where local authorities are looking to use linked data sets to plan the provision of Council services.

Public Perception and Concerns

The collection, storage and processing of personal information is regulated by UK and EU data protection laws (Box 4). Big data may contain personal information, such as medical records or genome data. Research on public attitudes to the use of personal data has identified privacy, security and discrimination as key concerns.¹⁷ For example, workshops looking at public views on data use found participants felt that the security of their personal data was very important, but that they had little control over it. Another study revealed concern about the potential for discrimination (for instance if employers gained access to data about mental health or HIV status). Good governance and regulation of big data can improve data quality and the reliability of results; ensure that legal requirements regarding data privacy are met; and encourage public interest in, and involvement with, big data projects.

Box 4. Data Protection Legislation**Data Protection Act 1998**

In the UK, big data sets containing personal data are subject to the Data Protection Act 1998 (DPA), which implements the EU Directive 95/46/EC. The DPA is based on eight core principles. Data should be: fairly and lawfully processed; processed for limited purposes; sufficient and relevant; accurate; not stored for longer than is necessary; processed in line with data subjects' rights; secure; and transferred only to countries with adequate security.

Fair processing requires that data subjects are informed of the identity of the data controller and the purposes of the processing. Health data may only be processed if explicit consent is obtained or if the processing is necessary for one of several defined conditions. One such specification is medical purposes undertaken by a health professional or person with an equivalent duty of confidence.

The EU Data Protection Regulation

EU Directive 95/46/EC is widely recognised as being outdated and is scheduled to be superseded by a new Data Protection Regulation (POSTnote 469). There are concerns that the version agreed by the European Parliament (but not expected to be approved by other European institutions until 2015) will impose administrative burdens on researchers and contains no clear provision for the use of identifiable data without consent (POSTnotes 473 and 474).

Privacy

The increase in the number of devices capable of collecting data means that data are being captured from previously private aspects of life. For instance, facial recognition technologies can identify people from pictures online, and GPS-enabled mobiles can track a person's location. There is the potential for infringements of privacy to occur if data are used for purposes other than those for which they were collected; for example, if data are collected from sensors without people's knowledge, if different data sets are linked together, or if data are unexpectedly sold to third parties.

Applying data protection tools before data analysis can help protect the privacy of individuals. These include various de-identification processes to remove identifying details from data (POSTnote 474). However, there is a risk that an individual may be re-identified if a de-identified data set is linked to another data set that contains identifiable data.¹⁸ A key focus here is the robustness of the de-identification process. Studies testing the effectiveness of robust de-identification found that re-identification was possible, but difficult and time-consuming.¹⁹

Data protection legislation attempts to create a balance between protecting individual privacy and allowing data use for purposes (such as some research) that are in the public interest. Privacy by Design is an approach that aims to meet the requirements of both data use and privacy, for example by building privacy protections into the design of data analytics technology. This approach is endorsed by the UK Information Commissioner's Office, which provides guidance on data protection and security breach management.

Box 5. Examples of Data Security Techniques

A variety of tools and procedures can help make data more secure.

- Implementing effective data governance processes across an organization (and external partners) can help to control access to, and use of, data. This may involve making individuals accountable for data security, providing relevant training, minimising the number of people with access and deleting data when appropriate.
- Encryption can make data more secure. For example, a message can be translated into an unreadable code for transmission, then later converted back into a readable form using a decryption 'key'.
- Multiple parties can contribute information to a communal activity without having to exchange data with each other directly. For example, countries wanting to calculate the risk of their satellites colliding without revealing exact positions can transmit data to a secure algorithm that automatically computes the risk.
- Access to sensitive data may be provided via a 'safe haven'. This can be a secure location or set of administrative arrangements to assist safe and secure communication of confidential information. The Clinical Practice Research Datalink allows accredited researchers to access records from GP practices for research.

Security

As with all data, the security of big data sets may be compromised if they are accessed by people without permission, or used inappropriately by rogue individuals within an organization. A range of tools and procedures can be used to reduce the risk of data being accessed without permission (Box 5). However, it is impossible to guarantee that data will be completely secure. Big data analytics may also help to improve system security, for instance by analysing very large log files to track patterns that may indicate attacks and unauthorised access to data.

Discrimination

Unless care is taken, the use of big data can lead to unintended discrimination.¹⁸ For example, a recent US report noted that an evaluation in 2009 of a database used by US employers to confirm the eligibility of new employees found that it took longer to clear non-US citizens. The same report also noted that big data may facilitate differential pricing, where individuals are offered different prices for online products depending on how 'affluent' they appear to be. However, big data technologies may also help some groups to enforce their rights by identifying and confirming instances of discrimination.¹⁸

Endnotes

- 1 *The digital universe of opportunities*, 2014, IDC, bit.ly/1jvQFck
- 2 Schroeck, M. et al., *Analytics: the real-world use of big data*, 2012, IBM
- 3 Yiu, C., *The big data opportunity*, 2012, Policy Exchange
- 4 Manyika, J. et al., *Big data: the next frontier*, 2011, McKinsey
- 5 *Social media data & real time analytics*, HoC Sci. & Tech. Com. bit.ly/1eMJcEK
- 6 *Improving the transparency & accountability of gov.*, 10/07/14, bit.ly/1hDCHG2
- 7 *Data Equity: unlocking the value of big data*, 2012, Cebr
- 8 *Towards a thriving data-driven economy*, European Commission, 02/07/14
- 9 Big data, RCUK, accessed 24/06/2014, bit.ly/W1CHbF
- 10 Hetherington, S. et al., *The thin data revolution*, 2014
- 11 Uprichard, E., *Big data, little questions?* 2013
- 12 Moat, H. S. et al., *Behavioral and brain sciences*, 37, 92-93, 2014
- 13 *Big data analytics adoption and employment trends*, 2013, e-Skills UK, SAS
- 14 *Big data analytics assessment of demand for labour*, 2013, e-Skills UK, SAS
- 15 *Market assessment for public sector information*, 2013, Deloitte, BIS
- 16 Frey, C. B. et al., *The future of employment*, 2013, University of Oxford
- 17 *Big data, public views*, 2014, Sciencewise, bit.ly/1rZyXV7
- 18 *Big data: seizing opportunities, preserving values*, 2014, The White House
- 19 Cavoukian, A. et al., *Big data and innovation*, June 2014, bit.ly/1uyQ9ha