Big Data Prophylactics

Roger Clarke

Big Data Prophylactics

Roger Clarke[1]

Xamax Consultancy Pty Ltd, Canberra, Australia,
UNSW Law, and ANU Research School of Computer Science
Roger.Clarke@xamax.com.au

Abstract. Data mining has been re-branded as 'big data analytics'. The techniques involved harbour a substantial set of risks, many of which will be borne by individuals. This chapter argues that safeguards are needed, to protect individuals against the potentially harmful acts that organisations will take against them. Alternative forms of such 'big data prophylactics' are outlined.

Keywords: Data quality, decision quality, risk assessment, risk management, transparency, regulation, activism

1 Big Data

Big data is a fashion-item. It was an invention of marketers, as a means of breathing fresh life into the flagging booms in successively data mining and mash-ups. It has been given an aura of excitement because of the vast array of sources of data. There has been massive expropriation of social media profiles and traffic, and of wellness data from individuals' self-monitoring of their physiological states. A parallel development has been the open access movement in the public sector, which government agencies in various countries are utilising as an opportunity to break down both data silos and privacy protections. Other prospects that have been heralded include flows of streams of telemetry data (fast data), and from the Internet of Things.

---

[1] Roger Clarke is Principal of Xamax Consultancy Pty Ltd, Canberra. He is also a Visiting Professor in Cyberspace Law & Policy at the University of N.S.W., and a Visiting Professor in the Research School of Computer Science' at the Australian National University.

In order to differentiate the big data concept, it was first proposed that its key characteristics were 'volume, velocity and variety' [20]. Subsequently, commentators added 'value', while a few have recently begun to add 'veracity' [28]. Such vague formulations even find their way into the academic literature, in forms along the lines of 'data that's too big, too fast or too hard for existing tools to process'. A somewhat more useful characterisation is "the capacity to analyse a variety of mostly unstructured data sets from sources as diverse as web logs, social media, mobile communications, sensors and financial transactions" [25, p.12]. This reflects the widespread use of the term to encompass not only data collections, but also the processes applied to those collections. To overcome this ambiguity of scope, this paper distinguishes between 'big data' and 'big data analytics'.

Some commentators consider that big data bears a stronger resemblance to an ideology than to a science. At [5, p.663], boyd & Crawford depict 'big data' as "a cultural, technological, and scholarly phenomenon that rests on the interplay of [three elements]". Their first two elements correspond to 'big data' and 'big data analytics'. Their third element, on the other hand, emphasises the importance of "mythology: the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy".

The mythology, or as its proponents would prefer, the meme, has been spruiked in business school magazines [12], by business school academics [22], and by academics in other disciplines who might have been expected to temper their enthusiasm [23]). The high level of enthusiasm, coupled with these authors' willing suspension of disbelief, has given rise to counter-arguments, and to accusations such as "There's a mania around big data ..." (David Searls, quoted in [30-31]).

2 Big Data Analytics

The term 'big data analytics' refers to the processes that are applied to big data collections. A substantial array of analytical tools pre-existed the big data era, and more are being developed.

One way to categorise big data analytics is according to the purpose for which the analysis is performed. In Table 1, two broad categories are first distinguished, according to whether the analysis is aiming to deliver insights into populations or about individual entities within those populations. Within each of these major categories, distinctly different kinds of problem-types can be addressed.

It would be reasonable to expect a highly-developed set of guidelines to exist, enabling analysts to recognise firstly which analytical techniques are suitable to which of those problem-categories (and, conversely, which are not); and secondly what attributes big data collections need in order to support each of those analytical techniques. However, it is very difficult to find such guidance. Despite the explosion in postgraduate degree offerings in the big data analytics area, the vague aura of art and craft has yet to be replaced by the clarity of science and engineering, the overtone of experimentation dressed up as innovation pervades, and the application of established expertise remains uncommon.

3 Risk Factors

A range of risks arise from the current spate of over-enthusiastic and uncritical adoption of the big data meme, and of its companion notions of open government data, social media exploitation, location and tracking of people and the devices that they use, and the Internet of Things.

