Data in Society: Challenging Statistics in an Age of Globalisation

Chapter draft: *Data Analytics*

Ifan D H Shepherd & Gary Hearne, Middlesex University, London, UK

Introduction

This chapter sets out to illustrate the dictum that there is (almost) nothing new under the sun. More specifically, its goal is to make the unfamiliar familiar within the field of data analytics. The need for such a treatment can be gauged from the plethora of terms currently vying for attention in the contemporary data analysis landscape, which can be puzzling even for seasoned researchers. These terms include: data mining, data science, data analytics, machine learning, deep learning, neural networks, and artificial intelligence. Hybrid terms such as 'big data analytics' are also emerging. As for the current front-runner term, data analytics, the evidence provided by the number of search engine hits reveals multiple competing versions subdivided by application domains, ranging from business analytics and crime analytics, to performance analytics, visual analytics, and many more. There is also an emerging software sub-industry providing tools for data analytics, many of which are named after the company which originally developed them.

The recent rise in popularity of data analytics can be charted by plotting the number of searches for the term over time using Google Trends. This reveals that interest in *data analytics* exhibited slow initial growth between 2004 and 2011, followed by explosive and sustained growth thereafter. At the same time, interest in *statistical analysis* has declined progressively, with a crossing point between the two search terms occurring in mid-2013. Amongst other related terms that have also been displaced by data analytics is *Business intelligence*, which dates from mid-nineteenth century. *Data analysis*, however, continues to retain its popularity. In the meantime, *data science* is gaining traction in academia, perhaps because of its greater acceptability as an academic subject.

A conclusion to be drawn from various lines of evidence is that the terminology used to describe data analytics and related practices are not at all helpful in identifying what goes on in their name, nor in discriminating clearly between them. It is probably best that data analytics be thought of as a fairly heterogeneous bundle of approaches, methods and technologies which is not very easily distinguished from similar practices competing in the same field with often closely related or overlapping labels. Recently published texts on the subject suggest that much the same methods and technologies tend to appear in lists drawn up to characterise many of the related fields mentioned above. Further exploration of the origins of these socially defined terms is beyond the scope of this chapter, but their interpretation in the context of the social construction of reality (e.g. Searle, 1996) would undoubtedly repay further study.

What's new about Data Analytics?

As far as data are concerned, the distinctive features may be characterised using the concept of the 'three Vs' of big data. (This concept was first proposed by Laney (2012), though others have subsequently extended the list.). The first V, volume, refers to the huge amount of data available to analysts. The second V is variety, which refers to the heterogeneous nature of the data being harvested, aggregated and analysed. For example, data are now available in a variety of types (numerical, text, image, audio, video, etc.), on several media (CDs, USB sticks, websites, streaming feeds, e-repositories, etc.), and in various formats (plain text, csv or cdf formatted text, pdf or Word

documents, image files, video files, and numerous proprietary formats). The third V, velocity, describes the real-time availability of data, especially online. To these may be added the role of data aggregation, fusion or integration, which finds its modern expression in data mashups, though users of geographical information systems (GIS) have been integrating spatial data since at least the 1980s (Shepherd, 1991). Further distinctive characteristics claimed for data analytics and data science are their predictive rather than explanatory focus, and the more contentious view that their approaches make the scientific method and the role of theory obsolete (Anderson, 2008).

We would argue that a historical perspective is useful in assessing these claims, because the concepts and practices of data analytics are significantly older than current depictions suggest. Big data, for example, is always relative to current data handling technology, with Laney (Ibid) suggesting the following definition: "Big data is data that's an order of magnitude bigger than you're accustomed to". A similar caveat can be applied to velocity, in that near real-time data gathering, analysis and decision taking has been with us since at least the 1970s, in the form of meteorologists making rolling weather forecasts based on data gathered from multiple ground, sea, aerial and space sources.

