2015

Quasi-Empirical Scenario Analysis and Its Application to Big Data Quality

Roger Clarke
Xamax Consultancy Pty Ltd, Australia Research School of Computer Science, ANU, Canberra, Visiting Professor, Faculty of Law, UNSW, Sydney, roger.clarke@xamax.com.au

28th Bled eConference

#eWellBeing

June 7 - 10, 2015; Bled, Slovenia

Quasi-Empirical Scenario Analysis and Its Application to Big Data Quality

Roger Clarke

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

Abstract

Big data and big data analytics have been the subject of a great deal of positive discussion, not only in traditionally upbeat popular management magazines but also in nominally scientific and therefore professionally sceptical academic journals. Research was undertaken to assess the impact on the quality of inferences drawn from big data of the quality of the underlying data and the quality of the processes applied to it.

Empirical study is difficult, however, because big data is emergent, and hence the phenomena are unstable, and likely to vary considerably across different settings. A research technique was accordingly sought that enabled theoretical treatment to be complemented by consideration of real-world data. The paper introduces Quasi-Empirical Scenario Analysis, which involves plausible story-lines, each commencing with a real-world situation and postulating lines of plot-development. This enables interactions among factors to be analysed, potential outcomes to be identified, and hypotheses to be generated.

The set of seven scenarios that was developed investigated the nature and impacts of shortfalls in data and decision quality in a range of settings. In all cases, doubts arise about the reliability of inferences that arise from big data analytics. This in turn causes concern about the impacts of big data analysis on return on investment and on public policy outcomes. The research method was found to offer promise in the challenging contexts of technologies in the process of rapid change.

Keywords: big data analytics, data quality, decision quality, scenario analysis

1 Introduction

As sensor technologies have matured, and as individuals have been encouraged to contribute data into organisations' databases, more transactions than ever before have been captured. Meanwhile, improvements in data-storage technologies have resulted in the cost of evaluating, selecting and destroying old data being now considerably higher than that of simply letting it accumulate. The glut of stored data has greatly increased the opportunities for data to be inter-related, and analysed. The moderate enthusiasm engendered by 'data warehousing' and 'data mining' in the 1990s has been replaced by unbridled euphoria about 'big data' and 'data analytics'. What could possibly go wrong?

Professional and management periodicals on big data topics have a very strong focus on opportunities, and so do most academic papers in the area to date. Far too little attention has been paid to the threats that arise from re-purposing data, consolidating data from multiple sources, applying analytical tools to the resulting collections, drawing inferences, and acting on them. The research reported on in this paper was conceived as a way to test the key claims made in the 'big data' literature, together with some of its implicit assumptions.

The paper commences by briefly reviewing the emergent theory and current practice of big data. Working definitions are provided that distinguish the processes whereby the data becomes available from the processes whereby inferences are drawn from it. A short summary is then provided of theory relating to the central concepts of data quality and decision quality. Formal empirical research techniques are difficult to apply to emergent phenomena. A research technique is described which is intended to provide a quasi-empirical base to complement the theoretical analysis. A set of scenarios is outlined, and the theoretical material is then used as a lens to gain insights into the nature of big data and big data analytics, and the risks that they entail.

2 Big Data and Big Data Analytics

During the 1980s, organisations had multiple data collections that were largely independent of one another. In order to enable the manipulation of data structures and content without disrupting underlying operational systems, copies of data were extracted from two or more collections and stored separately, in what was referred to as a 'data warehouse' – "a copy of transaction data specifically structured for query and analysis" (Jacobs 2010. See also Inmon 1992 and Kimball 1996). The processing of the contents of 'data warehouses' was dubbed 'data mining' (Fayyad et al. 1996, Ratner 2003, Ngai et al. 2009, Hall et al. 2009).

During the same era, government agencies that could gain access to data-sets from multiple sources practised an additional technique called data matching. This involves comparing machine-readable records from different data-sets that contain data that appears to relate to the same real-world entities, in order to detect cases of interest. In most data matching programs, the category of entities that is targeted is human beings (Clarke 1994b, 1995a).

A further technique that has long been applied in both the public and private sectors is profiling. This is "a technique whereby a set of characteristics of a particular class of person is inferred from past experience, and data-holdings are then searched for

individuals with a close fit to that set of characteristics" (Clarke 1993). Although profiles can be *ad hoc*, they have been increasingly "supported by analyses of existing data-holdings within and beyond the organisation, whereby individuals who are known to belong to that class are identified, their recorded characteristics examined, and common features isolated".

During the last few years, there has been a resurgence in enthusiasm for such techniques. The range and intensity of data capture has greatly increased in the intervening years. In addition, the economics of storage and destruction have shifted, such that retaining data is now cheaper than deleting it. This has led a wide variety of media commentators, consultants and excitable academics making enthusiastic pronouncements about revolutions, break-throughs and sparkling new opportunities.

One of the most-quoted authorities on big data is Mayer-Schonberger & Cukier (2013), which had accumulated over 700 citations in its first 21 months after publication. According to those authors, the cornerstone of big data thinking is that 'datafication' – the expression of phenomena "in a quantified format so it can be tabulated and analyzed" (p.78). This they say undermines the kinds of analyses conducted in the past: "With enough data, the numbers speak for themselves. Petabytes allow us to say: Correlation is enough ... [W]hen you are stuffed silly with data, you can tap that instead [of experience, expertise and knowledge], and to greater effect. Thus those who can analyze big data may see past the superstitions and conventional thinking not because they're smarter, but because they have the data ... [S]ociety 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 ... [L]et the data speak" (pp. 71, 143, 7, 52, 141).

