Chapter 5

# Big Data/Big Data Analytics

# Big Data/Big Data Analytics

**Key points**

- Big data/big data analytics is the lifeblood of AI, because it is what fuels AI algorithms.
- Big data is evaluated in terms of its volume, velocity, variety, veracity and value.
- Financial institutions use big data/big data analytics for activities ranging from marketing to credit assessment.
- Big data/big data analytics introduce new risks in financial services if not partnered with digital identification systems, because their widespread adoption could otherwise exacerbate financial exclusion.

## 5.1 Introduction

In all walks of life, individuals, organisations and governments rely on data to help them to make good decisions.[1] In the last 30 years, vast amounts of data have come to be collected on every possible aspect of modern life and these large datasets are now known as 'big data'—a phrase that many people began to use a decade or two ago without fully understanding what it meant.

While the amount of data collected and stored has increased exponentially, development of the skills and technology that allow us to analyse and use this data has gathered momentum more recently. Big data is the lifeblood of artificial intelligence (AI), for example. Without large datasets, AI models and algorithms cannot be refined and function effectively or accurately.

With detailed data that touches on almost every aspect of our lives never more readily available, we must be cautious of its potential to fuel both positive and negative outcomes. Solid governance of the collection and use of that data is critical if we are to ensure it is used only for positive ends, to enhance decision-making and to protect individuals' rights.

## 5.2 Context

The history of data analytics goes back to 18,000BCE, with evidence that Palaeolithic peoples marked notches into sticks or bones to keep track of and compare trading activity and supplies. In 1663, John Graunt conducted what is believed to be the first statistical analyses in trying to develop an early warning system for the bubonic plague.[2]

In 1928, the method for storing information magnetically on tape was invented, which led later, in 1965, to the first large data centres. These were limited, however, by the fact that data was at that time recorded and stored in physical form. In 1996, shortly after the birth of the internet, electronic storage became more cost-effective than paper storage, and in the early 2000s—as the capacity to store data skyrocketed and the need to adapt analysis to suit its volume and scope become obvious—the term 'big data' was coined.[3]

In 2010, chair of Google Eric Schmidt told a conference that as much data was now created every two days as had been created between the beginning of human civilisation and 2010.[4]

In 2014, consultancy IDC projected that, globally, there would be more than 44 zettabytes of data generated by 2020 compared with 4.4 zettabytes in 2013.[5] A zettabyte is a difficult measure to visualise: it is $2^{70}$ bytes—or as much data as can be stored on 250 billion DVDs. A gigabyte—1 million bytes—is a common unit in measuring computing memory: it has been said that if each gigabyte in a single zettabyte were a brick, those bricks would be enough to build 258 Great Walls of China.[6]

**As our ability to generate, collect and store information has grown, so has the potential for us to generate richer insights from this data. It has been clear that holding large and rich datasets alone are not an end in itself; the key to the value of data lies in its analysis.**

There are three key terms that we can define here.

- **Big data** is used in many different ways—to refer to large datasets, and to refer to the exponential increase of data and availability of data in the world today.[7] Big data is said to display the following '5Vs'.
    - *Volume* Big data—the size of the dataset—has to be large. There is no set agreement on how large 'large' is; it is relative and it is ever increasing.
    - *Velocity* This refers to the speed at which data is collected and analysed.
    - *Variety* With increasing volume and velocity comes increased variety. A dataset with wider variety can lead to richer insights.
    - *Veracity* This refers to the quality of the data: is it 'clean' and accurate?[8]
    - *Value* By this, we mean whether the data and its analysis lead to meaningful insights and inform good business decisions.[9]
- **Data science** is the field of studying data. The goal of data science is to improve decision-making through the analysis of data.[10]
- **Data analytics** is the multidimensional field that uses mathematics, statistical modelling and machine learning to find meaningful patterns in data.[11]

As technology evolves, so does our ability to collect, store and analyse data—but all of these processes are dependent on advances in hardware and software development, which act as the key enablers in a big data ecosystem.

While the positive implications of big data are vast, including enhanced decision-making, predictive technology and increased profitability, the potential for abuse is equally apparent. One instance emerged when political consulting firm Cambridge Analytica was found to have misused individual data, mined from Facebook, during the 2016 US presidential elections.[12] It is therefore essential that any big data ecosystem also include measures to prevent such abuse.