For those familiar with 'traditional' statistical methods, the huge amounts of data that have become available in recent years calls into question two of the fundamental assumptions of statistical analysis: probability distributions, and using results from samples to make inferences about a population. The standard use of inference for frequentists is based on null hypothesis significance testing (NHST). But statistical significance assessed in this way is affected by sample size. For example, a possible random effect can appear to be significant if the data set is very large, making the test too powerful. So a different approach is required, and this has come from the emerging discipline of data mining, many of whose methods, and especially those characterised as machine learning, have been driven by people from the inter-related worlds of computing, AI and databases.

Typically, a machine learning model can look exactly like a statistical model produced by conventional means, but it is not validated by significance testing. The machine learning model is built on training software with a subset of the data, and the model is then used to test a further data subset that was not used to build it. If the model does a good job of predicting the outcome on the test data, then it is considered valid. Many machine learning methods (e.g. neural networks and support vector machines) involve algorithms which are not built on traditional statistical theory, and do not produce results that are interpretable at the variable level. For a growing number of applications, the training/testing approach to validation is entirely appropriate, even for data sets that are of a suitable size for NHST.

Most people without the appropriate training are unlikely to understand the advanced algorithms that are becoming increasingly popular in data analytics toolkits, and this can lead to potential misuse. The use of such software imposes much the same responsibility on the shoulders of the user as statistical methods have in previous generations. Users should have sufficient understanding of the nature of their data, and make appropriate decisions as to which types of analysis or models can legitimately be used with that data. They should also be expected to restrict the interpretation of their analytical results to what is appropriate and valid.

The twin technologies of automatic data harvesting and data analytics software has led to commercial companies overtaking governments in terms of their intimate knowledge of citizens. The term 'surveillance society' can therefore no longer be thought of purely in terms of governments snooping on its citizens. Over the past couple of decades, consumer-facing businesses have developed sophisticated technologies based around a business model of offering a free service (e.g. a search engine or a community sharing platform) in exchange for personal data. This goes far beyond the capture of spending data from loyalty cards gleaned by an earlier generation of retail companies,

whose operations were typically limited to a single company, region or country. It now resembles what Zuboff (2018) refers to as 'surveillance capitalism'.

Not to be left behind, many national governments have begun to adopt these same techniques. In China, for example, which sponsors and/or controls many national equivalents to western online services, including Facebook (Tencent), Google (Baidu), YouTube (Youku Tudou), Amazon (Alibaba), eBay (Alibaba), Apple (Xiaomi) and Twitter (Sina Corporation), the government is now beginning to roll out an online social credit (*xinyong*) system (Harris, 2018). As in the west, many Chinese companies have been developing credit ratings systems. However, several Chinese companies, especially Ant Financial, with its *Sesame Credit* system, have gone further, and designed rating systems based on the concept of 'social integrity'. The government has been trialling its own system, which gathers both online and offline data to assess people quantitatively on various behavioural characteristics, and accumulated e-scores are likely to be used to determine whether individuals are granted or denied access to anything from credit to travel visas. Questions remain as to how far the government dictates what 'social integrity' or 'social credit' should be based on, and how far this initiative introduces technology-driven social control on a vast scale.

The contemporary data analytics landscape

Data analytics is used in a rapidly increasing number of application domains, including marketing, political influencing, epidemiology, crime analysis and industrial quality control. Modern data analytics has three quintessential components: data, methods of analysis, and technology to support the analysis of data. This trio has evolved hand-in-hand over several centuries, but has experienced rapid and major developments in the past half century or so. This trio provides the structure for this section.

Computational devices: personal and institutional

Members of the public have access to a full spectrum of computational devices (including desktops, laptops, tablets and smartphones) on which they use, often unwittingly, data analytics software. Admittedly, the arrays of thousands of dedicated computers and servers available to online giants may be beyond their budgets, but multi-core PCs with GPU accelerators and large-capacity storage devices can be acquired for all but the most demanding analytical work. As for social researchers in academia, many have access to networked university arrays which permit them to explore large volumes of data with a battery of analytics toolsets. Similarly, smaller businesses can acquire relevant processing power in the cloud. Cumulative hardware innovations, in processor speed, data storage capacity, graphical displays and networking, have meant that analytical operations once requiring days of computer time on specialised hardware are now capable of being completed in near real-time. The impact on business decision-making cycles has been dramatic.