At its most extreme, this argument postulates that understanding is no longer necessary, and that the ready availability of vast quantities of data justifies the abandonment of reason: "If the statistics ... say it is, that's good enough. No semantic or causal analysis is required. ... [M]assive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. ... [F]aced with massive data, [the old] approach to science — hypothesize, model, test — is becoming obsolete. ... Petabytes allow us to say: 'Correlation is enough'" (Anderson 2008).

Such bombast is common in papers on big data. More sceptical views do exist, however. For example, big data is "a cultural, technological, and scholarly phenomenon that rests on the interplay of [three elements]" boyd & Crawford (2012, p.663). Their first two elements are technical, and are examined below. 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".

Where definitions of big data are offered, they tend to use imprecise and upbeat explanations, and to use a single term to encompass both the data and the techniques applied to it. A common expression is 'data that's too big, too fast or too hard for existing tools to process', and many writers refer to the distinguishing features being 'volume, velocity and variety'. Even the definition offered by an international agency is vague: "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" (OECD 2013, p.12). An aspect that appears in some discussions is the accumulation of measurements of the same phenomenon over time, resulting in a longitudinal dimension within the data collection.

This paper distinguishes 'big data' from 'big data analytics', and proposes working definitions, listed in Table 1.

'Big Data'
A relatively large data collection that is associated with, or purports to be associated with, or can be interpreted as being associated with, particular entities. The following sub-categories can be usefully distinguished:

- A single very large data collection
- A consolidation of two or more data collections in such a manner that the association of data with specific entities is sustained or achieved. Three sub-categories can be usefully distinguished:
  - Merger into a single physical data collection
  - Interlinkage into a single virtual data collection, by means of:
    - Ephemeral Links, which exist only during the running of a particular analysis
    - Stored Links, which support multiple analyses

'Big Data Analytics'
The techniques whereby a big data collection is used to draw inferences about an entity, or a category of entities, or relationships among entities

Table 1: Working Definitions

Earlier research performed by the author distinguished a number of categories of purpose to which big data analytics may be applied. These are extracted in Table 2.

Big data analysis may test hypotheses about populations, which may be predictions from theory, existing heuristics, or hunches. Inferences may be drawn about the digital personae, which may be further applied to the population of entities that the data purports to represent, or to segments of that population. Profiles may be constructed for categories of entities that are of specific interest, such as heavy weather incidents, risk-prone card-holders, wellness recipes, or welfare cheats.

Further functions that can be performed are concerned not with populations but with individual instances. Outliers of many different kinds can be identified. Inferences can be drawn about individual entities, directly from a particular digital persona, or from apparent inconsistencies within a consolidated digital persona, or through comparison of a digital persona against the population of similar entities, or against a previously-defined profile, or against a profile created through analysis of that particular big data collection.

Population Focus

- **Hypothesis Testing**
  This approach evaluates whether a particular proposition is supported by the available data
- **Population Inferencing**
  This approach draws inferences about the entire population of (id)entities, or about sub-populations
- **Profile Construction**
  This approach identifies key characteristics of some category of (id)entities.

Individual Focus

- **Outlier Discovery**
  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'
- **Inferencing about Individuals**
  This approach draws inferences about individual entities within the population.

Table 2: Functions of Big Data Analytics – After Clarke (2014b)

The performance of these functions places a great deal of reliance on the analytical tools and the big data collections to which they are applied. The purpose of the research reported in this paper is to assess the impact of the quality of the big data, and of the big data analytics, on the quality of inferences drawn. The following section juxtaposes against big data notions the understanding accumulated within the information systems literature of the characteristics of data quality and decision quality.

3 Quality Theory

During its first 50 years, the information systems (IS) discipline has adjusted its focus over time. Underlying many threads of IS theory, however, are characteristics of data that influence the effectiveness and efficiency of decision-making based on it. This section briefly summarises key aspects of data and decision quality.

3.1 Data Quality

The computing and information systems literatures contain a body of material relating to data quality factors. See, for example, OECD (1980), Huh et al. (1990), Clarke (1995b, pp. 601-605), Wang & Strong (1996), Rusbridge et al. (2005), English (2006), Piprani & Ernst (2008), and the still-emergent ISO 8000 series of Standards. The primary factors are listed in Table 3.

- **Accuracy**

  The 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 'accurate to within 1 degree Celsius'

- **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'

- **Timeliness**, which comprises distinct elements:

  - **Temporal Applicability**
    This reflects, for example, the date and time when the temperature was measured, the period during which an income-figure was earned, and the date after which a qualification or licence was applicable

  - **Up-to-Dateness**
    This reflects the lag between the real-world occurrence and the recording of the corresponding data. This is relevant to volatile items such as the current temperature and the balance owing on a credit account

  - **Currency**
    This reflects when the data-item was captured or was last authenticated, or the period over which an average was computed. This is relevant to volatile data-items, such as average temperature, total rainfall for the last 12 months, age, marital status and fitness for work

- **Completeness**
  The availability of sufficient contextual information that the data is not liable to be misinterpreted

Table 3: The Primary Data Quality Factors

Each item of data is a measurement against some kind of scale. Some data arises from measurement against a ratio scale, and is capable of being subjected to analysis by powerful statistical tools. Frequently, however, data arises from measurements against cardinal, and against merely ordinal scales, such as Likert-scale data. Such data is capable of analysis using a smaller collection of statistical tools. However, researchers often assume for convenience that the data was gather against a ratio scale, in order to justify the application of more powerful inferencing techniques. Meanwhile, a great deal of data is collected against nominal scales, including text, sound, images and video. This supports only weak analytical tools, fuzzy matching and fuzzy logic. A further challenge arises where data that has been measured against different kinds of scales is consolidated and analysed. The applicability of analytical tools to mixed-scale data is more of a murky art than a precise science.