### 5.3 Description

The application of big data across financial services touches many aspects of our lives, ranging from payments, through lending and investment decisions, to impact more broadly on how financial services markets function. This section will present an overview of some of the relevant technologies and applications that support the use of big data in these contexts.

#### *5.3.1 Customer Segmentation and Personalised Marketing*

Big data offers financial and other services providers the ability to refine their customer segmentation at a more granular detail than was previously possible. This allows them to understand in detail what different customer segments need and it allows financial services organisations, such as credit card companies, to offer personalised marketing and tailored discounts to their users.[13]

#### *5.3.2 Credit Assessment*

Big data has given rise to alternative credit models aiming to address the role of existing credit models in financial exclusion.[14] Credit assessment procedures have long relied on regression analysis, which assumes that one behaviour predicts another, such as that companies and individuals who do not pay their loans on time will continue not to pay their loans on time. Alternative credit assessment models are now emerging that are based on an individual's education level or the reputation of their school as an indicator of whether or not they will default on a loan, while others use social media analysis.

These new models are not, however, without their challenges. Not all data is reliable for credit scoring, for example, and there are currently gaps in the law governing these new models to ensure that they are, and are used in ways that are, accurate, fair and transparent.[15]

The rise of behavioural analytics and social physics has fuelled a more robust approach to alternative modelling based on big data indicators to assess the likelihood of someone defaulting or repaying a loan.

- **Behavioural analytics** is the study of human behavioural data to identify meaningful patterns and draw inferences or make predictions based on those patterns.[16]

- **Social physics** applies the principles of physical sciences to study of how groups of people make decisions by analysing how information and ideas flow from person to person.[17]

Both of these can be applied to improve not only credit scoring, but also fraud detection, identity verification, regulatory compliance and enforcement, predictions of consumer behaviour, and stock trading and investing decisions.

#### *5.3.3 Stock and Commodity Trading*

Big data and machine learning have significantly influenced stock and commodity trading. Not only does big data improve the likely outcomes of financial services for individuals, but also the application of AI technology to capital markets reduces the barriers to entry for many individuals and widens participation in the market.[18] These types of new trading technology have fuelled adoption of electronic trading platforms and virtualised trading environments[19]—although this is not without risk. For example, algorithmic and high-frequency trading (HFT) have been known to cause flash crashes in the market, such as the 2010 US flash crash.

In foreign exchange (forex) brokerage, assessing risk is essential to successful operations. The broker must be able to see

data in real time and be alerted to specific preferences, market statuses, profit and loss, exposure and market volatility. The unified data that we can now gather from a wide spectrum of resources (Web, mobile, social, customer relationship management, affiliates, etc.) is the key to a holistic overview of platform performance on which managers can base their decisions.[20]

#### *5.3.4 Regulatory Compliance and Fraud Detection*

For financial services organisations, regulatory compliance is a key priority that typically involves significant amounts of paperwork, resulting in millions of user records. Machine learning and other AI technologies analyse this big data to detect compliance irregularities and even fraud more accurately and efficiently than ever before. In some cases, it can detect instances that would be beyond human capacity. For example, in 2017 Credit Suisse reported a 45-fold increase in productive alerts resulting from its predictive monitoring of transactions compared to the year before, and it measured resolution of the alerts as 60 per cent faster at a fraction of the historical cost.[21]

Some large financial services providers, such as JP Morgan Chase, use big data/big data analytics to detect fraud by analysing the activities of their own staff, including not only internet search histories, but also personal data including emails and call history. JP Morgan also applied big data in an optimised real-estate price-determination model for use when selling property the bank acquired as collateral on loans that are now in default (*see below*).[22]

#### *5.3.5 Other Contexts*

Other contexts in which big data can be valuable include real-estate markets. With big data/big data analytics, lenders can minimise social loss by analysing a local property market to determine the most marketable price at which a property will sell quickly, allowing a debtor to avoid insolvency.[23]

Big data also has the potential to underpin productivity and release consumer surplus. For example, McKinsey & Co. estimates that a retailer using big data can increase its operating margin by more than 60 per cent, while services enabled by personal-location data can allow consumers to capture US$600 billion in economic surplus.[24]

### 5.4 Key Considerations for Future Development

It is important to view big data not in isolation but as an ecosystem that includes the various sources from which data is gathered, the spaces in which this data is traded, the analysis of that data and the decisions that the analytics inform.

