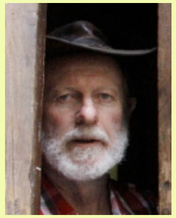
Thanks to Chris Slane, NZ
http://www.slane.co.nz/

Towards Responsible Data Analytics: A Process Approach



Roger Clarke
Xamax Consultancy Pty Ltd,
Canberra, RSCS ANU, UNSW Law

Kerry Taylor
RSCS ANU, Canberra

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

Bled eConference – 19 June 2018

Big Data Analytics
Vroom, Vroom, Vroom

- Volume
- Velocity
- Variety

- Value

- **Veracity**
- **Validity**
- **Visibility**

Laney 2001, Livingston 2013

Use Categories for Big Data Analytics

- **Population Focus**
  - Hypothesis Testing
  - Population Inferencing
  - Construction of Profiles
- **Individual Focus**
  - Application of Profiles
  - Discovery of Anomalies
  - Outlier Discovery
  - Discovery of Outliers

Data Quality Factors

Assessable at time of collection

D1 – Syntactic Validity

D2 – Appropriate (Id)entity Association

D3 – Appropriate Attribute Association

D4 – Appropriate Attribute Signification

D5 – Accuracy

D6 – Precision

D7 – Temporal Applicability

Information Quality Factors

Assessable only at time of use

I1 – Theoretical Relevance

I2 – Practical Relevance

I3 – Currency

I4 – Completeness

I5 – Controls

I6 – Auditability

Data Scrubbing (Wrangling / Cleaning / Cleansing)

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

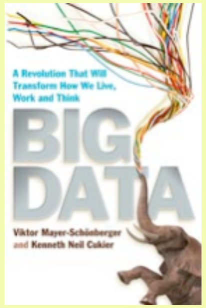
Key Decision Quality Factors

- Appropriateness of the Inferencing Technique
- Data Meaning
- Data Relevance
- Transparency
  - Process
  - Criteria

"[F]aced with massive data,
[the old] approach to science
-- hypothesize, model, test -- is ... obsolete.

"Petabytes allow us to say:
'Correlation is enough' "

Anderson C. (**2008**) 'The End of Theory:
The Data Deluge Makes the Scientific Method Obsolete'
Wired Magazine 16:07, 23 June 2008



"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 ..."

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

Transparency

- **Accountability** depends on clarity
  about the Decision Process
  and the Decision Criteria
- **In practice, Transparency is highly variable**:
  - **Manual decisions** – Often poorly-documented
  - **Algorithmic languages**
    Process & criteria explicit (or at least extractable)
  - **Rule-based 'Expert Systems' software**
    Process implicit; Criteria implicit
  - **'Neural Network' software**
    Process implicit; Criteria not discernible

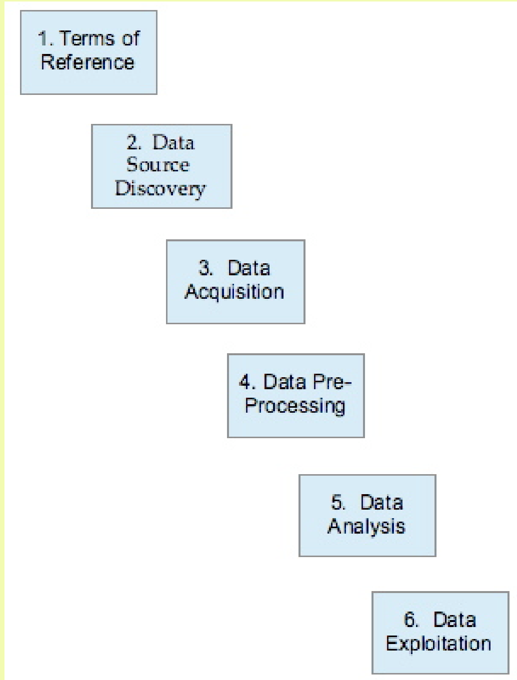
The Problem

- New techniques are escaping laboratories
  with limited maturity and few controls
- Over-enthusiasm by spruikers
  is about to collide with business risk
- There will be negative impacts on business
  and on people affected by decisions
- **Business needs guidance on how to cope**