A further consideration when assessing the quality of data is that what is collected, against what scales, and with what trade-offs between data quality and collection costs, all reflect the purpose of collection and the value-judgements of the sponsor. Particularly where data is collected frequently over time, data collection may also involve compression, through sampling, filtering and averaging. Further, where interesting outliers are being sought, compression is likely to ensure that the potentially most relevant data is absent from the collection.

To be useful, data needs to be associated with real-world entities, with the data currently held about any particular entity being that entity's 'digital persona' (Clarke 1994, 2014a). The reliability of the association depends on the attributes that are selected as identifiers, and the process of association has error-factors. In some circumstances the link between the digital persona and the underlying entity is unclear (the condition referred to as pseudonymity), and in some cases no link can be achieved (anonymity). In order to protect important interests and comply with relevant laws, the link may be broken (the process of de-identification or anonymisation – UKICO 2012, DHHS 2012). On the other hand, rich data-sets are vulnerable to re-identification procedures. These problems afflict all big data collections, particularly those that are intended to support longitudinal studies. Where the data is sensitive, significant public policy issues arise.

Over time, many threats arise to data integrity, including the loss of metadata such as the scale against which data was originally collected, the data's definition at the time of collection, the data's provenance, any supporting evidence for the data's quality, undocumented changes in meaning over time, and loss of contextual information that would have enabled its appropriate interpretation.

The big data movement commonly features the use of data for a purpose extraneous to its original purpose – and indeed many big data proponents champion this proposition. The many data quality issues identified above are exacerbated by the loss of context, the lack of clarity about the trade-offs applied at the time of collection, and the greatly increased likelihood of misinterpretation.

A further aspect of the big data movement is the step of physically or virtually consolidating data from multiple sources. This depends on linkages among data-items whose semantics and syntactics are different, perhaps substantially, perhaps subtly. The scope for misunderstandings and misinterpretation multiply.

Theorists and practitioners perceive deficiencies in the data, such as missing elements, and inconsistencies among seemingly similar data-items gathered from two or more sources. To address these concerns about the analysts' particular perceptions of data quality, they devise data 'scrubbing', 'cleansing' or 'cleaning' processes. Some of these processes use an external, authoritative reference-point, such as a database of recognised location-names and street-addresses. Most, however, lack any external referent, and are merely based on 'logical data quality', i.e. internal consistency within the consolidated data-sets: "rules are learned from data, validated and updated incrementally as more data is gathered and based on the most recent data" (Saha & Srivastava 2014. See also Jagadish et al. 2014). Such process guidance as exists overlooks intrinsic and contextual data quality, and omits controls and audit (e.g. Guo 2013). As a result, the notion of 'cleanliness' primarily relates to the ease with which analytical tools can be applied to the data, rather than to the quality of the data.

3.2 Decision Quality

It is feasible for big data analytics to be applied as a decision system. This may be done formally, by, for example, automatically sending infringement or 'show cause' notices. However, it is also possible for big data analytics to become a decision system not through conscious intent by an organisation, but by default. This can arise where a decision-maker becomes lazy, or is replaced by a less experienced person who is not in as good a position to apply human intelligence as a means of checking the reasonableness of the inferences drawn by software.

Where decisions are made by analytics, a number of concerns arise about decision-quality. In Clarke (2014b), three were highlighted:

- **Relevance**
  The relevance of the particular data to the particular decision needs to be demonstrated
- **Meaning**
  The meaning of each particular item of data needs to be clear, as does the meaning of each particular value that each data-item adopts. The meaning of each individual item of data is capable of definition at the time it is gathered. However, since the decline of the resource-intensive waterfall method of software development, it is much less common for a data dictionary to be even established, let alone maintained. As a result, data definitions may be unclear, ambiguous and even implicit. The lack of clarity about the original meaning increases the likelihood that the meaning will be subject to various interpretations at any given time, that its meaning(s) will change over time, and hence that different interpretations of the same data-item or its content may be current, and that conflict may exist among mutually inconsistent interpretations
- **Transparency**
  The decision-mechanism needs to be sufficiently transparent that those depending on it, those affected by it, and those reviewing it, can understand the basis on which the decision was made. Concerns have been expressed about transparency by a range of authors, e.g. Roszak (1986), Dreyfus (1992), boyd & Crawford (2012)

Those theoretical aspects lead to practical questions. On what scale, with what accuracy and what precision, was the data collected that was instrumental in leading to the decision, and was the inferencing mechanism that was used really applicable to those categories of data? Did the data mean what the inferencing mechanism implicitly treated it as meaning? To the extent that data was consolidated from multiple sources, were those sources compatible with one another in respect of the data's scale, accuracy, precision, meaning and integrity?

In many circumstances to which big data is claimed to be applicable, real-world decisions depend on complex models that feature confounding, intervening and missing variables. Correlations are commonly of a low grade, yet may nonetheless be treated, perhaps implicitly, as though the relationships were causal, and causal in one direction

rather than the other. Big data proponents blithely dismiss causality and champion correlation instead. Yet *mens rea* (i.e. intention to cause the outcome) is a fundamental element of most criminal prosecutions, and many decisions in civil jurisdictions also focus on the proximate cause of an event. Big data proponents are swimming against a strong tide of cultural and institutional history. If decisions are being made that have real impacts, is the decision-process, and are the decision-criteria, transparent? And are they auditable? Are they subject to appeal processes and review? And can society have confidence that risks and liabilities are appropriately allocated?