Among the key considerations in these regards is **data protection**. While effective AI systems of this type are dependent on personal data, users' rights must remain at the forefront of any policy governing big data. The leading example of regulation in this context is the General Data Protection Regulation (GDPR) of the European Union (EU): a benchmark against which policies covering data protection and privacy should be measured.[25] Its basic principles include user control of their own personal data, the requirement that users explicitly consent to others using their data and a right to be 'forgotten' (i.e., to request deletion of data).[26]

Policy considerations should also reinforce corporate responsibility for data governance. Businesses that are collecting data of any sort, but personal data in particular, are responsible—and should be held accountable—for the security and

> Big data has the potential to underpin productivity and release consumer surplus. It is estimated that a retailer using big data can increase its operating margin by more than 60 per cent.

legitimate use of this data. Any organisation gathering, storing and/or managing data must therefore put sound data governance structures in place.[27]

**Open data** can therefore be contrasted with personal data. Open data is (a) publicly available and (b) licensed for reuse, and it is ideally relatively easy to (re)use.[28] Open data is available to researchers and other organisations who will analyse it to extract the most value. Clearly, therefore, this type of data does not include personal data and data scientists must take care when selecting their sources.

In fact, in 2019 it was estimated that while the number of trained **data scientists** was increasing, demand for these skills was growing even more rapidly.[29] From a policy perspective, the recruitment and development of skilled data science professionals is consequently a fundamental component of the big data ecosystem and an effective digital economy.

Finally, while availability and types of data vary from country to country and personal data varies according to demographics including age groups, income brackets, gender and geographic locations, any policies focused on big data in financial services or elsewhere must take care to bridge the **digital divide**. We must take care to ensure that big data applications and solutions do not exclude any one or more groups, nor should we assume that our findings are universally applicable unless proved to be so.[30]

## Endnotes

1 United Nations (2020). 'Big Data for Sustainable Development' [online]. Retrieved from: www.un.org/en/sections/issues-depth/big-data-sustainable-development/index.html
2 World Economic Forum (2015). 'A Brief History of Big Data Everyone Should Read'. 25 February [online]. Retrieved from: www.weforum.org/agenda/2015/02/a-brief-history-of-big-data-everyone-should-read/
3 *Ibid.*
4 *Ibid.*
5 Adshead A (2014). 'Data Set to Grow 10-fold by 2020 as Internet of Things Takes off'. *ComputerWeekly.com*, 9 April [online]. Retrieved from: www.computerweekly.com/news/2240217788/Data-set-to-grow-10-fold-by-2020-as-internet-of-things-takes-off
6 Barnett T Jr (2016). 'The Zettabyte Era Officially Begins (How Much Is That?)'. *Cisco Blogs*, 9 September [online]. Retrieved from: https://blogs.cisco.com/sp/the-zettabyte-era-officially-begins-how-much-is-that
7 University of Wisconsin Data Science (2020). 'What Is Big Data?' [online]. Retrieved from: https://datasciencedegree.wisconsin.edu/data-science/what-is-big-data/
8 BBVA (2017). 'The Five V's of Big Data'. 8 May [online]. Retrieved from: www.bbva.com/en/five-vs-big-data/

9 Jain A (2016). 'The 5 V's of Big Data'. *Watson Health Perspectives*, 17 September [online]. Retrieved from: www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/

10 Kelleher J D, Tierney B (2018). *Data Science.* Cambridge, MA: MIT Press.

11 SAS (2020). 'Analytics: What It Is and Why It Matters' [online]. Retrieved from: www.sas.com/en_us/insights/analytics/what-is-analytics.html

12 Malan D (2018). 'The Law Can't Keep up with New Tech: Here's How to Close the Gap'. *World Economic Forum*, 21 June [online]. Retrieved from: www.weforum.org/agenda/2018/06/law-too-slow-for-new-tech-how-keep-up/

13 Kim S (2016). 'Big Data Use Cases in Finance'. *Samsung.com*, 6 September [online]. Retrieved from: www.samsungsds.com/global/en/support/insights/090617_Eng_BigData1.htm