3.1 Data

Tables 2 and 3 show the results of prior research, which drew on the literature – particularly [19, 7 pp. 601-605, 36, 35, 29, 26] – in order to identify and briefly define relevant quality categories. Some are 'data quality' factors (which are capable of being assessed at the time the data is collected) whereas others are 'information quality' factors (which are not assessable until the data is used).

Table 1. Purposes of Big Data Analytics [after 12]

Population Focus

- **Hypothesis Testing**

This approach evaluates whether propositions are supported by the available data. The propositions may be predictions from theory, existing heuristics, or hunches

- **Population Inferencing**

This approach draws inferences about the entire population of entities, or about sub-populations. In particular, correlations may be drawn among particular attributes

- **Construction of Profiles**

This approach identifies key characteristics of some category of entities. For example, attributes and behaviours of a target group, such as 'drug mules', sufferers from particular diseases, or children with particular aptitudes, may exhibit statistical consistencies

Individual Focus

- **Discovery of Outliers**

Statistical outliers are commonly disregarded, but this approach regards them instead as valuable needles in large haystacks, because they may herald a 'flex-point' or 'quantum shift'

- **Discovery of Anomalies**

This approach draws inferences about individual entities within the population. For example, a person may be inferred to have provided inconsistent information to two organisations, or to exhibit behaviour in one context inconsistent with behaviour in another

- **Application of Profiles**

A search can be conducted for individual entities that exhibit patterns associated with a particular, previously asserted or computed profile, thereby generating a set of suspect entities

In addition to quality, it is important that those conducting data analysis be clear about how data is to be interpreted. In syntactical terms, there must be clear answers to such questions as: Is each data-item mandatory or optional? What is the meaning of an empty (or 'null') field? What

values may each field contain? At the level of semantics, the signification of each item must also be unambiguous, i.e. with which real-world attribute of which real-world entity does it correspond, and what does it say about that attribute?

During the earlier 'data mining' era, low data quality was recognised as a matter of real concern. Data was modified in a variety of ways, using a process that was referred to as 'data scrubbing' [37]. One focus of data scrubbing is on missing data – although finding an appropriate basis for interpolating appropriate values is fraught with difficulty. Modifications are mostly made on the basis of various heuristics, or after comparison with characteristics derived from the data-holdings as a whole. In only rare cases is it possible to check item-content against an external authority. There are limits to the improvements that can actually be achieved, and almost all scrubbing, by its nature, involves a proportion of false positives. Hence, while the process may achieve some improvements, there is inevitably also some worsening in quality through mistaken modifications.

By the 'big data' era, the honest term 'data scrubbing' had been replaced by 'data cleaning' and 'data cleansing' [24]. These terms imply not only that an attempt has been made to achieve 'cleanliness', but also that cleanliness has been achieved. As is apparent from the use of heuristics and the inevitability of errors, the implication is false.

Problems with data quality and data meaning are major sources of risk. Of course, this applies to all forms of administrative data processing, whether or not the data-collection in question qualifies as 'big data'. The problems are compounded, however, when data from different sources, with different and potentially incompatible meanings, and with varying levels of quality, are consolidated, and then handled as though the melange of data constitutes a single, cohesive data-collection.

Table 2. Data Quality Factors [after 11]

- **D1 Syntactical Validity**
  Conformance of the data with the domain on which the data-item is defined
- **D2 Appropriate 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 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°C'
- **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 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

Table 3: Information Quality Factors [after 11]

Information Quality Factors

- **I1 Theoretical Relevance**
  Demonstrable capability of a category of data-item (a column in a table) to make a difference to the process in which the data is to be used
- **I2 Practical Relevance**
  Demonstrable capability of the content of a particular data-item (the content of a cell in a table) to make a difference to the 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 all 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's provenance, and supports assertions relating to data semantics, data quality and information quality

3.2 Analytics

Reference has already been made to the issues arising from uncertainty about the appropriateness of analytical techniques to problem-categories, and uncertainty about the suitability of any particular data-collection for processing using any particular analytical technique.

Further inadequacies that give rise to risks are lack of transparency. This is needed in relation to the process whereby inferences are drawn, the basis on which particular data is considered to be relevant to the drawing of the inference, and the criteria that gave rise to any particular judgement.