Big data analytics may be more commonly used as a form of decision support system, whereby a human decision-maker evaluates the inferences before applying them to any real-world purpose. However, the person may have great difficulty grasping the details of data provenance, data quality, data meaning, data relevance, and the rationale that have given rise to the recommendation. An even more problematic situation arises where the nominal decision-maker is not in a position to appreciate the rationale underlying the 'recommendation' made by the analytical procedure, and hence feels themselves to be incapable of second-guessing the system.

With what were once called 'third-generation' development tools, the rationale was evident in the form of an algorithm or procedure, which may have been explicitly documented externally to the software, but was at least extractable by any person who had access to the source-code and who had the capacity to read it. The fourth generation of development tools merely expressed the decision-model in a more generally-applicable manner. The advent of the fifth-generation adopted a different approach, however. Rather than a model of the decision, this involved a model of the problem-domain. It became much more difficult to understand how the model (commonly expressed in logic, rules or frames) applied to particular circumstances. Some tools can provide a form of explanation of the rationale (or at least a list of the rules that were fired), but many cannot.

Subsequently, with the sixth generation, an even greater barrier to understanding arose. With a neural network, there is no formal model of a decision or even of a problem-domain. There is just an empirical pile, and a series of inscrutable weightings that have been derived through mathematical processes, and that are then applied to each new instance (Clarke 1991). There have been expressions of concern from many quarters about the delegation of decision-making to software whose behaviour is fundamentally unauditable (e.g. Roszak 1986, Dreyfus 1992, boyd & Crawford 2012).

4 The Research Method

An examination of the big data literature, reported in Clarke (2014b), concluded that, to date, there is limited evidence of the body of knowledge about data quality and decision quality being applied by the 'big data' movement, either by practitioners or researchers. The research question addressed by this research was accordingly:

> *What is the impact of the quality of big data, and of big data analytics, on the quality of inferences drawn?*

In order to address that question, it is strongly desirable to conduct studies of real-world phenomena. However, conventional empirical research techniques are founded on some key assumptions, most relevantly that:

- stable theories exist relevant to the research domain;
- hypotheses about the research domain can be generated from those theories, which are explicit, unambiguous and refutable; and
- the relevant aspects of the domain can be observed and measured, in such a manner that the hypotheses, if false, can be shown to be so.

The big data research domain has a number of characteristics that present serious challenges to the conduct of empirical research. These are summarised in Table 4.

- **Unobservable Phenomena**:
  - To some extent the big data domain is speculative and the phenomena do not, or do not yet, exist
  - To the extent that big data is being practised, the phenomena are difficult for researchers to observe, in particular due to the desire for secrecy of the corporations, government agencies and service-providers that are managing the data and conducting the analyses, variously because of the potential for competitive advantage and the risk that the activity is in breach of confidentiality or data protection laws
- **Unstable Phenomena**:
  - The practices in relation to the collection and consolidation of data are changing
  - The analytical techniques being applied to data are also in a state of flux
  - To the extent that the data collection is longitudinal, the behaviours may be changing, rather than merely varying within a stable distribution or pattern
- **Highly Context-Dependent Phenomena**:
  - The practices depend on the nature of the data, the expertise of the analysts, and the nature of the organisation by which or on whose behalf the big data is being gathered and the analyses are being performed

Table 4: Characteristics of the Big Data Research Domain

These characteristics present serious theoretical challenges. The selection and application of a body of theory is made difficult because each theory is relevant only to particular entities, relationships and/or environmental circumstances, and the underlying model may map to the reality at one time, but not at another. In addition, inferences drawn from the new context may or may not be relevant to the body of theory that was used to guide the research design. The characteristics also present practical challenges, because there are difficulties in reliably defining populations, population segments, and sampling frames, and in defining what is to be observed and what the terms mean that are used in interviews, questionnaires and case study reports.

Another category of research can be envisaged, which is referred to here as 'quasi-empirical'. This draws on the technique of scenario analysis that has been used in futurology and long-term strategic planning since the 1960s, and is well-described in Wack (1985), Leemhuis (1985), Mobasheri et al. (1989), Schwarz (1991) and van der Heijden (1996).

A scenario is a story-line that represents a composite or 'imagined but realistic world'. The prefix 'quasi' is an appropriate qualifier, because it is from the Latin, in use in English since the 15th century, meaning 'resembling', or 'seemingly but not actually'. Scenario analysis involves the preparation of a small set of scenarios that have the potential to provide insights into an emergent new context.

The Quasi-Empirical Scenario Analysis (QuESA) technique proposed here involves a set of scenarios within which analytics are applied to big data collections. Each scenario is empirical to the extent that it commences with a setting-description that is drawn from contemporary phenomena. A story is then developed from the elements within the setting, by overlaying further data that is surmised, inferred or postulated, but is plausible or at least tenable, and by applying social and economic process sequences that are commonly observed in comparable settings.

The purpose of the QuESA technique is emphatically not to generate assertions of predictive, explanatory or even descriptive standing. The intention is to assist in the emergence of insights, and in the formulation of hypotheses for testing by means of conventional empirical techniques as phenomena emerge, stabilise and become observable. In some respects, the QuESA technique might be compared with focus group research, in that the purpose of both is not to formally test hypotheses, but to gain insights and to bring to the surface potential hypotheses.

