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

impact on the usability of the collection, and on the quality of inferences drawn from it.

Significantly, descriptions of big data analytics processes seldom make any provision for a pre-assessment of the nature and quality of the data that is to be processed. See, for example, Jagadish (2015) and Cao (2017). Proponents of big data analytics are prone to make claims akin to 'big trumps good', and that data quality is irrelevant if enough data is available. Circumstances exist in which such claims may be reasonable; but for most purposes they are not (Bollier, 2010; Boyd and Crawford, 2011; Clarke, 2016a), and data quality is an important consideration. McFarland and McFarland (2015) argue that 'precisely inaccurate' results arise from the 'biased samples' that are an inherent feature of big data.

A structured framework for assessing data quality is presented in Table 1. It draws on a range of sources, importantly Huh et al. (1990), Wang and Strong (1996), Müller and Freytag (2003) and Piprani and Ernst (2008). See also Hazen et al. (2014). Each of the factors in the first group can be assessed at the time of data acquisition and subsequently, whereas those in the second group, distinguished as 'information quality' factors, can only be judged at the time of use.

Underlying these factors are features of data that are often overlooked, but that become very important in the 'big data' context of data expropriation, re-purposing and merger. At the heart of the problem is the materially misleading presumption that data is 'captured'. That which pre-exists the act of data collection comprises real-world phenomena, not data that is available for 'capture'. Each item of data is created, by a process performed by a human or an artefact that senses the world and records a symbol that is intended to represent some aspect of the phenomena that is judged to be relevant. The choice of phenomena and of their attributes, and the processes for creating data to represent them, are designed and implemented by or on behalf of some entity that has some purpose in mind. The effort invested in data quality assurance at the time that it is created reflects the characteristics of the human or artefact that creates it, the process whereby it is created, the purpose of the data, the value of the data and of data quality to the relevant entity, and the available resources. Hence the relationship between the data-item and the real world phenomenon that it purports to represent is not infrequently tenuous, and is subject to limitations of definition, observation, measurement, accuracy, precision and cost.

The conduct of data analytics also depends heavily on the meanings imputed to data-items. Uncertainties arise even within a single data collection. Where a consolidated collection is being analysed, inferences may be drawn based on relationships among data-items that originated from different sources. The reasonableness of the inferences is heavily dependent not only on the quality and meaning of each item, but also on the degree of compatibility among their quality profiles and meanings.

A further serious concern is the propensity for proponents of big data to rely on correlations, without any context resembling a causative model. This even extends to championing the death of theory (Anderson, 2008; Mayer-Schonberger and Cukier, 2013). Further, it is all too common for proponents of big data analytics to interpret correlations as somehow being predictive, and then apply them as if they were prescriptive.

When big data analytics techniques are discussed, the notion of Artificial Intelligence (AI) is frequently invoked. This is a catch-all term that has been used since the mid-1950s. Various strands have had spurts of achievement, particularly in the pattern-matching field, but successes have been interspersed

Table 1 – Quality factors.

Data Quality Factors	
(assessable at the time of creation and subsequently)	
D1	**Syntactic Validity** Conformance of the data with the domain on which the data-item is defined
D2	**Appropriate (Id)entity Association** A high level of confidence that the data is associated with the particular real-world identity or entity whose attribute(s) it is intended to represent
D3	**Appropriate Attribute Association** The absence of ambiguity about which real-world attribute(s) the data is intended to represent
D4	**Appropriate Attribute Signification** The absence of ambiguity about the particular state of the particular real-world attribute(s) that the data is intended to represent
D5	**Accuracy** A high degree of correspondence of the data with the real-world phenomenon that it is intended to represent, typically measured by a confidence interval, such as '±1 degree Celsius'
D6	**Precision** The level of detail at which the data is captured, reflecting the domain on which valid contents for that data-item are defined, such as 'whole numbers of degrees Celsius'
D7	**Temporal Applicability** The absence of ambiguity about the date and time when, or the period of time during which, the data represents or represented a particular real-world phenomenon. This is important in the case of volatile data-items such as total rainfall for the last 12 months, marital status, fitness for work, age, and the period during which an income-figure was earned or a licence was applicable
Information Quality Factors	
(assessable only at the time of use)	
I1	**Theoretical Relevance** A demonstrable capability of the data-item to make a difference to the inferencing process in which the data is to be used
I2	**Practical Relevance** A demonstrable capability of the data-item's content to make a difference to the inferencing process in which the data is to be used
I3	**Currency** The absence of a material lag between a real-world occurrence and the recording of the corresponding data
I4	**Completeness** The availability of sufficient contextual information that the data is not liable to be misinterpreted
I5	**Controls** The application of business processes that ensure that the data quality and information quality factors have been considered prior to the data's use
I6	**Auditability** The availability of metadata that evidences the data quality and information quality factors

Adapted version of Table 1 of Clarke (2016a)

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

within a strong record of failure, and considerable dispute (e.g. Dreyfus, 1992, Katz, 2012). Successive waves of enthusiasts keep emerging, to frame much the same challenges somewhat differently, and win more grant money from parallel new waves of funding decision-makers. Meanwhile, the water has been muddied by breathless, speculative extensions of AI notions into the realms of metaphysics. In particular, an aside by von Neumann about a 'singularity' has been elevated to spirituality (Moravec, 2000; Kurzweil, 2005), and longstanding sci-fi notions of 'super-intelligence' have been re-presented as philosophy (Bostrom, 2014).

