Risk Management for Big Data Projects

Roger Clarke

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

8 December 2014

http://www.rogerclarke.com/EC/BDQF {.html, .ppt}

Risk Management for Big Data Projects

Agenda

- Big Data, Big Data Analytics
- Data
- Data Quality
- Decision Quality
- Quality Factors and Big Data
- Risk Exposure for Organisations
- RA / RM and DQM

How 'Big Data' Came To Be

Data Capture Developments

- 'Self-Service' Tx, Self-Exposure
- Web-Page, Mobile-Phone Usage
- Bar-Code Scanning
- Toll-Road Monitoring
- Payment and Ticketing Schemes
- Environmental Sensors

Storage Developments

- Disk-drives (Speed of Access, Storage Capacity)
- Solid-State Storage (Cost)

==>> **Economic Developments**
Data Retention cf. Data Destruction

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'

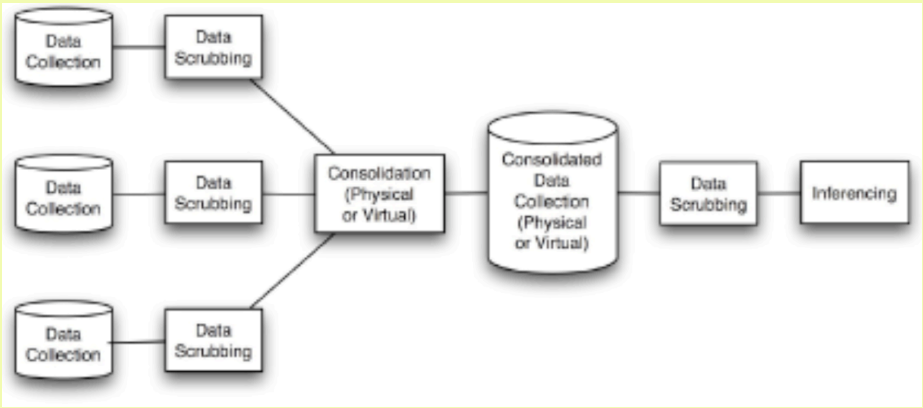
The Third Element

- **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)

Big Data – A Process View



Use Categories for Big Data Analytics

- Population Focus
  - Hypothesis Testing
  - Population Inferencing
  - Profile Construction
- Individual Focus
  - Outlier Discovery
  - Inferencing about Individuals

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

Risk Management for Big Data Projects

Agenda

- Big Data, Big Data Analytics
- **Data**
- Data Quality
- Decision Quality
- Quality Factors and Big Data
- Risk Exposure for Organisations
- RA / RM and DQM

Data

A symbol, sign or measure that is accessible to a person or an artefact

- Empirical Data represents or purports to represent a real-world phenomenon; Synthetic Data does not
- Quantitative Data gathered against Ordinal, Cardinal or Ratio Scales is suitable for various statistical techniques
- Qualitative Data gathered against a Nominal scale is subject to limited analytical processes
- Data is collected selectively and for a purpose
- Data may be compressed at or after the time of collection, through sampling, averaging and filtering of outliers

Identified Data

- **Entities** may exhibit one or more **Identities**
- Entities and Identities have **Attributes**
- An Attribute may be represented by a **Data-Item**
- Data-Items that represent Attributes of a particular (Id)Entity may be gathered into a **Record**
- A Record is associated with an (Id)Entity by means of an **(Id)Entifier** – one or more Data-Items that distinguish that (Id)Entity from others of the same category
- A **Digital Persona** is the impression of an (Id)Entity contained in a Record

Data Collections

- A **Record** contains Data-Items that represent Attributes of a particular (Id)Entity
- A **Data-File** contains a set of like Records
- A **Database** is a complex of Records in which data about each (Id)Entity is distributed, but in a sufficiently coordinated manner that a Digital Persona can be extracted whenever it is needed
- Conceptually, a **Data Collection** can be regarded as a two-dimensional table, with rows relating to (Id)Entities and columns containing Data-Items relating to Attributes of each of those (Id)Entities

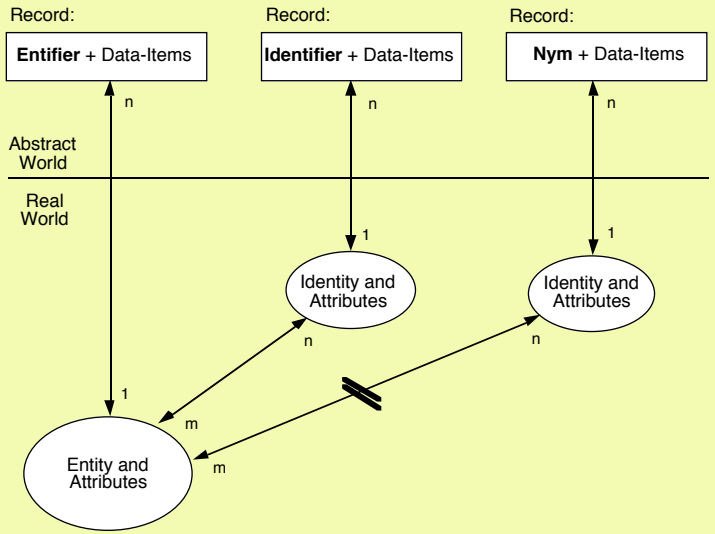
The Identifiability of (Id)Entities

- **Identified Data**: The (Id)Entity with which the data is associated is apparent from the Record alone
- **Identifiable Data**: The (Id)Entity with which the data is associated is apparent from the Record together with other Data from the context of use
- **Pseudonymous Data**: The association between the Record and the (Id)Entity is subject to technical, organisational and legal protections
- **Anonymous Data**: No association can be achieved between the Record and the (Id)Entity