In the QuESA technique, each scenario is built from a factual base, utilises real or at least realistic story elements, and is presented using narrative logic, i.e. plausible interactions and sequences. A single scenario is too limiting, firstly because it cannot capture sufficient of the richness of the research domain, and secondly because of the risk that the researcher or readers of reports on the research will lapse into predictive thinking rather than recognising the technique's purposes and limitations. Hence multiple scenarios are developed, addressing a requisite diversity of contexts, postulating events, interactions and sequences that are plausible within the particular context associated with the factual base.

5 Conduct of the Research

This section describes the manner in which the QuESA technique was applied to big data and big data analytics, and presents the findings that resulted from the work.

5.1 The Scenarios

A range of contexts was identified from the big data literature. In most cases, reports exist of applications of the specific kind referred to in the scenario, although in a few cases the application may be at this stage only aspirational. An endeavour was made to achieve diversity in the settings, in the nature of the data, in the nature of the analytical tools applied to it, and in the type of function being performed. It is infeasible to attempt representativeness in the sample, because the population is as-yet ill-defined, and indeed the dimensions across which the population varies are still under investigation.

The text for the seven Scenarios is in Appendix 1. The basis on which each Scenario was developed is explained in Supplementary Materials on the author's web-site. See: http://www.rogerclarke.com/EC/BDSA.html#App2

Some of the Scenarios involve data that identifies or may identify specific individuals. This is the case with (2) Creditworthiness, (4) Foster Parenting, (6) Fraud Detection and (7) Insider Detection. In (3) Ad Targeting, the data relates to online identities that may but may not relate to a particular human being. One instance depends on de-identified patient data – (5) Cancer Treatment; whereas in Scenario (1) Precipitation Events, the data relates to environmental phenomena.

In each case, starting with the factual base, various events or trajectories were postulated, which appeared to be tenable for that particular context. A narrative was then developed, and a story-line written. For example, Scenario (5) Cancer Treatment is based on a widely-quoted case in Mayer-Schonberger & Cukier (2013). An impact is postulated on research funding policies, which is a natural corollary of such a discovery. Because of the tight linkage between medical science research and industry, an associated development in the pharmaceutical industry is postulated. The impact on the population of young researchers is a natural demographic consequence. It is commonly the case in empirical research that subsequent work establishes that the correlations that were initially discovered were not as they seemed, and that the inferences drawn from the correlations need substantial qualification; and it is surmised that just such an eventuality will occur in this context as well.

5.2 Findings

The inferences that are drawn from Quasi-Empirical Scenario Analysis are by definition not empirically based, but are 'insights' that arise from a plausible story-line, not from observation of the real world or of some more or less carefully-controlled proxy for the real world. The 'findings' reported in this section are accordingly hypotheses. The credibility of the hypotheses is weaker than that which arises from the application of a demonstrably relevant theory to a research domain – which are of course to be preferred if such a theories are available. On the other hand, the credibility of the hypotheses is considerably greater than that of *ad hoc* propositions and hunches, and greater than ideas generated on the basis of anecdotes alone.

The first data quality issue discussed above related to the analysis of big data that was gathered against varying data scales. Challenges of these kinds are evident in Scenarios (3) Ad Targeting, (6) Fraud Detection and (7) Insider Detection. For example, data about online identities includes nominal data (interests), binary data although possibly with some data missing (gender), ordinal data (age-group), ratio-scale data (transaction counts and proportions), and longitudinal data. Few analytical techniques that are supported by mathematical statistics theory are available to cope with such mixed-mode data.

The relationship between purpose of collection and the investment in original data quality was identified as a factor that may undermine the quality of inferences drawn. Administrative actions in relation to fraud, and particularly prosecutions, are undermined by poor-quality evidence. In some contexts, commercial liability might arise, e.g. Scenario (2) Creditworthiness, while in others a duty of care may be breached, e.g. Scenario (4) Foster Parenting.

Loss of quality through inappropriate or inconsistent data compression, and through inappropriate or inconsistent handling of missing data over a timeline, arise in Scenario (1) Precipitation Events.

Various uncertainties arise about identity, and hence about whether the data from different sources, and data collected over time, actually relate to the same real-world entity. This is significant in big health data contexts generally, e.g. Scenario (5) Cancer Treatment, and big social data applications, e.g. Scenario (4) Foster Parenting. The issues also loom large in Scenario (3) Ad Targeting, and – given the prevalence of identity fraud – in Scenario (2) Creditworthiness.

Public demands for anonymisation/deidentification, and active endeavours to obfuscate and falsify identity, can be reasonably expected to exacerbate these challenges. Retrospective studies across long periods, as in Scenario (1) Precipitation Events, (4) Foster Parenting and (6) Fraud Detection, are confronted by problems arising from multiple sources with inconsistent or unclear syntax and semantics.

The vagaries of 'data cleansing' techniques ensure that there will be instances of worsened data quality, resulting in both wrong inferences and spurious matches. This is a problem shared by all big data projects that depend on data consolidated from multiple sources. It afflicts Scenarios as diverse as (1) Precipitation, (5) Cancer Treatment and (7) Insider Detection.

Of the decision quality issues, dubious relevance appears as an issue in (2) Creditworthiness, (5) Cancer Treatment, (6) Fraud Detection and (7) Insider Detection. In Scenario (6) Fraud Detection, data meaning is an issue, in that suspects may be unjustifiably identified as a result of inconsistencies arising from attempts to deceive, and from semantic issues even within a single database, let alone within a consolidation of multiple, inherently incompatible databases. Meanwhile, transparency of the decision mechanism results in the misallocation of resources in Scenario (5) Cancer Treatment, and in unfair discrimination in (2) Creditworthiness and (7) Insider Detection.