Multiple threads of AI are woven into big data mythology. Various words with a similarly impressive sound to 'intelligent' have been used as marketing banners, such as 'expert', 'neural', 'connectionist', 'learning' and 'predictive'. Definitions are left vague, with each new proposal applying Arthur C. Clarke's Third Law, and striving to be 'indistinguishable from magic' and hence to gain the mantle of 'advanced technology'. Within the research community, expressions of scepticism are in short supply, but Lipton (2015) encapsulates the problem by referring to "an unrealistic expectation that modern feed-forward neural networks exhibit human-like cognition".

One cluster of techniques is marketed as 'machine learning'. A commonly-adopted approach ('supervised learning') involves some kind of (usually quite simple) data structure being provided to a piece of generic software, often one that has an embedded optimisation function. A 'training set' of data is fed in. The process of creating this artefact is claimed to constitute 'learning'. Aspects of the "substantial amount of 'black art'" involved are discussed in Domingos (2012).

Even where some kind of objective is inherent in the data structure and/or the generic software, application of the metaphor of 'learning' is something of stretch for what is a sub-human and in many cases a non-rational process (Burrell, 2016). A thread of work that hopes to overcome some of the weaknesses expands the approach from a single level to a multi-layered model. Inevitably, this too has been given marketing gloss by referring to it as 'deep learning'. Even some enthusiasts are appalled by the hyperbole: "machine learning algorithms [are] not silver bullets, . . . not magic pills, . . . not tools in a toolbox – they are method{ologie}s backed by rational thought processes with assumptions regarding the datasets they are applied to" (Rosebrock, 2014).

A field called 'predictive analytics' over-claims in a different way. Rather than merely extrapolating from a data-series, it involves the extraction of patterns and then extrapolation of the patterns rather than the data; so the claim of 'prediction' is bold. Even some enthusiasts have warned that predictive analytics can have "'unintended side effects' – [things] you didn't really count on when you decided to build models and put them out there in the wild" (Perlich, quoted in Swoyer (2017)).

There is little doubt that there are specific applications to which each particular approach is well-suited – and also little doubt that each is neither a general approach nor deserving of the pretentious title used to market it. As a tweeted aphorism has it: "Most firms that think they want advanced AI/ML really just need linear regression on cleaned-up data" (Hanson, 2016).

The majority of big data analytics activity is performed behind closed doors. One common justification for this is commercial competitiveness, but other factors are commonly at work, in both private and public sector contexts. As a result of the widespread lack of transparency, it is far from clear that practices take into account the many challenges that are identified in this section.

Transparency is in any case much more challenging in the contemporary context than it was in the past. During the early decades of software development, until c.1990, the rationale underlying any particular inference was apparent from the independently-specified algorithm or procedure implemented in the software. Subsequently, so-called expert systems adopted an approach whereby the problem-domain is described, but the problem and solution, and hence the rationale for an inference, are much more difficult to access. Recently, purely empirical techniques such as neural nets and the various approaches to machine learning have attracted a lot of attention. These do not even embody a description of a problem domain. They merely comprise a quantitative summary of some set of instances (Clarke, 1991). In such circumstances, no humanly-understandable rationale for an inference exists, and in many cases none can be created. As a result, transparency is non-existent, and accountability is impossible (Burrell, 2016; Knight, 2017). To cater for such problems, Broeders et al. (2017), writing in the context of national security applications, called for the imposition of a legal duty of care and requirements for external reviews, and the banning of automated decision-making.

This brief review has identified a substantial set of risk factors. Critique is important, but critique is by its nature negative in tone. It is incumbent on critics to also offer positive and sufficiently concrete contributions towards resolution of the problems that they perceive. The primary purpose of this paper is to present a set of Guidelines whose application would address the problems and establish a reliable professional basis for the practice of data analytics.

3. The Guidelines

The Guidelines presented here avoid the word 'big', and refer simply to 'data' and 'data analytics'. These are straightforward and generic terms whose use conveys the prescriptions' broad applicability. The Guidelines are of particular relevance to personal data, because data analytics harbours very substantial threats when applied to data about individuals. The Guidelines are expressed quite generally, however, because inferences drawn from any form of data may have negative implications for individuals, groups, communities, societies, polities, economies or the environment. The purpose of the Guidelines is to assist in the avoidance of harm to all values of all stakeholders. In addition to external stakeholders, shareholders and employees stand to lose where material harm to a company's value arises from poorly-conducted data analytics, including not only financial loss and compliance breaches but also reputational damage.

The Guidelines are presented in Table 2, divided into four segments. Three of the segments correspond to the

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

Table 2 – Guidelines for the responsible application of data analytics.