Digital data: proprietary and open

With the emergence of nation states since the eighteenth century, and their associated apparatus of government, the gathering of citizen data has evolved at an ever-increasing pace (Woolf, 1989). Modern data collection by CCTV cameras linked to ANPR and facial recognition software, has added near-real time tracking of individuals to the periodic collection of census data. The recent growth of private corporate databases maintained by online business giants has arguably overtaken in scale and content richness the data repositories of individual national states, with the possible exception of their surveillance arms. Moreover, online companies have amassed databases that are global in scope, and harvest data continuously from users in real time. At the time of writing, the 'Big Four' of Apple, Amazon, Facebook and Google (Galloway, 2017) have succeeded in disrupting the supply chains of entire business sectors. By leveraging the value of their users' personal data, they have created

enormous economic, social and political power based on capturing advertising revenue that previously went to traditional business operators. Salganik (2018) describes the distinguishing features of proprietary data holdings, and suggests that while these directly serve the marketing needs of these companies extremely well, only three of them are beneficial for social research.

An alternative to proprietary data, whether commercial or governmental, exists in the form of open data, which potentially provides citizens with raw material that can be interrogated with data analytic tools. For the best part of half a century, the US government has made available data whose collection was paid for by the 'federal dollar', simply for the cost of distribution. This has included remotely sensed land surface data, maps and population census data at global and national levels, and at a local level, a wealth of social data is routinely made public on local government websites, though not (as in the case of registered paedophiles' place of residence) without controversy. In the more supposedly socialist UK, in contrast, the trading status of the national mapping agency, the Ordnance Survey, has for several decades meant that its publicly financed data has been largely available on a paid-for basis. In 2010. however, after protracted lobbying, and with perhaps an eye on open-source competition in the form of OpenStreetMap (www.openstreetmap.org), the OS launched its OpenData initiative, in which it released large quantities of digital spatial data on a dozen broad topics, for use by the general public (Ordnance Survey, 2018). Many other local, regional, national and super-national governments have launched similar initiatives. This is highly significant, because the precise spatial description of features in the real world provides a universal frame of reference for much of the digital data acquired by businesses and governments alike. On the down side, however, most open data are generally collected for purposes other than for specific social research requirements, so its users often have to make the best use of what are essentially secondary data.

The issue of public trust is becoming increasingly significant in the data domain. However, this trust is not limited to whether data held by corporations and governments are correct in some definable and measurable sense. More importantly, it extends to whether specific data are: honestly and openly gathered; securely held; not combined with data from other sources without the knowledge or permission of the persons described by the data; not used for purposes other than those for which they were originally collected; and only analysed by algorithms that can be publicly verified as being free from bias (Angwin et al., 2016). On all six counts, public trust has been steadily eroding in recent years, on account of: revealed inaccuracies in official and commercial data; covert data harvesting; data leaks from online websites; fusion or integration of personal data from different sources; trading in personal data by commercial and public organisations; and revelations about racial and other biases in the automated processing of personal data. Paradoxically, perhaps, this erosion of trust appears not to have led to a reduction in the public's use of the online services provided by data-rich, online organisations.

Data analytics software: proprietary and open

Up until the 1970s, most of what we now refer to as data analytics involved 'number crunching'. More recently, however, computers have become increasingly adept at processing all kinds of information, with specialised tools being made available for text analytics, video analytics, music analytics, visual analytics, amongst many others. It is not our intention to provide an exhaustive list of data analytics software, or an evaluation of their capabilities. For this kind of information, interested readers should consult relevant web lists, and those textbooks which explore the software side of data analytics. Rather, this section will briefly describe the two main types of data analytics software currently available.