6 Conclusions

Many applications of big data and big data analytics are not currently able to be studied because the organisations conducting them do not provide ready access. This paper has introduced and applied a new research technique, dubbed QuESA. Its purpose is to enable research to be undertaken in circumstances in which the gathering of the data necessary to support empirical research is not possible, not practicable, or not economic.

The scenarios that have been used in this research are not case studies of specific instances of big data at work. They are story-lines, devised in order to encompass a range of issues not all of which are likely to arise in each particular real-life application. Each plot-line builds on factual foundations, then infiltrates into the stories additional elements that are plausible in the particular context. The intention of the scenarios was to test the assumptions underlying the big data value-proposition, not to pretend to be a substitute for deep case studies of actual experience.

The big data movement involves data-collections that are of uncertain original quality, and lower current quality, and that have uncertain associations with real-world entities. Data collections are combined, by means of unclear validity, and modified by unaudited means, in order to achieve consolidated digital personae that have uncertain validity, and have uncertain relationships with any particular real-world entity. To this melange, powerful analytical tools are then applied. Theoretical analysis, complemented by the findings from the QuESA study, identifies a wide range of circumstances in which these problems can arise.

Big data proponents claim that 'more trumps better'. More specifically: "With big data, the sum is more valuable than its parts, and when we recombine the sums of multiple datasets together, that sum too is worth more ... [W]e no longer need to worry so much about individual data points biasing the overall analysis" (Mayer-Schonberger & Cukier 2013, pp. 108, 40). The over-claiming by those authors extends beyond the general (where, in some circumstances, the much-overstated 'law of large numbers' actually does apply) to the specific: "Big data gives us an especially clear view of the granular" (p.13). As summarised by Leonelli (2014, p.3): "Big Data is viewed, through its mere existence, as countering the risk of bias in data collection and interpretation". These assertions lack adequate support by either theoretical argument or empirical evidence. In the face of the analysis reported on in this paper, they must be regarded as being at best wishful thinking, or self-delusional, or – because they are so often framed as recommendations to executives – reckless or even fraudulent.

Given the uncertain quality of data and of decision processes, it appears that many inferences from big data projects may currently be being accorded greater credibility than they actually warrant. If that is the case, then resources will be misallocated. Within corporations, the impact will ultimately be felt in lower return on investment, whereas in public sector contexts there will be negative impacts on public policy outcomes, such as unjustified discrimination against particular population segments.

When big data analytics are inappropriately applied to population inferencing and profile-construction, the harm that can arise includes resource misallocation and unjustified discrimination. When, on the other hand, inappropriate inferencing is about specific individuals, the costs are inevitably borne not by the organisation(s) involved but by the individuals, sometimes in the form of inconvenience, but sometimes with

financial, service-availability, psychological, discrimination, or natural justice dimensions.

When profiles generated by big data analytics are applied in order to generate suspects, the result is an obscure and perhaps counter-intuitive "predetermined characterisation or model of infraction" (Marx & Reichman 1984, p.429), based not on 'probable cause', but on a merely 'probabilistic cause' (Bollier 2010, pp.33-34). Not only does this result in unjustified impositions on the individuals concerned, but it also denies them natural justice because the lack of transparency relating to data and decision criteria means that the accusations are mysterious and even undefendable and they cannot get a fair hearing.

The computer science and management literatures are remarkably lacking in discussion of the issues and the impacts examined in this paper. As a result, they have yet to mature into a literature on appropriate business processes for acquiring and consolidating big data, and applying big data analytics. The information systems literature could be expected to be more sceptical, and more helpful to decision-makers in business and government. Yet, as at the end of 2014, the entire AIS electronic library contained but one paper whose Abstract included both 'big data' and either 'risk assessment' or 'risk management'. The results of the research reported in this paper, limited though it was to an analytical and quasi-empirical base, make clear that much more care is warranted.

Appendix 1: Big Data Scenarios

(1) Precipitation Events

Historical rainfall data has been gathered from many sources, across an extended period, and across a range of geographical locations. The collectors, some of them professional but mostly amateurs, used highly diverse collection methods, with little calibration and few controls. The data is consolidated into a single collection. A considerable amount of data manipulation is necessary, including the interpolation of data for empty cells, the arbitrary disaggregation of long-period data into the desirable shorter periods, and, in some cases, arbitrary disaggregation and reaggregation of data (e.g. to reconcile midnight-to-midnight with dusk-to-dusk recording times).

Attempts are made to conduct quality audits against such sources as contemporaneous newspaper reports. However, these prove to be too slow and expensive, and are curtailed. Analytical techniques are applied to the data. Conclusions are reached about historical fluctations and long-term trends, with appropriate qualifications expressed. Analysts then ignore the qualifications and apply the data as though it were factual rather than a mix of facts and interpolations. Climate-change sceptics point to the serious inadequacies in the database, and argue that climate-change proponents, in conducting their crusade, have played fast and loose with scientific principles.

(2) Creditworthiness

A financial services provider combines its transactions database, its records of chargebacks arising from fraudulent transactions, and government statistics regarding the geographical distribution of income and wealth. It draws inferences about the risks that its cardholders create for the company. It uses those inferences in its decision-making about individual customers, including credit-limits and the issue of replacement and upgraded cards.

Although not publicised by the company, this gradually becomes widely known, and results in negative media comments and recriminations on social media. Questions are raised about whether it conflicts with 'redlining' provisions in various laws. Discrimination against individuals based on the behaviour of other customers of merchants that they use is argued to be at least immoral, and possibly illegal, but certainly illogical from an individual consumer's perspective. The lender reviews the benefits arising from the technique, the harm done to its reputation, and the trade-off between the two.