1. General

DO's

1.1 **Governance**
Ensure that a comprehensive governance framework is in place prior to, during, and for the relevant period after data acquisition, analysis and use activities, that it is commensurate with the activities' potential impacts, and that it encompasses:
a. risk assessment and risk management from the perspectives of all affected parties
b. express assignments of accountability, at an appropriate level of granularity

1.2 **Expertise**
Ensure that all individuals participating in the activities have education, training, and experience in relation to the real-world systems about which inferences are to be drawn, appropriate to the roles that they play

1.3 **Compliance**
Ensure that all activities are compliant with all relevant laws and established public policy positions within relevant jurisdictions, and with public standards of behaviour

2. Data Acquisition

DO's

2.1 **The Problem Domain**
Understand the real-world systems about which inferences are to be drawn and to which data analytics are to be applied

2.2 **The Data Sources**
Understand each source of data, including:
a. the data's provenance
b. the purposes for which the data was created
c. the meaning of each data-item at the time of creation
d. the data quality at the time of creation
e. the data quality and information quality at the time of use

2.3 **Data Merger**
If data is to be merged from multiple sources, assess the compatibility of the various collections, records and items of data, taking into account the data's provenance, purposes, meaning and quality, and the potential impact of mis-matching and mistaken assumptions

2.4 **Data Scrubbing**
If data is to be scrubbed, cleaned or cleansed, assess the reliability of the processes for the intended purpose and the potential impacts of mistaken assumptions and erroneous changes

2.5 **Identity Protection**
If the association of data with an entity is sensitive, apply techniques to the data whose effectiveness is commensurate with the risks to those entities, in order to ensure pseudonymisation (if the purpose is to draw inferences about individual entities), or de-identification (if the purpose is other than to draw inferences about individual entities)

2.6 **Data Security**
Minimise the risks arising from data acquisition, storage, access, distribution and retention, and manage the unavoidable risks

DON'Ts

2.7 **Identifier Compatibility**
Don't merge data-sets unless the identifiers in each data-set are compatible with one another at a level of reliability commensurate with the potential impact of the inferences drawn

2.8 **Content Compatibility**
Don't merge data-sets unless the reliability of comparisons among the data-items in the sources reaches a threshold commensurate with the potential impact of the inferences drawn

3. Data Analysis

DO's

3.1 **Expertise**
Ensure that all staff and contractors involved in the analysis have:
a. appropriate professional qualifications
b. training in the specific tools and processes
c. sufficient familiarity with the real-world system to which the data relates and with the manner in which the data purports to represent that real-world system
d. accountability for their analyses

3.2 **The Nature of the Tools**
Understand the origins, nature and limitations of data analytic tools that are considered for use

(*continued on next page*)

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

Table 2 – (*continued*)

3.3 **The Nature of the Data Processed by the Tools**
Understand the assumptions that data analytic tools make about the data that they process, and the extent to which the data to be processed is consistent with those assumptions. Important areas in which assumptions may exist include:
a. the presence of values in relevant data-items
b. the presence of only specific, pre-defined values in relevant data-items
c. the scales against which relevant data-items have been measured
d. the precision with which relevant data-items have been expressed

3.4 **The Suitability of the Tool and the Data**
Demonstrate the applicability of each particular data analytic tool to the particular data that it is proposed be processed using it

DON'Ts

3.5 **Inappropriate Data**
Don't apply data analytics unless the data satisfies threshold tests commensurate with the potential impact of the inferences drawn, in relation to data quality, internal consistency, and reliable correspondence with the real-world systems about which inferences are to be drawn

3.6 **Humanly-Understandable Rationale**
Don't apply an analytical tool that lacks transparency, by which is meant that the rationale for inferences that it draws is expressible in humanly-understandable terms

4. Use of the Inferences

DO's

4.1 **The Impacts**
Understand the potential negative impacts on stakeholders of reliance on the inferences drawn, taking into account the quality of the data and the data analysis process

4.2 **Evaluation**
Where decisions based on inferences from data analytics may have material negative impacts, evaluate the advantages and disadvantages of proceeding, by conducting cost-benefit analysis and risk assessment from an organisational perspective, and impact assessments from the perspectives of other internal and external stakeholders

4.3 **Reality Testing**
Test a sufficient sample of the results of the analysis against the real world, in order to gain insight into the reliability of the data as a representation of relevant real-world entities and their attributes

4.4 **Safeguards**
Design, implement and maintain safeguards and mitigation measures, together with controls that ensure the safeguards and mitigation measures are functioning as intended, commensurate with the potential impacts of the inferences drawn

4.5 **Proportionality**
Where specific decisions based on inferences from data analytics may have material negative impacts on individuals, consider the reasonableness of the decisions prior to committing to them

4.6 **Contestability**
Where actions are taken based on inferences drawn from data analytics, ensure that the rationale for the decisions is transparent to people affected by them, and that mechanisms exist whereby stakeholders can access information about, and if appropriate complain about and dispute interpretations, inferences, decisions and actions

4.7 **Breathing Space**
Provide stakeholders who perceive that they will be negatively impacted by the action with the opportunity to understand and to contest the proposed action