---

- **Re-Identification**: The association of nominally Anonymous Data with am (Id)Entity

The Identity Model



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

Beyond Data

- **Information** is Data that has value.
  The value of Data depends upon Context.
- The most common such Context is a **Decision**, i.e. selection among a number of alternatives
- **Knowledge** is the matrix of impressions within which a human situates new Information
- **Wisdom** is the capacity to exercise judgement by selecting and applying Decision Criteria to Knowledge combined with new Information

Risk Management for Big Data Projects

Agenda

- Big Data, Big Data Analytics
- Data
- **Data Quality**
- Decision Quality
- Quality Factors and Big Data
- Risk Exposure for Organisations
- RA / RM and DQM

Key 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

Completeness

- The availability of sufficient contextual information that the data is not liable to be misinterpreted
- The notions of context, sufficiency and interpretation are highly situation-dependent

Data Quality Processes

- Data Integrity tends to deteriorate as a result of:
  - Efflux of time
  - Degradation of the storage medium
  - Processing
  - Loss of associated (meta)data
    e.g. the data's provenance, the scale against which it was measured, the valid domain-values at the time it was recorded, the context within which it needs to be interpreted
- Measures are necessary to sustain Data Integrity

Risk Management for Big Data Projects

Agenda

- Big Data, Big Data Analytics
- Data
- Data Quality
- **Decision Quality**
- Quality Factors and Big Data
- Risk Exposure for Organisations
- RA / RM and DQM

Key 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

Risk Management for Big Data Projects

Agenda

- Big Data, Big Data Analytics
- Data
- Data Quality
- Decision Quality
- **Quality Factors and Big Data**
- Risk Exposure for Organisations
- RA / RM and DQM

Quality Factors in Big Data Inferences

- Data Quality in each data collection:
  - Accuracy, Precision, Timeliness, Completeness
- Data Meaning Compatibilities
- Data Scrubbing Quality
- ...

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

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

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

Risk Management for Big Data Projects

Agenda

- Big Data, Big Data Analytics
- Data
- Data Quality
- Decision Quality
- Quality Factors and Big Data
- **Risk Exposure for Organisations**
- RA / RM and DQM

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

- Big Data, Big Data Analytics
- Data
- Data Quality
- Decision Quality
- Quality Factors and Big Data
- Risk Exposure for Organisations
- **RA / RM and DQM**

Risk Assessment / Risk Management

- ISO 31000 – Risk Management Process Standards
- ISO 27000 – Information Security Process Standards

- Generic Strategies:
  - Avoidance
  - Exploitation
  - Removal
  - Amelioration
  - Sharing
  - Acceptance

Data Quality Assurance

- ISO 8000 – Data Quality Process Standard
- "But ISO 8000 simply requires that the data elements and coded values be explicitly defined. ... ISO 8000 is a method that seeks to keep the metadata and the data in sync"

Risk Management for Big Data Projects

1. **Frameworks**

- Incorporate Big Data Programs within the organisation's RA/RM framework
- Incorporate Big Data Programs within the organisation's DQM framework
- Ensure that the organisation's DQM framework addresses intrinsic and contextual data quality factors

Risk Management for Big Data Projects

2. **Data Consolidation**

- Ensure that data collections are not consolidated unless:
  - they satisfy threshold data quality tests
  - their purposes, their quality and the meanings of relevant data-items are compatible
  - relevant legal, moral and public policy constraints are respected

Risk Management for Big Data Projects

3. **Effective Anonymisation**

- Ensure that, where sensitive data is involved, particularly personal data, anonymisation techniques are applied, and the data that is submitted to analysis is not re-identifiable

Risk Management for Big Data Projects

4. Data Scrubbing

- Ensure that, where data scrubbing operations are undertaken:
  - they are undertaken within the context of the organisation's data quality assurance framework
  - they involve external reference-points, and are not limited to internal consistency checks
  - their accuracy and effectiveness are audited
  - the results are not used for decision-making unless the audits demonstrate that the results satisfy threshold data quality tests

Risk Management for Big Data Projects

5. Decision-Making

- Ensure that inferencing mechanisms are not relied upon to make decisions, unless their applicability to the data in question has been subjected to independent review and they have been found to be suitable
- Ensure that, when 'big data' is applied to decision-making:
  - the criteria of relevance, meaning, and transparency of decision mechanisms are all satisfied
  - the results are audited, including by testing against known instances
  - the outcomes are subjected to post-implementation assessment, including through transparency arrangements and complaints mechanisms