(3) Ad Targeting

A social media services provider accumulates a vast amount of social transaction data, and some economic transaction data, through activity on its own web-sites and those of strategic partners. It applies complex data analytics techniques to this data to infer attributes of individual digital personae. Based on the inferred attributes of online identities and the characteristics of the available materials, the service-provider allocates third-party ads and its own promotional materials to the available space on web-pages.

The 'brute force' nature of the data consolidation and analysis means that no account is taken of the incidence of partial identities, conflated identities, and obfuscated and falsified profiles. This results in mis-placement of a significant proportion of ads, to the

detriment mostly of advertisers, but to some extent also of individual consumers. It is challenging to conduct audits of ad-targeting effectiveness, and hence advertisers remain unaware of the low quality of the data and of the inferences. The nature of the data exploitation achieves a considerably higher level of public consciousness as a result of the increasing incidence of inappropriate content appearing on childrens' screens.

(4) Foster Parenting

A government agency responsible for social welfare programs consolidates data from foster-care and unemployment benefits databases, and discovers a correlation between having multiple foster parents and later being chronically unemployed. On the basis of this correlation, it draws the inference that the longstanding practice of moving children along a chain of foster-parents should be discontinued. It accordingly issues new policy directives to its case managers.

Because such processes lack transparency, and foster-children are young and largely without a voice, the new policy remains 'under the radar' for some time. Massive resistance then builds from social welfare NGOs, as it becomes apparent that children are being forced to stay with foster-parents who they are fundamentally incompatible with, and that accusations of abuse are being downplayed because of the forcefulness of policy directives based on mysterious 'big data analytics'.

(5) Cancer Treatment

Millions of electronic medical records reveal that cancer sufferers who take a certain combination of aspirin and orange juice see their disease go into remission. Research funding agencies are excited by this development, and transfer resources to 'big health data analytics' and away from traditional systemic research into causes, pathways and treatments of disease.

Pharmaceutical companies follow the trend by purchasing homeopathic suppliers and patenting herb genes. The number of doctoral and post-doctoral positions available in medical science drops sharply. After 5 years, enough data has become available for the conclusion to be reached that the health treatments 'recommended' by these methods are ineffectual. A latter-day prophet emerges who decries 'the flight from reason', fashion shifts back to laboratory rather than digital research, and medical researchers slowly regain their previous high standing. The loss of momentum is estimated to have delayed progress by 10-15 years and generated a shortage of trained medical scientists.

(6) Fraud Detection

A company that has large sums of money flushing through its hands is under pressure from regulators, knows that stock exchanges run real-time fraud detection schemes, and accepts at face value the upbeat claims made by the proponents of big data analytics. It combines fraud-detection heuristics with inferences drawn from its large transaction database, and generates suspects. It assigns its own limited internal investigation resources to these suspects, and refers some of them to law enforcement agencies.

The large majority of the cases investigated internally are found to be spurious. Little is heard back from law enforcement agencies. Some of the suspects discover that they are being investigated, and threaten to take their business elsewhere and to initiate

defamation actions. The investigators return to their tried-and-true methods of locating and prioritising suspicious cases.

(7) Insider Detection

A government agency receives terse instructions from the government to get out ahead of the whistleblower menace, with Macbeth, Brutus, Iago, Judas Iscariot, Manning and Snowden invoked as examples of trusted insiders who turned. The agency increases the intrusiveness and frequency of employee vetting, and lowers the threshold at which positive vetting is undertaken. It applies big data analytics to a consolidated database comprising all internal communications, and all postings to social media gathered by a specialist external services corporation. To increase the pool of available information, it exercises powers to gain access to border movements, credit history, court records, law enforcement agencies' persons-of-interest lists, and financial tracking alerts.

The primary effect of these measures is to further reduce employee loyalty to the organisation. To the extent that productivity is measurable, it sags. The false positives arising from data analytics explode, because of the leap in negative sentiments expressed on internal networks and in social media, and in the vituperative langauge the postings contain. The false positives greatly increase the size of the haystack, making the presumed needles even harder to find. The poisonous atmosphere increases the opportunities for a vindictive insider to obfuscate their activities and even to find willing collaborators. Eventually cool heads prevail, by pointing out how few individuals ever actually leak information without authority. The wave of over-reaction slowly subsides, leaving a bruised and demotivated workforce with a bad taste in its mouth.

Acknowledgements

This paper has benefited from valuable feedback from Kasia Bail, Lyria Bennett Moses, Russell Clarke, David Vaile and Graham Greenleaf, and from comments by reviewers. Aspects were presented to a Symposium on Legal Implications of Big Data, run by Prof. Dan Svantesson of Bond University, in Sydney, on 12 December 2014.

References

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

Bollier D. (2010) 'The Promise and Peril of Big Data' The Aspen Institute, 2010, at http://www.ilmresource.com/collateral/analyst-reports/10334-ar-promise-peril-of-big-data.pdf

boyd D. & Crawford K. (2012) 'Critical Questions for Big Data' Information, Communication & Society, 15, 5 (June 2012) 662-679, DOI: 10.1080/1369118X.2012.678878, at http://www.tandfonline.com/doi/abs/10.1080/1369118X.2012.678878#.U_0X7kaLA4M

Clarke R. (1991) 'A Contingency Approach to the Software Generations' Database 22, 3 (Summer 1991) 23 - 34, PrePrint athttp://www.rogerclarke.com/SOS/SwareGenns.html

Clarke R. (1993) 'Profiling: A Hidden Challenge to the Regulation of Data Surveillance' Journal of Law and Information Science 4,2 (December 1993), PrePrint at http://www.rogerclarke.com/DV/PaperProfiling.html