14 Pan W, Aharony N, Pentland A (2011). 'Composite Social Network for Predicting Mobile Apps Installation'. *arXiv.org*, 2 June [online]. Retrieved from: https://arxiv.org/abs/1106.0359

15 Hurley M, Adebayo J (2017). 'Credit Scoring in the Era of Big Data'. *Yale Journal of Law and Technology*, 18(1) [online]. Retrieved from: https://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1122&context=yjolt

16 Biddle S (2019). 'Thanks to Facebook, Your Cellphone Company Is Watching You More Closely than Ever'. *The Intercept*, 20 May [online]. Retrieved from: https://theintercept.com/2019/05/20/facebook-data-phone-carriers-ads-credit-score/

17 Wladawsky-Berger I (2018). 'Social Physics: Reinventing Analytics to Better Predict Human Behaviors'. *Wall Street Journal*, 14 September [online]. Retrieved from: https://blogs.wsj.com/cio/2018/09/14/social-physics-reinventing-analytics-to-better-predict-human-behaviors/

18 Cuen L (2017). 'Fintech is Rebuilding Capital Markets, from AI to Crowdfunding Startups'. *International Business Times*, 20 October [online]. Retrieved from: www.ibtimes.com/fintech-rebuilding-capital-markets-ai-crowdfunding-startups-2559398

19 Deutsche Börse AG, Celent (2016). *Future of Fintech in Capital Markets* [online]. Retrieved from: www.deutsche-boerse.com/resource/blob/37024/ed055219caeb553f43950609d29e1bb3/data/future-of-fintech-in-capital-markets_en.pdf

20 Levy T L (2017). 'How Trading Companies are Leveraging Behavioral Analytics to Win Conversions and Retention'. *Datafloq*, 20 July [online]. Retrieved from: https://datafloq.com/read/trading-companies-leveraging-behavioral-analytics/3442

21 Credit Suisse (2017). 'How Big Data Analytics Is Transforming Regulatory Compliance'. 30 November [online]. Retrieved from: www.credit-suisse.com/about-us-news/en/articles/news-and-expertise/how-big-data-analytics-is-transforming-regulatory-compliance-201711.html

22 Kim S (2016). 'Big Data Use Cases in Finance'. *Samsung.com*, 6 September [online]. Retrieved from: www.samsungsds.com/global/en/support/insights/090617_Eng_BigData1.htm

23 *Ibid.*

24 Manyika J *et al.* (2011). 'Big Data: The Next Frontier for Innovation, Competition, and Productivity'. *McKinsey.com*, 1 May [online]. Retrieved from: www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

25 Allen C (2016). 'The Path to Self-sovereign Identity'. *Coindesk*, 27 April [online]. Retrieved from: www.coindesk.com/path-self-sovereign-identity; Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC, 4 May 2016, OJ L 119/1.

26 European Commission (2020). 'Complete Guide to GDPR Compliance' [online]. Retrieved from: https://gdpr.eu/; Violino B (2019). 'Get Ready for More Data Privacy Regulations'. *ZDNet*, 12 March [online]. Retrieved from: www.zdnet.com/article/get-ready-for-more-data-privacy-regulations/

27 Sohail O, Sharma P, Ciric B (2018). '4 Pillars to Guide Data Governance for New Platforms'. *Wall Street Journal*, 10 October [online]. Retrieved from: https://deloitte.wsj.com/cio/2018/10/10/4-data-governance-pillars-for-modern-data-platforms/

28 Maaroof A (2015). *Big Data and the 2030 Agenda for Sustainable Development* [online]. Retrieved from: www.unescap.org/sites/default/files/Final%20Draft_%20stock-taking%20report_For%20Comment_301115.pdf

29 LinkedIn (2018). *LinkedIn Workforce Report: United States—August 2018* [online]. Retrieved from: https://economicgraph.linkedin.com/resources/linkedin-workforce-report-august-2018

30 Maaroof A (2015). *Big Data and the 2030 Agenda for Sustainable Development* [online]. Retrieved from: www.unescap.org/sites/default/files/Final%20Draft_%20stock-taking%20report_For%20Comment_301115.pdf