Quality Factors in Big Data and Big Data Analytics and Their Legal Implications

Roger Clarke

Xamax Consultancy, Canberra
Visiting Professor in Computer Science, ANU
and in Cyberspace Law & Policy, UNSW

Privacy Experts Symposium: Bigger Data in a Smaller World
UTS – 12 December 2014

http://www.rogerclarke.com/EC/BDLI {.html, .pdf}

Risk Management for Big Data Projects

Agenda

1. Big Data, Big Data Analytics
2. Data Quality
3. Decision Quality
4. Quality Factors and Big Data
5. Legal Risk Exposures

Vroom, Vroom
The 'Hype' Factor in Big Data

- Volume
- Velocity
- Variety

- Value

- Veracity

Laney 2001

Working Definitions

Big Data

- A single large data-collection
- A consolidation of data-collections:
  - Merger (Physical)
  - Interlinkage (Virtual)
    - Stored
    - Ephemeral

Big Data Analytics

Techniques for analysing 'Big Data'

The Third Element

- **Mythology**

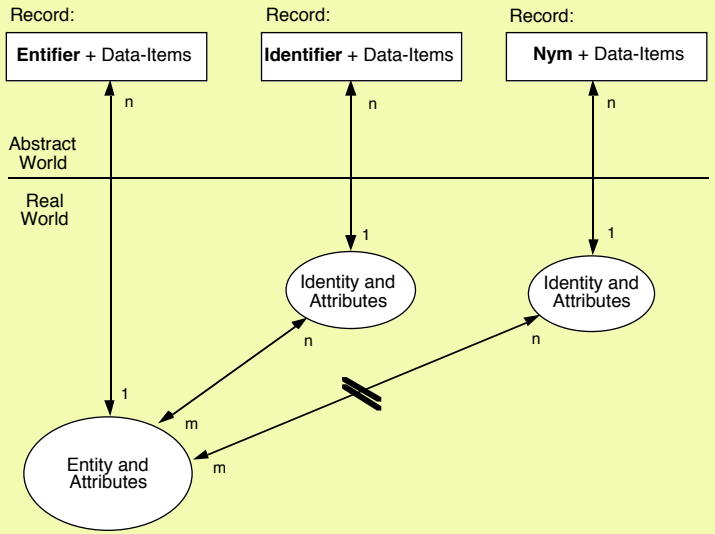
  "The widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy"

boyd & Crawford (2012, p.663)

Use Categories for Big Data Analytics

- **Hypothesis Testing**
  Evaluate whether propositions are supported by available data
  Propositions may be predictions from theory, heuristics, hunches
- **Population Inferencing**
  Draw inferences about the entire population or sub-populations, in particular correlations among particular attributes
- **Profile Construction**
  Identify key characteristics of a category, e.g. attributes and behaviours of 'drug mules' may exhibit statistical consistencies
- **Outlier Discovery**
  Find valuable needle in large haystack (flex-point, quantum shift)
- **Inferencing about Individuals**
  Inconsistent information or behaviour
  Patterns associated with a previously computed profile

The Identity Model



http://www.rogerclarke.com/ID/IdModel-1002.html

2. Data Quality Factors

- Accuracy
- Precision
- Timeliness
- Completeness

Accuracy

The **degree of correspondence** of a Data-Item with the real-world phenomenon that it is intended to represent

Measured by a confidence interval
e.g. 'accurate to within 1 degree Celsius'

Precision

The **level of detail** at which the data is captured
e.g. 'whole numbers of degrees Celsius'

Precision reflects the domain on which valid contents for that data-item are defined, e.g. Numeric fields may contain 'multiples of 5', 'integers', 'n digits after the decimal point', etc.

Date-of-Birth may be DDMMYYYY, DDMM, or YYYY, and may or may not include an indicator of the relevant time-zone

Timeliness

Temporal Applicability

- e.g. the period during which an income-figure was earned; the date after which a marriage, a qualification or a licence was applicable

Up-to-Dateness

- The absence of a material lag between a real-world occurrence and the recording of the corresponding data

Currency

- e.g. when the data-item was captured or last authenticated, or the period over which an average was computed. This is critical for volatile data-items, such as rainfall for the last 12 months, age, marital status, fitness for work

3. Decision Quality Factors

- Data Meaning
- Data Relevance
- Transparency
  - Process
  - Criteria

Data Meaning

- For each Data-Item, clear definition is needed of:
  - its meaning
  - the values that it can contain
  - the format in which the values are expressed
  - the meaning of each of those values
- Frequently, however:
  - meaning is not explicitly defined
  - the semantics are ambiguous
    e.g. 'spouse includes husband and wife' is silent on the questions of temporality, de facto relationships and same-gender relationships
  - meaning is subject to change, without recording of the changes and when they they took effect
  - valid content of the data-item is not defined

Data Relevance

- In Principle:

  Could the Data-Item make a difference **to the category of decision**?