4.8 **Post-Implementation Review**
Ensure that actions and their outcomes are audited, and that adjustments are made to reflect the findings

DON'Ts

4.9 **Humanly-Understandable Rationale**
Don't take actions based on inferences drawn from an analytical tool in any context that may have a material negative impact on any stakeholder unless the rationale for each inference is readily available to those stakeholders in humanly-understandable terms

4.10 **Precipitate Actions**
Don't take actions based on inferences drawn from data analytics until stakeholders who perceive that they may be materially negatively impacted by the action have had a reasonable opportunity to understand and to contest the proposed action. Denial of a reasonable opportunity is only justifiable on the basis of emergency, as distinct from urgency or mere expediency or efficiency. Where a reasonable opportunity is not provided, ensure that stringent safeguards, mitigation measures and controls are designed, implemented and maintained in relation to justification, reporting, review, and recourse in the case of unjustified or disproportionate actions

4.11 **Automated Decision-Making**
Don't delegate to a device any decision that has potentially harmful effects without ensuring that it is subject to specific human approval prior to implementation, by a person who is acting as an agent for the accountable organisation

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

successive processes involved – acquisition of the data, analysis of the data in order to draw inferences, and use of the inferences. The first segment specifies generic requirements that apply across all of the phases.

Each Guideline is expressed in imperative mode, some in the positive and others in the negative. However, they are not statements of law, nor are they limited to matters that are subject to legal obligations. They are declarations of what is needed in order to manage the risks arising from data quality issues, data meaning uncertainties, incompatibilities in data meaning among similar data-items sourced from different data-collections, misinterpretations of meaning, mistakes introduced by data scrubbing, approaches taken to missing data that may solve some problems but at the cost of creating or exacerbating others, erroneous matches, unjustified assumptions about the scale against which data has been measured, inappropriate applications of analytical tools, lack of review, and confusions among correlation, causality, predictive power and normative force.

The organisations and individuals to whom each Guideline is addressed will vary depending on the context. In some circumstances, a single organisation, a single small team within an organisation, or even a single individual, might perform all of the activities involved. On the other hand, multiple teams within one organisation, or across multiple organisations, may perform several of the activities.

The Guidelines are intended to be comprehensive. As a result, in any particular context, some of them will be redundant, and some would be more usefully expressed somewhat differently. In particular, some of the statements are primarily relevant to data that refers to an individual human being. Such statements may be irrelevant, or may benefit from re-phrasing, where the data relates to inanimate parts of the physical world (e.g. meteorological, geophysical, vehicular traffic or electronic traffic data), or to aggregate economic or social phenomena. In such circumstances, careful sub-setting and adaptation of the Guidelines is appropriate.

4. Ways to apply the Guidelines

These Guidelines, in their current or some adapted form, can be adopted by any organisation. Staff and contractors can be required to demonstrate that their projects are compliant, or, to the extent that they are not, to explain why not. In practice, adoption may be driven by staff and contractors, because many practitioners are concerned about the implications of their work, and would welcome the availability of an instrument that enables them to raise issues in the context of project risk management.

Organisational self-regulation of this kind has the capacity to deliver value for the organisation and for shareholders, but it has only a mediocre track-record in benefiting stakeholders outside the organisation. A stronger institutional framework is needed if preventable harm arising from inappropriate data, analysis and use is to be avoided.

Industry associations can adopt or adapt the Guidelines, as can government agencies that perform oversight functions. Industry regulation through a Code of Practice may achieve some positive outcomes for organisations in terms of the quality of work performed, and particularly by providing a means of defending against and deflecting negative media reports, public concerns about organisational actions, and acts by any regulator that may have relevant powers. In practice, however, such Codes are applied by only a proportion of the relevant organisations, are seldom taken very seriously (such as by embedding them within corporate policies, procedures, training programs and practices), are unenforceable, and generally offer very limited benefits to external stakeholders. Nonetheless, some modest improvements would be likely to accrue from adoption, perhaps at the level of symbolism, but more likely as a means of making it more difficult for data analytics issues to be ignored.

Individual organisations can take positive steps beyond such, largely nominal, industry sector arrangements. They can embed consideration of the factors identified in these Guidelines into their existing business case, cost/benefit and/or risk assessment and management processes. In order to fulfil their corporate social responsibility commitments, they can also evaluate proposed uses of data analytics from the perspectives of external stakeholders. A very narrow and inadequate approach to this merely checks legal compliance, as occurs with the pseudo-PIA processes conventional in the public sector throughout much of North America (Clarke, 2011 s.4), and in the new European 'Data Protection Impact Assessment' (DPIA) mechanism (Clarke, 2017a). Much more appropriately, a comprehensive Privacy Impact Assessment can be performed (Clarke, 2009; Wright and de Hert, 2012). In some circumstances, a much broader social impact assessment is warranted (Raab and Wright, 2012; Wright and Friedewald, 2013). Raab & Wright (2012 pp. 379–381) calls for extension of the scope of PIAs firstly to a wide range of impacts on the individual's "relationships, positions and freedoms", then to "impacts on groups and categories", and finally to "impacts on society and the political system".