The first group consists of commercial software. Two examples are worthy of mention, because they provide easy-to-use graphical user interfaces (GUIs). SAS is a huge commercial package, both in

terms of its market penetration and its breadth of functionality. It has a range of add-ins which extend the base product, one of which, SAS Enterprise Miner, is a workflow-based data mining tool. SAS has made this available for academic use, although it does require some buy-in to the SAS ecosystem, including their proprietary file type. The second example, RapidMiner is specifically designed for data mining and related work. A free version is available for general use with small datasets, although it is possible for these limits to be removed for academic users. While RapidMiner doesn't have its own proprietary file format, it reads data into a format that is stored in a repository for future use or editing. Like SAS, and other commercial data mining products, RapidMiner is designed to facilitate complex analyses by means of a graphical interface that can be used to create a workflow, rather than requiring programming skills.

The second broad group of data analytics software consists of freely available programs and software platforms which can be used by anyone with the requisite knowledge and skills. Several of these are extremely powerful and well-respected alternatives to commercial packages, and fall into two subgroups. In the first subgroup are open-source platforms which were originally developed to provide support for specific research projects, but then grew into more mature products. Two of the best-known are Weka and KNIME, both of which are available under the GNU General Public License. Both are driven by the computing and machine learning aspects of data analytics, and both require some level of programming ability to make the fullest use of their capabilities, although both also provide graphical interfaces.

The second subgroup of free software consists of general programming languages or platforms. One of these, Python (https://www.python.org/), has been available since the early 1990s. Because it allows curated contributions from the wider community in the form of libraries, Python is now one of the most-used tools for a wide range of analyses. The scikit-learn library, which includes algorithms for a huge range of statistical, data mining and machine learning tasks, is now one of the most popular data analytics tools. The other leading member of this subgroup is the R language (https://www.r-project.org/foundation/). Although originally intended to be used for statistics, as a genuine programming language it has applications far beyond this. Like Python, it allows curated additions to its collection of several thousand 'packages', which now include many tools for data mining, text mining and machine learning.

With so much data analytics software now readily available, the question arises as to whether and how users and the general public can trust the correctness of their algorithms. With open source software, serious flaws are unlikely to remain hidden for long, because many open source websites encourage users to submit formal bug reports and fixes (e.g. https://www.r-project.org/bugs.html). Proprietary software, in contrast, is typically not open to independent scrutiny, because it is developed and used by commercial companies, governments, intelligence agencies, and others with a vested interest in keeping their intellectual property (IP) secret. As for deep learning algorithms, these can be extremely difficult to validate, particularly when their authors find it difficult to explain their inner workings. To all intents and purposes, these are black boxes (Pasquale, 2015).

Conclusions: Towards responsible data analytics?

At the time of writing, two distinct cultures appear to be emerging in relation to data analytics. These can be distinguished largely on the basis of their stance on what might be called responsible use. On the one hand there are the large online corporations and government intelligence wings that see themselves as being largely beyond public oversight, albeit for different reasons. On the other hand, there are public bodies and scientific establishments (including academia) which subject themselves, perhaps not always willingly, to the rule of law. Positioned somewhere between these two cultures are

the multitude of small businesses and software enthusiasts who are perhaps unaware of any degree of public responsibility and accountability in their use of data analytics.

Attempts to reduce some of the socially negative aspects of data analytics come from two directions. The first is the tightening of legal regulatory frameworks by nation states and super-national blocs, typically in relation to data privacy. A significant example is provided by the GDPR regulations published by the European Commission in May 2018 (EC, 2018). Although this is likely to lead many smaller collectors of individual data to behave more responsibly, the global online giants are more likely to ignore legislative and other government controls over their behaviour, for example by moving their user databases to locations outside relevant government jurisdictions. Only a few mainstream corporations (e.g. Unilever) have withdrawn advertising from online companies that automatically position ads inappropriately on pages containing material posted by religious and political extremists, or engaging in unlawful behaviour. But even occasional push-back responses by online advertisers have done little to deter online giants from capitalising on the compelling competitive foundation provided by a combination of consumer data and sophisticated analytics software. It remains to be seen whether future indiscretions by online business giants who operate at a global level will lead to more effective legislative oversight by governments who do not.