Humans who make decisions can be called to account, and required to explain the basis on which they drew inferences, made decisions and took action. Computer systems programmed in algorithmic or procedural languages (as was the norm from the 1960s to the 1980s) embody

explicit processes and criteria, and hence they are also capable of being interrogated.

Later forms of programming language, however, embody increasing layers of mystery and inscrutability [6]. With so-called 'expert systems' approaches (which most commonly involve the expression of sets of rules), both the decision processes and the decision criteria are implicit. The most that can be re-constructed is that a particular set of rules 'fired', and that particular data was what caused them to be invoked. This is seldom a clear basis for justifying an action, and in any case many rule-based applications aren't designed to support the extraction of even this inadequate form of explanation.

The current vogue in software development can be reasonably described as 'empirical'. Neural nets and machine-learning 'algorithms' do not have anything that can sensibly be described as a decision process, and it is not feasible to extract or infer decision criteria. The issues are discussed in some depth in ss.2.1 and 2.2 of [10]. Use of these techniques represents blind faith, by the 'data scientists', by the organisations that apply them, by any regulator that attempts to review them, and ultimately by everyone affected by them.

3.3 Decision and Action

The challenges identified in the preceding sub-sections give rise to risk if they are not understood and managed. Where any such inadequacies carry forward into decisions made and actions taken, whether about resource allocation or about relationships between organisations and particular individuals, risks arise that the decisions may be unjustified, disproportionate, or just plain wrong.

It has been a fundamental tenet of democracy that dealings between government agencies and individuals must be subject to review and recourse. This principle has also found its way into consumer rights laws in many jurisdictions and many contexts, particularly the financial sector and health care. Purely empirical data analytic techniques are completely at odds with these public expectations. Yet these expectations are in some

cases expressed as legal requirements. Hence many potential applications of AI and machine learning 'algorithms' are arguably in breach of existing laws.

3.4 Consequences

The proponents of big data analytics may rail against these restraints, and complain that conservative attitudes and slow-changing laws are stifling innovation. They may protest that human rights, anti-discrimination laws and privacy protections constrain the freedom of corporations, government agencies and 'data scientists' to act in economically efficient ways – which implies the scope to impose their will on people. They may even invoke the anti-humanitarian credo that 'logic is dead ... get over it'. Or to use their own words: "Faced with massive data, [the old] approach to science -- hypothesize, model, test -- is ... obsolete. Petabytes allow us to say: 'Correlation is enough'" [1] and "Society will need to shed some of its obsession for causality in exchange for simple correlations: not knowing why but only what. Knowing why might be pleasant, but it's unimportant ..." [23].

Inherent in the boisterous claims of Anderson, Mayer-Schönberger and others is the abandonment of balance between the empirical and the rational, and its replacement by empiricism dominant and systemic explanations deprecated. This 'flight from reason' has consequences. The proponents of big data analytics focus on the value that they assert society can, and that they assert society will, extract from massive, low-quality data-collections using more or less ad hoc analytical techniques. Even if their assertions turned out to be right – and such positive evidence as exists to date is merely anecdotal and unaudited – the benefits would be accompanied by massive negative impacts – for some, if not for all.

The resource-allocation and administrative decision-making errors that follow on from poor-quality inferencing will produce losers. Review will be at least hampered, and where AI and machine-learning are involved, actually not possible. The losers will be forced to 'like it or lump it', without recourse.

But social equity cannot exist in a world in which rationality has been abandoned, and decisions are made mysteriously and enforced by the powerful. This breaches the social contract. The losers' natural reaction is to stop trusting the institutions that they deal with. Some are likely to become sullen non-compliers with the diktats of powerful organisations, while others will be more aggressive in their avoidance measures. There is then an unpleasant scale up through active falsification, via electronic forms of sabotage, to violence.

The consequences of the Anderson / Mayer-Schönberger thesis are the breakdown of social cohesion, and serious challenges to the social order on which economies and polities depend. That such seers fail to look ahead to the consequences of their wild enthusiasms is quite extraordinary.