A further step that individual organisations can take is to enter into formal undertakings to comply with a Code, combined with submission to the decisions of a complaints body, ombudsman or tribunal that is accessible by any aggrieved party that has the resources to conduct investigations, that has enforcement powers, and that uses them. Unfortunately, such arrangements are uncommon, and it is not obvious that suitable frameworks exist within which an enforceable Code along the lines of these Guidelines could be implemented.

Another possibility is for a formal and sufficiently precise Standard to be established, and for this to be accepted by courts as the measuring-stick against which the behaviour of organisations that conduct data analytics is to be measured. A loose mechanism of this kind is declaration by an organisation that it is compliant with a particular published Standard. In principle, this would appear to create a basis for court action by aggrieved parties. In practice, however, it appears that such mechanisms are seldom effective in protecting either internal or external stakeholders.

As discussed earlier, some documents exist that at least purport to provide independent guidance in relation to data analytics activities. These Guidelines can be used as a yardstick against which such documents can be measured. The UK Cabinet Office's 'Data Science Ethical Framework' (UKCO,

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

2016) was assessed against an at-that-time-unformalised version of these Guidelines, and found to be seriously wanting (Raab and Clarke, 2016). For different reasons, and in different ways, the Council of Europe document (CoE 2017) falls a very long way short of what is needed by professionals and the public alike as a basis for responsible use of data analytics. The US Government Accountability Office has identified the existence of "possible validity problems in the data and models used in [data analytics and innovation efforts – DAI]" (GAO, 2016, p. 38), but has done nothing about them. An indication of the document's dismissiveness of the issues is this quotation: "In automated decision making [using machine learning], monitoring and assessment of data quality and outcomes are needed to gain and maintain trust in DAI processes" (p.13, fn.8). Not only does the statement appear in a mere footnote, but the concern is solely about 'trust' and not at all about the appropriateness of the inferences drawn, the actions taken as a result of them, or the resource efficiency and equitability of those actions. The current set of documents from the US National Institute of Standards and Technology (NIST, 2015) is also remarkably devoid of discussion about data quality and process quality, and offers no process guidance along the lines of the Guidelines proposed in this paper.

Another avenue whereby progress can be achieved is through adoption by the authors of text-books. At present, leading texts commonly have a brief, excusatory segment, usually in the first or last chapter. Curriculum proposals commonly suffer the same defect, e.g. Gupta et al. (2015), Schoenherr and Speier-Pero (2015). Course-designers appear to generally follow the same pattern, and schedule a discussion or a question in an assignment, which represents a sop to the consciences of all concerned, but does almost nothing about addressing the problems, and nothing about embedding solutions to those problems within the analytics process. It is essential that the specifics of the Guidelines in Table 2 be embedded in the structure of text-books and courses, and that students learn to consider each issue at the point in the acquisition/analysis/use cycle at which each challenge needs to be addressed.

None of these approaches is a satisfactory substitute for legislation that places formal obligations on organisations that apply data analytics, and that provides aggrieved parties with the capacity to sue organisations where they materially breach requirements and there are material negative impacts. Such a scheme may be imposed by an activist legislature, or a regulatory framework may be legislated and the Code negotiated with the relevant parties prior to promulgation by a delegated agency. It is feasible for organisations themselves to submit to a parliament that a co-regulatory scheme of such a kind should be enacted, for example where scandals arise from inappropriate use of data analytics by some organisations, which have a significant negative impact on the reputation of an industry sector as a whole.

5. Conclusions

This paper has not argued that big data and big data analytics are inherently evil. It has also not argued that no valid applications of the ideas exist, nor that all data collections are of such low quality that no useful inferences can be drawn from them, nor that all mergers of data from multiple sources are necessarily logically invalid or necessarily deliver fatally flawed consolidated data-sets, nor that all data scrubbing fails to clean data, nor that all data analytics techniques make assumptions about data that can under no circumstances be justified, nor that all inferences drawn must be wrong. Expressed in the positive, some big data has potential value, and some applications of data analytics techniques are capable of realising that potential.

What this paper has done is to identify a very large fleet of challenges that have to be addressed by each and every specific proposal for the expropriation of data, the re-purposing of data, the merger of data, the scrubbing of data, the application of data analytics to it, and the use of inferences drawn from the process in order to make, or even guide, let alone explain, decisions and action that affect the real world. Further, it is far from clear that measures are being adopted to meet these challenges.

Ill-advised applications of data analytics are preventable by applying the Guidelines proposed in this paper. As the 'big data' mantra continues to cause organisations to have inflated expectations of what data analytics can deliver, both shareholders and external stakeholders need constructive action to be taken in order to get data analytics practices under control, and avoid erroneous business decisions, loss of shareholder value, inappropriate policy outcomes, and unjustified harm to individual, social, economic and environmental values. The Guidelines proposed in this paper therefore provide a basis for the design of organisational and regulatory processes whereby positive benefits can be gained from data analytics, but undue harm avoided.