The second approach focuses on the design of data analytics software. Unlike data, there are fewer legislative controls placed on the use of data analytics software, though the GDPR regulations do address the issue of 'lawful processing' of personal data. A complementary approach is being played by professional bodies in the field of computing, many of whom have drawn up codes of practice for their members that encourage their adoption of ethical design principles in developing data analytics and related software. Key examples of these codes are the ACM/IEEE CS Code of Ethics and Professional Practice (Gotterbarn et al., 1997), and the more recent consultation document on ethically aligned design from the IEEE (2018). These codes are supplemented by thought pieces from special interest groups (e.g. AI Now Institute, 2017) and experts (e.g. Shneiderman, 2017) who address issues ranging from algorithmic bias to algorithmic accountability. It remains to be seen whether such guidance carries any real weight among professionals working for companies whose operations are underpinned by neoliberal values and business disruptor principles. Responsible use of data analytics may only happen when the principal users and their governments join forces in making it happen. However, ethical data analytics is not simply a technological problem requiring a technological solution; it is fundamentally a social goal that has to be worked for. Meaningful change in the role of data analytics requires no less than a change in attitudes among its many beneficiaries, whether they are online businesses, governments or, perhaps most important of all, individual citizens.

References

AI Now Institute (2017) AI Now 2017 Report, https://ainowinstitute.org/AI_Now_2017_Report.pdf (Accessed 5 June 2018).

Anderson, 2008) The end of theory: The data deluge makes the scientific method obsolete, https://www.wired.com/2008/06/pb-theory/, 23 June 2008 (Accessed: 1 August 2014).

Angwin, J., Larson, J. Mattu, S. & Kirchner, L. (2016) Machine Bias, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Accessed 10 June 2018).

Chen, C. W & Koufaris, M (2015) The impact of decision support system features on user overconfidence and risky behaviour, *European Journal of Information Systems*, 24(6), pp.607-623.

European Commission (2018) 2018 Reform of EU Data Protection Rules, https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en#abouttheregulationanddataprotection (Accessed 10 June 2018).

Galloway, S. (2017) *The Four: The hidden DNA of Amazon, Apple, Facebook, and Google* (New York, NY: Portfolio).

Gotterbarn, D., Miller, K. & Rogerson, S. (1997) Software engineering code of ethics, *Communications of the ACM*, 40(11), pp.110-118.

Harris, J. (2018) The tyranny of algorithms is part of our lives: soon they could rate everything we do, *The Guardian*, 5 March 2018. Available online at: https://www.theguardian.com/commentisfree/2018/mar/05/algorithms-rate-credit-scores-finances-data (ACCESSED: 10 June 2018).

IEEE (2017) Ethically Aligned Design: A vision for prioritizing human well-being with autonomous and intelligent systems, Version II (Institute of Electrical and Electronics Engineers).

Laney, D. (2012) Deja VVVu: Others claiming Gartner's construct for Big Data, https://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/, 14 January 2012 (Accessed: 14 April 2018).

Ordnance Survey (2018) OS Open Data, https://www.ordnancesurvey.co.uk/business-and-government/products/opendata.html (Accessed: 12 April 2018).

Pasquale, E. (2015) *The Black Box Society: The secret algorithms that control money and information* (Boston, MA: Harvard University Press).

Salganik , M. J. (2018) *Bit by Bit: Social research in the digital age* (Princeton, NJ: Princeton University Press).

Shneiderman, B. (2017) Algorithmic accountability, https://www.youtube.com/watch?v=H2iiHiK-hJ0 (Accessed 17 June 2018)..

Searle, J. R. (1995) *The Construction of Social Reality* (London: Allen Lane).

Shepherd, I. D. H. (1991) Information integration and GIS, in: D. Maguire, D. Rhind & M. Goodchild (Eds), *Geographical Information Systems: principles and applications*, Vol. 1, Chapter 22, pp.337-360 (London: Longmans).

Woolf, S. (1989) Statistics and the Modern State, *Comparative Studies in Society and History*, 31(3), pp.588-604.

Zuboff, S. (2018) *The Age of Surveillance Capitalism: The fight for a human future at the new frontier of power* (New York, NY: PublicAffairs).