4 Big Data Prophylactics

Many of the risks identified in the previous section will be borne by individuals rather than by the organisations that make big-data-originated mistakes. So if 'big data analytics' is to be more than just a passing fad, it needs to be accompanied by 'big data prophylactics', to provide people with protections against organisations' potentially harmful acts against them.

4.1 Evaluation

One of the most important forms of protection is the conduct of evaluations of big data initiatives prior to their implementation. These should identify in advance ideas whose potential benefits do not justify the negative impacts and risks, which should then lead to their substantial reworking or their abandonment.

An examination of business case preparation gives rise to serious doubts about its effectiveness as a means of protecting organisations against bad big data. Business cases evidence many variants, some disciplined and formalised, but most pragmatic and informal. Typically, they involve spreadsheet modelling, often with primarily financial data, and perhaps

cost-benefit analysis, but internal-only. The focus is on payback / Return on Investment (RoI), or on alignment with corporate strategy. However, all such approaches are more or less explicitly designed to provide support for the proposal [18]. Business case preparation provides inadequate protection even for the organisations that conduct them; still less do such processes protect against unjustifiable negative impacts on other parties.

Previous research by this author has considered big data risks from the perspective of the organisations that conduct the analytics and/or rely on the inferences they lead to. Data quality assurance should in principle be the means whereby the risks to those organisations can be avoided, detected, investigated and managed. Further, risk assessment should be the means of tackling not only data quality issues, but also the risks arising from data semantics, non-relevance, inappropriate data analytic techniques, and lack of transparency [12].

In practice, however, a large proportion of the negative outcomes of poor-quality big data and poor-quality big data analytics arise in the form of 'externalities': rather than the relevant organisation suffering them, someone else will. Most commonly the entities bearing the harm will be those that lack institutional and market power, sometimes small business enterprises, but most commonly people.

A much broader form of evaluation is needed than that provided by organisation-internal risk assessment processes. However, the incentives are such that organisations are not going to perform them, at least not of their own volition. Market failure exists, so government intervention is necessary.

4.2 Regulation

Where a proposal harbours serious threats, the protection of parties other than the proposal's sponsor depends on the conduct of a form of evaluation that takes into account the interests of all stakeholders. A consolidation of mainstream 'meta-principles' in relation to the evaluation of potentially harmful initiatives is in [2].

Beyond an obligation to conduct an appropriate form of evaluation, an effective regulatory scheme needs to be in place, to ensure that the findings from the process are carried through into actions. A variety of forms of regulatory arrangement exist, commonly referred to as organisational self-regulation, industry self-regulation, co-regulation and formal regulation. Nomatter which approach is adopted, an effective regulatory scheme needs to satisfy a range of requirements. A consolidation of criteria found in the literature is in Table 2 of [14].

Such industry and professional codes as exist in the big data analytics arena fail comprehensively when tested against criteria such as these. Examples include [34], and a flurry of recent initiatives whose intention is quite bare-facedly to hold off demands for formal regulatory measures [3, 16, 33]. For a scathing assessment of the UK 'ethical framework', see [27].

A specific regulatory mechanism that has been making considerable progress over the last two decades is Privacy Impact Assessment (PIA). A PIA is a systematic process, which identifies and evaluates from the perspectives of all stakeholders the potential effects on privacy of a project, initiative or proposed system or scheme and which includes a search for ways to avoid or mitigate negative privacy impacts [8, 38]. Unfortunately, evidence also exists that PIAs are not being effective in exercising control over inappropriate initiatives, particularly in national security contexts [13].

A recent development that has raised some people's hopes is the 'Data Protection Impact Assessment' (DPIA) requirement within the EU's General Data Protection Regulation [17], which is to come into force in 2018. The provisions are, however, very weak. The trigger is limited to "high risks" (Art. 35.1-35.6). The impacts to be assessed are only those on the protection of personal data (35.1) – which is a poor proxy even for data privacy, and which excludes other dimensions of privacy. It appears inevitable that DPIA will be interpreted as a mere Data Protection Law Compliance Assessment. Moreover, seeking civil society's views is optional, and there is no requirement that they be reflected in the design (35.9). There is a complete exemption for authorised programs (35.10), rather than merely an exemption from the justification requirement. And

it is far from clear whether any enforcement of design features will be feasible, and whether any review will ever be undertaken of the performance of schemes against the data used to justify them (35.7(d), 35.11).