  Do applicable law, policy and practice permit the Data-Item to make a difference?
- In Practice:

  Could the value that the Data-Item adopts in the particular context make a difference **to the particular decision** being made?

  Do applicable law, policy and practice permit the value of the Data-Item to make a difference?

Transparency

- Accountability requires clarity about the decision process and the decision criteria
- However:
  - Manual decisions are often poorly-documented
  - Algorithmic languages provide explicit or at least extractable process and criteria
  - Rule-based 'Expert Systems' software has implicit process and implicit criteria
  - 'Neural Network' software has implicit process and no discernible criteria

4. Quality Factors in Big Data Inferences

- Data Quality in each data collection:
  - Accuracy, Precision, Timeliness, Completeness
- Data Meaning Compatibilities
- Data Scrubbing Quality
- Data Consolidation Logic Quality
- Inferencing Process Quality
- Decision Process Quality:
  - Relevance, Meaning, Transparency

Data Scrubbing / Cleaning / Cleansing

- **Problems It Tries to Address**
  - Differing Definitions, Domains, Applicable Dates
  - Missing Data
  - Low and/or Degraded Data Quality
  - Failed Record-Matches due to the above
- **How It Works**
  - Internal Checks
  - Inter-Collection Checks
  - Algorithmic / Rule-Based Checks
  - Checks against Reference Data
- **Its Implications**
  - Better Quality and More Reliable Inferences
  - Worse Quality and Less Reliable Inferences

Factors Resulting in Bad Decisions

Assumption of Causality

- Inferencing Techniques seldom discover causality
- In complex circumstances, a constellation of factors are involved, none of which may be able to be meaningfully isolated as 'the cause', or 'the proximate cause', or even 'a primary cause'

Low-Grade Correlations

- Models with large numbers of intervening and confounding variables give low-grade correlations

Inaequate Models

- Key Variables and relationships may be missing from the model, resulting in misleading correlation
- There may not be a Model

Impacts of Bad Decisions based on Big Data

Resource Misallocation

- Negative Impacts on ROI
- Negative Impacts on public policy outcomes

Unjustified Discrimination

Breaches of Trust

- Re-Purposing of data
- Data Consolidation
- Data Disclosure

Reduced Security

- Multiple Copies
- Attacks on consolidated data-collections

Big Data Analytics – Population Focus

- Hypothesis Testing
- Population Inferencing
- Profile Construction

Anonymisation & Non-Reidentifiability are Vital

- Omission of specific rows and columns
- Generalisation / Suppression of particular values and value-ranges
- Data Falsification / 'Data Perturbation'
  - micro-aggregation, swapping, adding noise, randomisation

Slee 2011, DHHS 2012, **UKICO 2012**

Big Data Analytics – Individual Focus

- Outlier Discovery
- Inferencing about Individuals (e.g. Tax/Welfare Fraud Control)

Impacts on Individuals

- "A predermined model of infraction" "Probabilistic Cause cf. Probable Cause"
- A Non-Human Accuser, Poorly-Understood, Uncorrectable, Unchallengeable, and with Reversed Onus of Proof (i.e. Kafkaesque)
- Inconvenience, Harm borne by the Individual

S

6. Risk Exposure for Organisations

- Prosecution / Regulatory Civil Actions:
  - Against the Organisation
  - Against Directors
- Public Civil Actions, e.g. in Negligence
- Media Coverage / Harm to Reputation
- Public Disquiet / Complaints / Customer Retention / Brand-Value

The Effectiveness of Legal Controls

- Unknown Decisions
- Opaque Decision Processes and Criteria
- Lack of a Cause of Action
- Market and Institutional Power
- Lack of Effective Regulatory Agencies
- The Rapid Demise of Journalism
- Lack of Consumer/Citizen Power

Risk Management for Big Data Projects

Agenda

1. Big Data, Big Data Analytics
2. Data Quality
3. Decision Quality
4. Quality Factors and Big Data
5. Legal Risk Exposures

Quality Factors in Big Data and Big Data Analytics and Their Legal Implications

Roger Clarke

Xamax Consultancy, Canberra
Visiting Professor in Computer Science, ANU
and in Cyberspace Law & Policy, UNSW

Privacy Experts Symposium: Bigger Data in a Smaller World
UTS – 12 December 2014

http://www.rogerclarke.com/EC/BDLI {.html, .pdf}