COMPUTER LAW & SECURITY REVIEW ■■ (2017) ■■–■■

Available online at www.sciencedirect.com

ScienceDirect

www.compseconline.com/publications/prodclaw.htm

Computer Law & Security Review

Guidelines for the responsible application of data analytics

Roger Clarke [a,b,c,*]

[a] *Xamax Consultancy Pty Ltd, Canberra, Australia*
[b] *University of NSW Law, Sydney, Australia*
[c] *Research School of Computer Science, Australian National University, Canberra, Australia*

Keywords:
Big data
Data science
Data quality
Decision quality
Regulation

ABSTRACT

The vague but vogue notion of 'big data' is enjoying a prolonged honeymoon. Well-funded, ambitious projects are reaching fruition, and inferences are being drawn from inadequate data processed by inadequately understood and often inappropriate data analytic techniques. As decisions are made and actions taken on the basis of those inferences, harm will arise to external stakeholders, and, over time, to internal stakeholders as well. A set of Guidelines is presented, whose purpose is to intercept ill-advised uses of data and analytical tools, prevent harm to important values, and assist organisations to extract the achievable benefits from data, rather than dreaming dangerous dreams.

1. Introduction

Previous enthusiasms for management science, decision support systems, data warehousing and data mining have been rejuvenated. Fervour for big data, big data analytics and data science has been kindled, and is being sustained, by high-pressure technology salesmen. Like all such fads, there is a kernel of truth, but also a large penumbra of misunderstanding and misrepresentation, and hence considerable risk of disappointment, and worse.

A few documents have been published that purport to provide some advice on how to avoid harm arising from the practice of these techniques. Within the specialist big data analytics literature, the large majority of articles focus on techniques and applications, with impacts and implications relegated to a few comments at the end of the paper rather than even being embedded within the analysis, let alone a driving factor in the design. But see Agrawal et al. (2011), Saha and Srivastava (2014), Jagadish et al. (2014), Cai and Zhu (2015) and Haryadi et al. (2016), and particularly Merino et al. (2016).

Outside academe, most publications that offer advice appear to be motivated not by the avoidance of harm to affected values, but rather the protection of the interests of organisations conducting analyses and using the results. Examples of such documents in the public sector include DoFD (2015) – subsequently withdrawn, and UKCO (2016). Nothing resembling guidelines appears to have been published to date by the relevant US agencies, but see NIST (2015) and GAO (2016).

Some professional codes and statements are relevant, such as UNSD (1985), DSA (2016), ASA (2016) and ACM (2017). Examples also exist in the academic research arena, e.g. Rivers and Lewis (2014), Müller et al. (2016) and Zook et al. (2017). However, reflecting the dependence of the data professions on the freedom to ply their trade, such documents are oriented towards facilitation, with the protection of stakeholders commonly treated as a constraint rather than as an objective.

* *Corresponding author.* Xamax Consultancy Pty Ltd, 78 Sidaway St, Chapman ACT 2611 Canberra, Australia.
E-mail address: Roger.Clarke@xamax.com.au (R. Clarke).

https://doi.org/10.1016/j.clsr.2017.11.002
0267-3649/

Please cite this article in press as: Roger Clarke, Guidelines for the responsible application of data analytics, Computer Law & Security Review: The International Journal of Technology Law and Practice (2017), doi: 10.1016/j.clsr.2017.11.002

Documents have begun to emerge from government agencies that perform regulatory rather than stimulatory functions. See, for example, a preliminary statement issued by Data Protection Commissioners (WP29, 2014), a consultation draft from the Australian Privacy Commissioner (OAIC, 2016), and a document issued by the Council of Europe Convention 108 group (CoE 2017). These are, however, unambitious and diffuse, reflecting the narrow statutory limitations of such organisations to the protection of personal data. For a more substantial discussion paper, see ICO (2017).

It is vital that guidance be provided for at least those practitioners who are concerned about the implications of their work. In addition, a reference-point is needed as a basis for evaluating the adequacy of organisational practices, of the codes and statements of industry and professional bodies, of recommendations published by regulatory agencies, and of the provisions of laws and statutory codes. This paper's purpose is to offer such a reference-point, expressed as guidelines for practitioners who are seeking to act responsibly in their application of analytics to big data collections.