A further factor that some argue to be a regulatory arrangement is the 'precautionary principle'. A strong form of this exists in some jurisdictions' environmental laws, along the lines of 'When human activities may lead to morally unacceptable harm that is scientifically plausible but uncertain, actions shall be taken to avoid or diminish that potential harm' [32]. However, no such strong form is applicable in human rights contexts. All that can be argued is that a moral responsibility exists, whereby, 'if an action or policy is suspected of causing harm, and scientific consensus that it is not harmful is lacking, the burden of proof arguably falls on those taking the action'.

These relevant and current examples all indicate that market failure is matched by regulatory failure. Neither parliaments nor regulatory agencies are providing effective restraints on organisational misbehaviour in the field of big data analytics.

4.3 Public Activism

In order to protect people's interests, public activism is needed. This could take the form of civil disobedience, in particular the obfuscation and falsification of data, traffic, location and identity. Further, pressure could be brought to bear on organisations, regulators and politicians, through coordinated actions focussed on a specific target, and the use of whatever communications channels are judged to have the greatest impact on that target at that particular point in time.

However, there are limited prospects of action by the general public, or even by the population segments most seriously affected by big data blunders. The issues are too complex and obscure for public discourse to cope with, and reduction to the simple slogans compatible with popular uprisings is very challenging. Another problem is that the regulatory failure noted in the previous section means that appropriate evaluation does not take place, and hence transparency is denied.

Rather than the general public, it appears more likely that the battle will be fought by advocacy organisations, called (originally in UN contexts) non-government organisations (NGOs), and collectively referred to as civil society. A review of privacy advocacy organisations around the world is in [4]. These associations have very little direct power, and limited resources. However, they have a wide range of techniques at their disposal [4, 15].

This author has previously proposed a further form that public activism can take. Civil society could abandon its half-hearted and ineffectual involvement in 'official' Standards processes conducted by industry and government. NGOs could develop, adopt, promulgate and promote their own series of Civil Society Standards [9]. Importantly in the big data context, these would specify principles and processes for evaluation, processes for quality assurance and audits, and checklists of mitigation measures and controls.

5 Conclusions

All uses of data involve issues with data semantics, data quality and information quality. However, the issues arising in conventional administrative systems are reasonably well-understood, and safeguards, controls, reviews, recourse and audit are factored into system designs.

Big data compounds the problems of data semantics, data quality and information quality. It merges data of uncertain and often low quality, and of often incompatible semantics, and it projects mystique rather than being founded on any real 'data science'.

Big data analytics then heaps further problems on the bonfire. One is the failure to provide a reliable way to identify appropriate techniques and to clearly specify the attributes that data must possess in order to justify processing in that manner. A second and very substantial problem is the lack of transparency inherent in contemporary analytical methods, whose rationality is not penetrable even by 'data science' specialists.

Inferences drawn by software that uses incomprehensible processes may be relied on to make decisions, and to take actions, variously affecting

categories of people, particularly through resource allocation, and affecting individuals, particularly through administrative decison-making. These decisions are, quite fundamentally, unreviewable, because the rationale underlying them cannot be communicated – and a rationale, in the sense in which humans understand the notion, may not even exist.

The consequences of big data inferences, and decisions and actions based on them, will inevitably be negative for some entities, and some categories of entities. Those entities will mostly be normal human beings. They will be denied meaningful review and recourse processes, because no comprehensible information is available. Weak regulators will be cowed by the accusation of stultifying innovation. Social equity, the social contract, and ultimately social order, will be the victims.

Protections are needed, which I've referred to in this paper as 'prophylactics', to underline the fact that they are a counterpoint to 'analytics'. In the foreground are evaluation processes; but these are shown not to work, and hence market failure is evident. Various forms of regulatory mechanism should in principle come into play; but multiple examples show that market failure is matched by regulatory failure. All that remains is public activism. The conditions are not right for the general public, or even the affected segments of the public, to take decisive action. Civil society is likely to have to fill the void, if big data prophylactics are to arise, to protect the public from inappropriate applications of big data analytics.