Acknowledgement

The author received valuable feedback from Prof. Louis de Koker of La Trobe University, Melbourne, David Vaile and Dr. Lyria Bennett Moses of UNSW, Sydney, Dr. Kerry Taylor of the ANU, Canberra, Dr. Kasia Bail of the University of Canberra, Prof. Charles Raab of Edinburgh University, and an anonymous reviewer. Evaluative comments are those of the author alone.

REFERENCES

ACM. Statement on algorithmic transparency and accountability. Association for Computing Machinery; 2017. Available from: https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf. [Accessed November 24, 2017].

Agrawal D, Philip Bernstein P, Bertino E, Davidson S, Dayal U, Franklin M, Gehrke J, et al. Challenges and opportunities with big data 2011-1. Cyber Center Technical Reports, Paper 1; 2011. Available from: http://docs.lib.purdue.edu/cctech/1. [Accessed November 24, 2017].

Anderson C. The end of theory: the data deluge makes the scientific method obsolete. Wired Magazine 16:07; 2008.

ASA. Ethical guidelines for statistical practice. American Statistical Association; 2016. Available from: http://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx. [Accessed November 24, 2017].
Bollier D. The promise and peril of big data. The Aspen Institute; 2010. Available from: https://www.emc.co.tt/collateral/analyst-reports/10334-ar-promise-peril-of-big-data.pdf. [Accessed November 24, 2017].
Bostrom N. Superintelligence: paths, dangers, strategies. Oxford University Press; 2014.
Boyd D, Crawford K. Six provocations for big data. Proc. Symposium on the Dynamics of the Internet and Society; 2011. Available from: http://ssrn.com/abstract=1926431. [Accessed November 24, 2017].
Broeders D, Schrijvers E, van der Sloot B, van Brakel R, de Hoog J, Ballina EH. Big data and security policies: towards a framework for regulating the phases of analytics and use of big data. Comput Law Secur Rev 2017;33:309–23.
Burrell J. How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc 2016;3(1):1–12.
Cai L, Zhu Y. The challenges of data quality and data quality assessment in the big data era. Data Sci J 2015;14(2):1–10. Available from: https://datascience.codata.org/articles/10.5334/dsj-2015-002/.
Cao L. Data science: a comprehensive overview. ACM Computing Surveys; 2017. Available from: http://dl.acm.org/ft_gateway.cfm?id=3076253&type=pdf. [Accessed November 24, 2017].
Clarke R. A contingency approach to the software generations. Database 1991;22(3):23–34. PrePrint available from: http://www.rogerclarke.com/SOS/SwareGenns.htmlSummer 1991.
Clarke R. Dataveillance by governments: the technique of computer matching. Inf Tech People 1994;7(2):46–85. PrePrint available from: http://www.rogerclarke.com/DV/MatchIntro.html.
Clarke R. Privacy impact assessment: its origins and development. Comput Law Secur Rev 2009;25(2):123–35. PrePrint available from: http://www.rogerclarke.com/DV/PIAHist-08.html.
Clarke R. An evaluation of privacy impact assessment guidance documents. Int Data Priv Law 2011;1(2):111–20. PrePrint available from: http://www.rogerclarke.com/DV/PIAG-Eval.html.
Clarke R. Big data, big risks. Inf Syst J 2016a;26(1):77–90. PrePrint available from: http://www.rogerclarke.com/EC/BDBR.html.
Clarke R. Quality assurance for security applications of big data. Proc. European Intelligence and Security Informatics Conference (EISIC), Uppsala, 17–19 August 2016; 2016b. PrePrint available from: http://www.rogerclarke.com/EC/BDQAS.html. [Accessed November 24, 2017].
Clarke R. The distinction between a PIA and a Data Protection Impact Assessment (DPIA) under the EU GDPR. Working Paper, Xamax Consultancy Pty Ltd; 2017a. Available from: http://www.rogerclarke.com/DV/PIAvsDPIA.html. [Accessed November 24, 2017].
Clarke R. Big data prophylactics, chapter 1. In: Lehmann A, Whitehouse D, Fischer-Hübner S, Fritsch L, Raab C, editors. Privacy and identity management. Facing up to next steps. Springer; 2017b. p. 3–14 [chapter 1]. PrePrint available from: http://www.rogerclarke.com/DV/BDP.html.