This paper draws heavily on previous research reported in Wigan and Clarke (2013), Clarke (2016a, 2016b), Raab and Clarke (2016) and Clarke (2017b). It also reflects literature critical of various aspects of the big data movement, notably Bollier (2010), Boyd and Crawford (2011), Lazer et al. (2014), Metcalf and Crawford (2016), King and Forder (2016) and Mittelstadt et al. (2016). It first provides a brief overview of the field, sufficient to provide background for the remainder of the paper. It then presents a set of Guidelines whose intentions are to filter out inappropriate applications of data analytics, and provide a basis for recourse by aggrieved parties against organisations whose malbehaviour or misbehaviour results in harm. An outline is provided of various possible applications of the Guidelines.

2. Background

The 'big data' movement is largely a marketing phenomenon. Much of the academic literature has been cavalier in its adoption and reticulation of vague assertions by salespeople. As a result, definitions of sufficient clarity to assist in analysis are in short supply. This author adopts the approach of treating as 'big data' any collection that is sufficiently large that someone is interested in applying sophisticated analytical techniques to it. However, it is important to distinguish among several categories:

- a single large data collection; and
- a consolidation of two or more data collections, which may be achieved through:
  - merger into a single physical data collection; or
  - interlinkage into a single virtual data collection

The term 'big data analytics' is distinguishable from its predecessor 'data mining' primarily on the basis of the decade in which it is used. It is subject to marketing hype to almost the same extent as 'big data'. So all-inclusive are its usages that a reasonable working definition is:

Big data analytics encompasses all processes applied to big data that may enable inferences to be drawn from it.

The term 'data scientist' emerged two decades ago as an upbeat alternative to 'statistician' (Press, 2013). Its focus is on analytic techniques, whereas the more recent big data movement commenced with its focus on data. The term 'data science' has been increasingly co-opted by the computer science discipline and business communities in order to provide greater respectability to big data practices. Although computer science has developed some additional techniques, a primary focus has been the scalability of computational processes to cope with large volumes of disparate data. It may be that the re-capture of the field by the statistics discipline will bring with it a recovery of high standards of professionalism and responsibility – which, this paper argues, are sorely needed. In this paper, however, the still-current term 'big data analytics' is used.

Where data is not in a suitable form for application of any particular data analytic technique, modifications may be made to it in an attempt to address the data's deficiencies. This was for many years referred to as 'data scrubbing', but it has become more popular among proponents of data analytics to use the misleading terms 'data cleaning' and 'data cleansing' (e.g. Rahm and Do, 2000, Müller and Freytag, 2003). These terms imply that the scrubbing process reliably achieves its aim of delivering a high-quality data collection. Whether that is actually so is highly contestable, and is seldom demonstrated through testing against the real world that the modified data purports to represent. There are many challenging aspects of data quality. What should be done where data-items that are important to the analysis are empty ('null')? And what should be done where they contain values that are invalid according to the item's definition, or have been the subject of varying definitions over the period during which the data-set has been collected? Another term that has come into currency is 'data wrangling' (Kandel et al., 2011). Although the term is honest and descriptive, and the authors adopt a systematic approach to the major challenge of missing data, their processes for 'correcting erroneous values' are merely computationally-based 'transforms', neither sourced from nor checked against the real world. The implication that data is 'clean' or 'cleansed' is commonly an over-claim, and hence such terms should be avoided in favour of the frank and usefully descriptive term 'data scrubbing'.

Where data is consolidated from two or more data collections, some mechanism is needed to determine which records in each collection are appropriately merged or linked. In some circumstances there may be a common data-item in each collection that enables associations between records to be reliably postulated. In many cases, a combination of data-items (e.g., in the case of people, the set of first and last name, date-of-birth and postcode) may be regarded as representing the equivalent of a common identifier. This process has long been referred to as computer or data matching (Clarke, 1994). Other approaches can be adopted, but generally with even higher incidences of false-positives (matches that are made but that are incorrect) and false-negatives (matches that could have been made but were not). A further issue is the extent to which a consolidated collection should contain all entries or only those for which a match has (or has not) been found. This decision may have a significant

Please cite this article in press as: Roger Clarke, Guidelines for the responsible application of data analytics, Computer Law & Security Review: The International Journal of Technology Law and Practice (2017), doi: 10.1016/j.clsr.2017.11.002