References

1. Anderson C. (2008) 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete' Wired Magazine 16:07, 23 June 2008, at http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory
2. APF (2013) 'Meta-Principles for Privacy Protection' Australian Privacy Foundation, March 2013, at http://www.privacy.org.au/Papers/PS-MetaP.html
3. ASA (2016) 'Ethical Guidelines for Statistical Practice' American Statistical Association, April 2016, at http://ww2.amstat.org/about/pdfs/EthicalGuidelines.pdf
4. Bennett C.J. (2008) 'The Privacy Advocates' The MIT Press, 2008
5. boyd D. & Crawford K. (2011) `Six Provocations for Big Data' Proc. Symposium on the Dynamics of the Internet and Society, September 2011, at http://ssrn.com/abstract=1926431

6. Clarke R. (1991) 'A Contingency Approach to the Software Generations' Database 22, 3 (Summer 1991) 23 - 34, PrePrint at http://www.rogerclarke.com/SOS/SwareGenns.html
7. Clarke R. (1995) 'A Normative Regulatory Framework for Computer Matching' J. of Computer & Info. L. 13,3 (June 1995), PrePrint at http://www.rogerclarke.com/DV/MatchFrame.html
8. Clarke R. (2009) 'Privacy Impact Assessment: Its Origins and Development' Computer Law & Security Review 25, 2 (April 2009) 123-135, PrePrint at http://www.rogerclarke.com/DV/PIAHist-08.html
9. Clarke R. (2010) 'Civil Society Must Publish Standards Documents' Proc. Human Choice & Computers (HCC9), IFIP World Congress, Brisbane, September 2010, pp. 180-184, PrePrint at http://www.rogerclarke.com/DV/CSSD.html
10. Clarke R. (2014) 'What Drones Inherit from Their Ancestors' Computer Law & Security Review 30, 3 (June 2014) 247-262, PrePrint at http://www.rogerclarke.com/SOS/Drones-I.html
11. Clarke R. (2016a) 'Big Data, Big Risks' Information Systems Journal 26, 1 (January 2016) 77-90, PrePrint at http://www.rogerclarke.com/EC/BDSA.html
12. Clarke R. (2016b) 'Quality Assurance for Security Applications of Big Data' Proc. European Intelligence and Security Informatics Conference (EISIC), Uppsala, 17-19 August 2016, PrePrint at http://www.rogerclarke.com/EC/BDQAS.html
13. Clarke R. (2016c) 'Privacy Impact Assessments as a Control Mechanism for Australian National Security Initiatives' Computer Law & Security Review 32, 3 (May-June 2016) 403-418, PrePrint at http://www.rogerclarke.com/DV/IANS.html
14. Clarke R. & Bennett Moses L. (2014) 'The Regulation of Civilian Drones' Impacts on Public Safety' Computer Law & Security Review 30, 3 (June 2014) 263-285, PrePrint at http://www.rogerclarke.com/SOS/Drones-PS.html
15. Davies S. (2014) 'Ideas for Change: Campaign principles that shift the world' The Privacy Surgeon, December 2014, at http://www.privacysurgeon.org/resources/ideas-for-change
16. DSA (2016) 'Data Science Code Of Professional Conduct' Data Science Association, undated but apparently of 2016, at http://www.datascienceassn.org/sites/default/files/datascicencecodeofprofessionalconduct.pdf
17. GDPR35 (2016) 'EU General Data Protection Regulation (EU-GDPR) Article 35 - Data protection impact assessment' at http://www.privacy-regulation.eu/en/35.htm
18. Humphrey W.S. (2000) 'Justifying a Process Improvement Proposal' SEI Interactive, March 2000, at http://northhorizons.com/Reference%2520Materials/%2520Justifying%2520a%2520PIP.pdf
19. Huh Y.U., Keller F.R., Redman T.C. & Watkins A.R. (1990) 'Data Quality' Information and Software Technology 32, 8 (1990) 559-565
20. Laney D. (2001) '3D Data Management: Controlling Data Volume, Velocity and Variety' Meta-Group, February 2001, at http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
21. LaValle S., Lesser E., Shockley R., Hopkins M.S. & Kruschwitz N. (2011) 'Big Data, Analytics and the Path From Insights to Value' Sloan Management Review (Winter 2011Research Feature), 21 December 2010, at http://sloanreview.mit.edu/article/big-data-analytics-and-the-path-from-insights-to-value/