Clarke R. (1994) 'The Digital Persona and Its Application to Data Surveillance' The Information Society 10,2 (June 1994), at http://www.rogerclarke.com/DV/DigPersona.html

Clarke R. (1995a) 'Computer Matching by Government Agencies: The Failure of Cost/Benefit Analysis as a Control Mechanism' Information Infrastructure & Policy 4,1 (March 1995), PrePrint at http://www.rogerclarke.com/DV/MatchCBA.html

Clarke R. (1995b) '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

Clarke R. (2014a) 'Promise Unfulfilled: The Digital Persona Concept, Two Decades Later' Information Technology & People 27, 2 (Jun 2014) 182-207, PrePrint at http://www.rogerclarke.com/ID/DP12.html

Clarke R. (2014b) 'Quality Factors in Big Data and Big Data Analytics' Working Paper, Xamax Consultancy Pty Ltd, September 2014, at http://www.rogerclarke.com/EC/BDQF.html

DHHS (2012) 'Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule' Department of Health & Human Services, November 2012, at http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html

Dreyfus H.L. (1992) 'What Computers Still Can't Do: A Critique of Artificial Reason' MIT Press, 1992

English L.P. (2006) 'To a High IQ! Information Content Quality: Assessing the Quality of the Information Product' IDQ Newsletter 2, 3, July 2006, at http://iaidq.org/publications/doc2/english-2006-07.shtml

Fayyad U., Piatetsky-Shapiro G. & Smyth P. (1996) 'From Data Mining to Knowledge Discovery in Databases ' AI Magazine 17, 3 (1996) 37-54, at http://aaaipress.org/ojs/index.php/aimagazine/article/download/1230/1131..

Guo P. (2013) 'Data Science Workflow: Overview and Challenges' ACM Blog, 30 October 2013, at http://cacm.acm.org/blogs/blog-cacm/169199-data-science-workflow-overview-and-challenges/fulltext

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P. & Witten I.H. (2009) 'The WEKA Data Mining Software: An Update' SIGKDD Explorations 11, 1 (2009), at http://www.sigkdd.org/sites/default/files/issues/11-1-2009-07/p2V11n1.pdf

van der Heijden K. (1996) 'Scenarios: The Art of Stategic Conversation' Wiley, 1996

Huh Y.U., Keller F.R., Redman T.C. & Watkins A.R. (1990) 'Data Quality' Information and Software Technology 32, 8 (1990) 559-565

Inmon B. (1992) 'Building the Data Warehouse' Wiley, 1992

Jacobs A. (2009) 'The Pathologies of Big Data' Communications of the ACM 52, 8 (August 2009) 36-44

Jagadish H.V., Gehrke J., Labrinidis A., Papakonstantinou Y., Patel J.M., Ramakrishnan R. & Shahabi C. (2014) 'Big data and its technical challenges' Communications of the ACM 57, 7 (July 2014) 86-94

Kimball R. (1996) 'The Data Warehouse Toolkit' Wiley, 1996

Leemhuis J.P. (1985) 'Using scenarios to develop strategies' Long Range Planning 18 (1985) 30-37

Leonelli S. (2014) 'What difference does quantity make? On the epistemology of Big Data in biology ' Big Data & Society (April–June 2014) 1–11, at http://bds.sagepub.com/content/1/1/2053951714534395.full.pdf+html

Marx G.T. & Reichman N. (1984) 'Routinising the Discovery of Secrets' Am. Behav. Scientist 27,4 (Mar/Apr 1984) 423-452

Mayer-Schonberger V. & Cukier K. (2013) 'Big Data: A Revolution That Will Transform How We Live, Work and Think' John Murray, 2013

Mobasheri F., Orren L.H. & Sjoshansi F.P. (1989) 'Scenario Planning at Southern California Edison' Interfaces (September/October 1989) 31-44

Ngai E.W.T., Xiu L. & Chau D.C.K. (2009) 'Application of data mining techniques in customer relationship management: A literature review and classification' Expert Systems with Applications, 36, 2 (2009) 2592-2602

OECD (1980) 'Guidelines on the Protection of Privacy and Transborder Flows of Personal Data' OECD, Paris, 1980

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

Piprani B. & Ernst D. (2008) 'A Model for Data Quality Assessment' Proc. OTM Workshops (5333) 2008, pp 750-759

Ratner B. (2003) 'Statistical Modeling and Analysis for Database Marketing: Effective Techniques for Mining Big Data' CRC Press, June 2003

Roszak T. (1986) 'The Cult of Information' Pantheon 1986

Rusbridge C., Burnhill P., Seamus R., Buneman P., Giaretta D., Lyon L. & Atkinson M. (2005) 'The Digital Curation Centre: A Vision for Digital Curation' Proc. Conf. From Local to Global: Data Interoperability--Challenges and Technologies, Sardinia, 2005, pp. 1-11, at http://eprints.erpanet.org/archive/00000082/01/DCC_Vision.pdf

Saha B. & Srivastava D. (2014) 'Data Quality: The other Face of Big Data', Proc. ICDE Conf., March-April 2014, pp. 1294 - 1297

Schwartz P. (1991) 'The Art of the Long View: Planning for the Future in an Uncertain World' Doubleday, 1991

UKICO (2012) 'Anonymisation: managing data protection risk: code of practice' Information Commissioners Office, November 2012, at http://ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Practical_application/anonymisation-codev2.pdf

Wack P. (1985) 'Scenarios: Uncharted Waters Ahead' Harv. Bus. Rev. 63, 5 (September-October 1985) 73-89

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