CoE. Guidelines on the protection of individuals with regard to the processing of personal data in a world of big data. Convention 108 Committee, Council of Europe; 2017. Available from: https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000016806ebe7a. [Accessed November 24, 2017].
DoFD. Better practice guide for big data. Australian Dept of Finance & Deregulation, v.2; 2015. Available from: http://www.finance.gov.au/sites/default/files/APS-Better-Practice-Guide-for-Big-Data.pdf. [Accessed November 24, 2017].
Domingos P. A few useful things to know about machine learning. Commun ACM 2012;55(10):78–87.
Dreyfus H. What computers still can't do. MIT Press; 1992.
DSA. Data science code of professional conduct. Data Science Association, undated but apparently of 2016; 2016. Available from: http://www.datascienceassn.org/sites/default/files/datasciencecodeofprofessionalconduct.pdf. [Accessed November 24, 2017].
GAO. Emerging opportunities and challenges data and analytics innovation. Government Accountability Office, Washington DC; 2016. Available from: http://www.gao.gov/assets/680/679903.pdf. [Accessed November 24, 2017].
Gupta B, Goul M, Dinter B. Business intelligence and big data in higher education: status of a multi-year model curriculum development effort for business school undergraduates, MS graduates, and MBAs. Commun Assoc Inf Syst 2015;36(23):Available from: https://www.researchgate.net/profile/Babita_Gupta4/publication/274709810_Communications_of_the_Association_for_Information_Systems/links/557ecd4b08aeea18b7795225.pdf.
Hanson R. This AI boom will also bust. Overcoming Bias Blog; 2016. Available from: http://www.overcomingbias.com/2016/12/this-ai-boom-will-also-bust.html. [Accessed November 24, 2017].
Haryadi AF, Hulstijn J, Wahyudi A, van der Voort H, Janssen M. Antecedents of big data quality: an empirical examination in financial service organizations. Proc. IEEE Int'l Conf. on Big Data; 2016. pp. 116–21. Available from: https://pure.tudelft.nl/portal/files/13607440/Antecedents_of_Big_Data_Quality_IEEE2017_author_version.pdf. [Accessed November 24, 2017].
Hazen BT, Boone CA, Ezell JD, Jones-Farmer LA. Data quality for data science, predictive analytics, and big data in supply chain management: an introduction to the problem and suggestions for research and applications. Int J Prod Econ 2014;154:72–80. Available from: https://www.researchgate.net/profile/Benjamin_Hazen/publication/261562559_Data_Quality_for_Data_Science_Predictive_Analytics_and_Big_Data_in_Supply_Chain_Management_An_Introduction_to_the_Problem_and_Suggestions_for_Research_and_Applications/links/0deec534b4af9ed874000000.
Huh YU, Keller FR, Redman TC, Watkins AR. Data quality. Inf Softw Tech 1990;32(8):559–65.
ICO. Big data, artificial intelligence, machine learning and data protection. UK Information Commissioner's Office, Discussion Paper v.2.2; 2017. Available from: https://ico.org.uk/for-organisations/guide-to-data-protection/big-data/. [Accessed November 24, 2017].
Jagadish HV. Big data and science: myths and reality. Big Data Res 2015;2(2):49–52.
Jagadish HV, Gehrke J, Labrinidis A, Papakonstantinou Y, Patel JM, Ramakrishnan R, et al. Big data and its technical challenges. Commun ACM 2014;57(7):86–94.
Kandel S, Heer J, Plaisant C, Kennedy J, van Ham F, Henry-Riche N, et al. Research directions for data wrangling: visualizations and transformations for usable and credible data. Information Visualization 10. 4; 2011. 271–88. Available from: https://idl.cs.washington.edu/files/2011-DataWrangling-IVJ.pdf. [Accessed November 24, 2017].
Katz Y. Noam Chomsky on where artificial intelligence went wrong: an extended conversation with the legendary linguist. The Atlantic; 2012. Available from: https://www.theatlantic.com/technology/archive/2012/11/noam-chomsky-on-where-artificial-intelligence-went-wrong/261637/. [Accessed November 24, 2017].