22. McAfee A. & Brynjolfsson E. (2012) 'Big Data: The Management Revolution' Harvard Business Review (October 2012) 61-68
23. Mayer-Schönberger V. & Cukier K. (2013) 'Big data: A revolution that will transform how we live, work, and think' Houghton Mifflin Harcourt, 2013
24. Müller H. & Freytag J.-C. (2003) 'Problems, Methods and Challenges in Comprehensive Data Cleansing' Technical Report HUB-IB-164, Humboldt-Universität zu Berlin, Institut für Informatik, 2003, at http://www.informatik.uni-jena.de/dbis/lehre/ss2005/sem_dwh/lit/MuFr03.pdf
25. OECD (2013) 'Exploring Data-Driven Innovation as a New Source of Growth: Mapping the Policy Issues Raised by "Big Data"' OECD Digital Economy Papers, No. 222, OECD Publishing, at http://dx.doi.org/10.1787/5k47zw3fcp43-en
26. Piprani B. & Ernst D. (2008) 'A Model for Data Quality Assessment' Proc. OTM Workshops (5333) 2008, pp 750-759
27. Raab C. & Clarke R. (2016) 'Inadequacies in the UK Cabinet Office's Data Science Ethical Framework' Euro. Data Protection L. 2, 4 (Dec 2016) 555-560, PrePrint at http://www.rogerclarke.com/DV/DSEFR.html
28. Schroeck M., Shockley R., Smart J., Romero-Morales D. & Tufano P. (2012) 'Analytics : The real world use of big data' IBM Institute for Business Value / Saïd Business School at the University of Oxford, October 2012, at http://www.ibm.com/smarter-planet/global/files/se__sv_se__intelligence__Analytics_-_The_real-world_use_of_big_data.pdf
29. Shanks G. & Darke P. (1998) 'Understanding Data Quality in a Data Warehouse' The Australian Computer Journal 30 (1998) 122-128
30. Stilgherrian (2014a) ' Big data is just a big, distracting bubble, soon to burst' ZDNet, 11 July 2014, at http://www.zdnet.com/big-data-is-just-a-big-distracting-bubble-soon-to-burst-7000031480/
31. Stilgherrian (2014b) 'Why big data evangelists should be sent to re-education camps' ZDNet, 19 September 2014, at http://www.zdnet.com/why-big-data-evangelists-should-be-sent-to-re-education-camps-7000033862/
32. TvH (2006) 'Telstra Corporation Limited v Hornsby Shire Council' NSWLEC 133 (24 March 2006), esp. paras. 113-183, at http://www.austlii.edu.au/au/cases/nsw/NSWLEC/2006/133.htm
33. UKCO (2016) 'Data Science Ethical Framework' UK Cabinet Office, v.1, 19 May 2016, at https://www.gov.uk/government/publications/data-science-ethical-framework
34. UNSD (1985) 'Declaration of Professional Ethics' United Nations Statistical Division, August 1985, at http://unstats.un.org/unsd/dnss/docViewer.aspx?docID=93#start
35. Wand Y. & Wang R.Y. (1996) 'Anchoring Data Quality Dimensions in Ontological Foundations' Commun. ACM 39, 11 (November 1996) 86-95
36. Wang R.Y. & Strong D.M. (1996) 'Beyond Accuracy: What Data Quality Means to Data Consumers' Journal of Management Information Systems 12, 4 (Spring, 1996) 5-33
37. Widom J. (1995) 'Research Problems in Data Warehousing' Proc. 4th Int'l Conf. on Infor. & Knowledge Management, November 1995, at http://ilpubs.stanford.edu:8090/91/1/1995-24.pdf

38. Wright D. & De Hert P. (eds) (2012) 'Privacy Impact Assessments' Springer, 2012

Acknowledgements. This paper was developed from my opening keynote invited for the IFIP Summer School on Privacy and Identity Management, Karlstad, Sweden, 22 August 2016.