Risk Management for Big Data Projects

Agenda

- Big Data, Big Data Analytics
- Data
- Data Quality
- Decision Quality
- Quality Factors and Big Data
- Risk Exposure for Organisations
- RA / RM and DQM

Risk Management for Big Data Projects

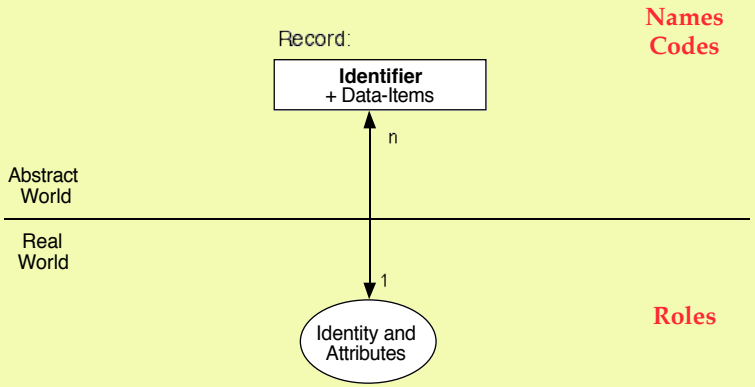
Roger Clarke

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

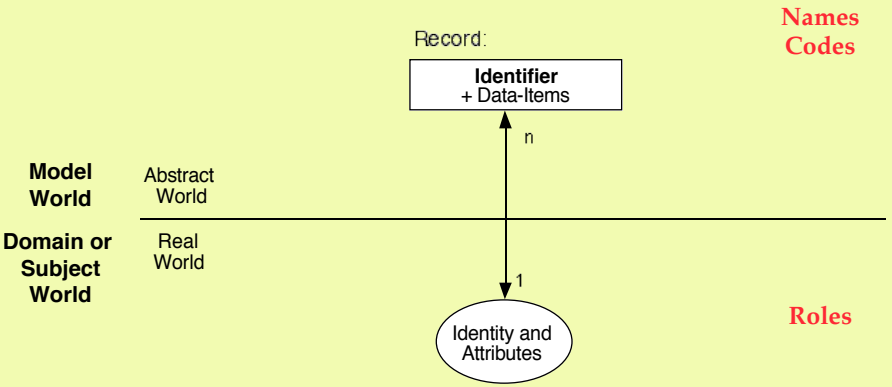
8 December 2014

http://www.rogerclarke.com/EC/BDQF {.html, .ppt}

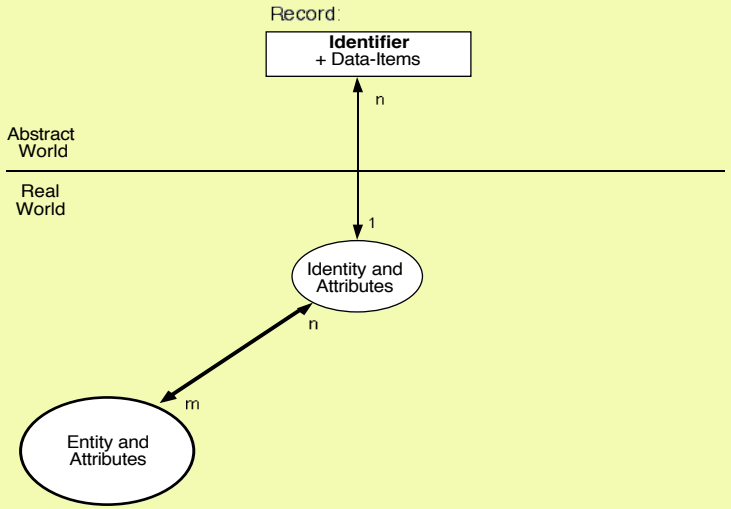
Identity and Identifier



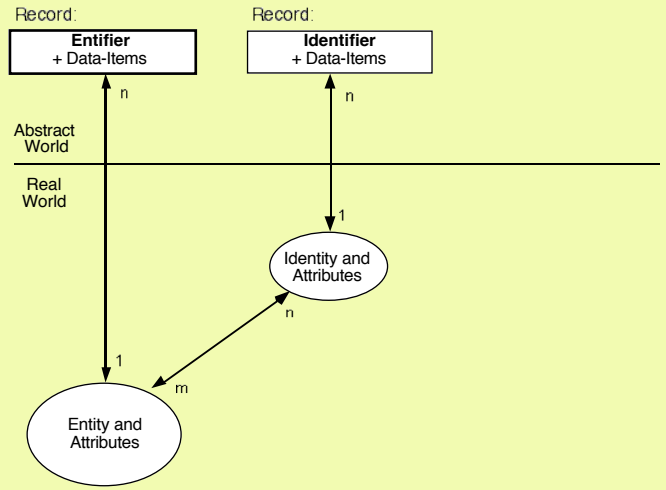
Identity and Identifier



The Entity/ies underlying an Identity



Entity and Entifier



Nymity