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

King NJ, Forder J. Data analytics and consumer profiling: finding appropriate privacy principles for discovered data. Comput Law Secur Rev 2016;32:696–714.
Knight W. The dark secret at the heart of AI. 11 April 2017, MIT Technology Review; 2017. Available from: https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/. [Accessed November 24, 2017].
Kurzweil R. The singularity is near: when humans transcend biology. Viking; 2005.
Lazer D, Kennedy R, King G, Vespignani A. The parable of Google flu: traps in big data analysis. Science 2014;343(6176):1203–5. Available from: https://dash.harvard.edu/bitstream/handle/1/12016836/The%20Parable%20of%20Google%20Flu%20%28WP-Final%29.pdf.
Lipton ZC. (Deep Learning's Deep Flaws)'s Deep Flaws. KD Nuggets; 2015. Available from: http://www.kdnuggets.com/2015/01/deep-learning-flaws-universal-machine-learning.html. [Accessed November 24, 2017].
Mayer-Schonberger V, Cukier K. Big data, a revolution that will transform how we live, work and think. John Murray; 2013.
McFarland DA, McFarland HR. Big data and the danger of being precisely inaccurate. Big Data Soc 2015;2(2):1–4.
Merino J, Caballero I, Bibiano R, Serrano M, Piattini M. A data quality in use model for big data. Fut Gen Comput Syst 2016;63:123–30.
Metcalf J, Crawford K. Where are human subjects in big data research? The emerging ethics divide. Big Data Soc 2016;3(1):1–14.
Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: mapping the debate. Big Data Soc 2016;3(2):1–21.
Moravec H. Robot: mere machine to transcendent mind. Oxford University Press; 2000.
Müller H, Freytag J-C. Problems, methods and challenges in comprehensive data cleansing. Technical Report HUB-IB-164, Humboldt-Universität zu Berlin, Institut für Informatik; 2003. Available from: http://www.informatik.uni-jena.de/dbis/lehre/ss2005/sem_dwh/lit/MuFr03.pdf. [Accessed November 24, 2017].
Müller O, Junglas I, vom Brocke1 J, Debortoli S. Utilizing big data analytics for information systems research: challenges, promises and guidelines. Eur J Inf Syst 2016;25(4):289–302. Available from: https://www.researchgate.net/profile/Oliver_Mueller5/publication/290973859_Utilizing_Big_Data_Analytics_for_Information_Systems_Research_Challenges_Promises_and_Guidelines/links/56ec168f08aee4707a384fff/Utilizing-Big-Data-Analytics-for-Information-Systems-Research-Challenges-Promises-and-Guidelines.pdf.
NIST. NIST big data interoperability framework. Special Publication 1500-1, v.1, National Institute of Standards and Technology; 2015. Available from: https://bigdatawg.nist.gov/V1_output_docs.php. [Accessed November 24, 2017].
OAIC. Consultation draft: guide to big data and the Australian Privacy Principles. Office of the Australian Information Commissioner; 2016. Available from: https://www.oaic.gov.au/engage-with-us/consultations/guide-to-big-data-and-the-australian-privacy-principles/consultation-draft-guide-to-big-data-and-the-australian-privacy-principles. [Accessed November 24, 2017].
Piprani B, Ernst D. A model for data quality assessment. Proc. OTM Workshops (5333); 2008. pp 750–9.
Press G. A very short history of data science. Forbes; 2013. Available from: https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#375c75e355cf. [Accessed November 24, 2017].
Raab C, Clarke R. Inadequacies in the UK's data science ethical framework. Euro Data Protect L 2016;2(4):555–60. PrePrint available from: http://www.rogerclarke.com/DV/DSEFR.html.
Raab CD, Wright D, de Hert P. editors. Surveillance: extending the limits of privacy impact assessment. 2012. p. 363–83 [Ch. 17].
Rahm E, Do HH. Data cleaning: problems and current approaches. IEEE Data Eng Bull 2000;23:Available from: http://dc-pubs.dbs.uni-leipzig.de/files/Rahm2000DataCleaningProblemsand.pdf.
Rivers CM, Lewis BL. Ethical research standards in a world of big data. F1000Res 2014;3:38. Available from: https://f1000research.com/articles/3-38.
Rosebrock A. Get off the deep learning bandwagon and get some perspective. PY Image Search; 2014. Available from: https://www.pyimagesearch.com/2014/06/09/get-deep-learning-bandwagon-get-perspective/. [Accessed November 24, 2017].
Saha B, Srivastava D. Data quality: The other face of big data. Proc. Data Engineering (ICDE); 2014. pp. 1294–7. Available from: https://people.cs.umass.edu/~barna/paper/ICDE-Tutorial-DQ.pdf. [Accessed November 24, 2017].
Schoenherr T, Speier-Pero C. Data science, predictive analytics, and big data in supply chain management: current state and future potential. J Bus Logist 2015;36(1):120–32. Available from: http://www.logisticsexpert.org/top_articles/2016/2016%20-%20Research%20-%20JBL%20-%20Data%20Science,%20Predictive%20Analytics,%20and%20Big%20Data%20in%20Supply%20Chain%20Managementl.pdf.
Shanks G, Darke P. Understanding data quality in a data warehouse. Aust Comput J 1998;30:122–8.
Swoyer S. The shortcomings of predictive analytics. TDWI; 2017. Available from: https://tdwi.org/articles/2017/03/08/shortcomings-of-predictive-analytics.aspx. [Accessed November 24, 2017].
UKCO. Data science ethical framework. U.K. Cabinet Office, v.1.0; 2016. Available from: https://www.gov.uk/government/publications/data-science-ethical-framework. [Accessed November 24, 2017].
UNSD. Declaration of professional ethics. United Nations Statistical Division; 1985. Available from: http://unstats.un.org/unsd/dnss/docViewer.aspx?docID=93#start. [Accessed November 24, 2017].
Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. J Manag Inf Syst 1996;12(4):5–33. Spring, 1996.
Wigan MR, Clarke R. Big data's big unintended consequences. IEEE Comput 2013;46(6):46–53. PrePrint available from: http://www.rogerclarke.com/DV/BigData-1303.html.
WP29. Statement of the WP29 on the impact of the development of big data on the protection of individuals with regard to the processing of their personal data in the EU. Article 29 Working Party, European Union; 2014. Available from: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp221_en.pdf. [Accessed November 24, 2017].
Wright D, de Hert P, editors. Privacy impact assessments. Springer; 2012.
Wright D, Friedewald M. Integrating privacy and ethical impact assessments. Sci Public Policy 2013;40(6):755–66. Available from: http://spp.oxfordjournals.org/content/40/6/755.full.
Zook M, Barocas S, boyd d, Crawford K, Keller E, Gangadharan SP, Goodman A, et al. Ten simple rules for responsible big data research. PLoS Comput Biol 2017;13(3):Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5373508/. [Accessed November 24, 2017].

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