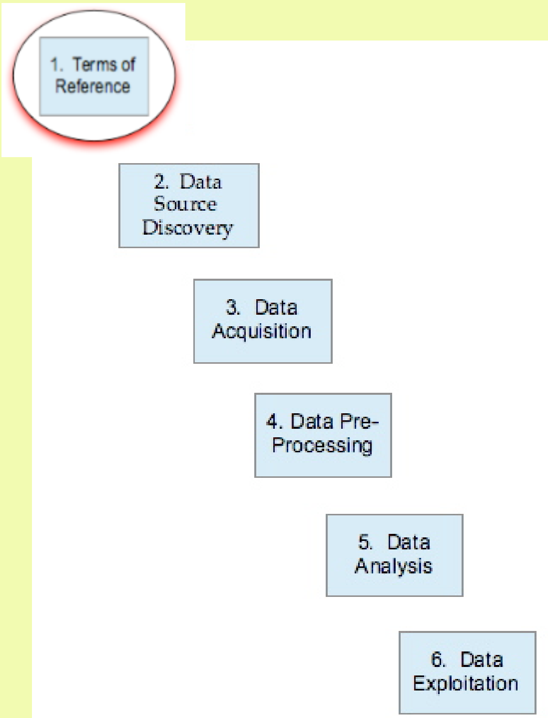
The Project Method
A Design Science Approach

- Identify conventional business processes for applying data analytics
- Apply risk assessment, risk management
- Identify shortfalls
- Propose an adapted business process
- Illustrate through a case study

A Conventional Business Process for Big Data Analytics Projects



A Conventional Business Process for Big Data Analytics Projects



Risks & Responsibilities

- Data Quality at time of creation
- Information Quality at time of use
- Data Scrubbing impacts
- Data Merger errors
- Analytical Technique applicability
- Inferencing Quality
- **Decision Rationale Transparency == >> Accountability**
- **Usee Impacts**
- Organisational Impacts

Risk Assessment

For Organisations

- ISO 31000/10 – Risk Mngt Process Standards
- ISO 27005 etc. – Information Security Risk Mngt
- NIST SP 800-30 – Risk Mngt Guide for IT Systems
- ISO 8000 – Data Quality Process Standard
- ISACA COBIT, ITIL, PRINCE2, ...

Risk Assessment

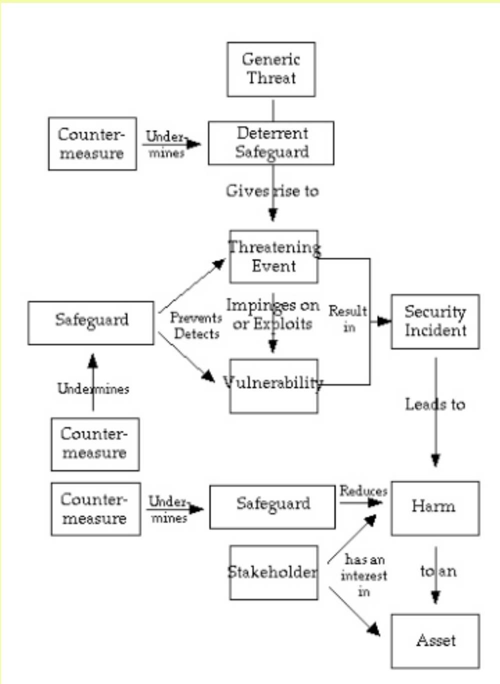
For Organisations

- ISO 31000/10 – Risk Mngt Process Standards
- ISO 27005 etc. – Information Security Risk Mngt
- NIST SP 800-30 – Risk Mngt Guide for IT Systems
- ISO 8000 – Data Quality Process Standard
- ISACA COBIT, ITIL, PRINCE2, ...

For 'Usees'

- Technology Assessment (TA)
- Privacy Impact Assessment (PIA)

The Conventional Model Underlying Risk Assessment



Generic Risk Management Strategies

Proactive Strategies

- Avoidance
- Deterrence
- Prevention
  e.g. Redundancy

Reactive Strategies

- Detection
- Isolation / Mitigation
- Recovery
- Transference
  e.g. Insurance

Non-Reactive Strategies

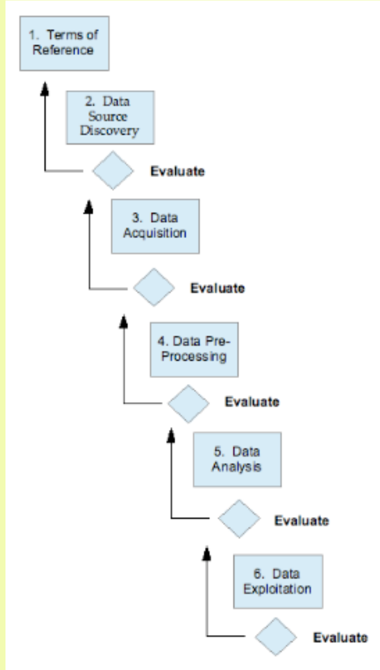
- Tolerance / Acceptance
  e.g. Self-Insurance
- Abandonment
- Dignified Demise / Graceful Degradation
- Abandonment / Grace<u>less</u> Degradation

Conventional Business Process for Data Analytics

MISSING ELEMENTS

1. A preliminary, planning Phase
2. Evaluation steps after each Phase
3. Criteria for deciding whether the project needs to be looped back to an earlier Phase

An Adapted Business Process



'Guidelines for Responsible Application of Data Analytics'

1. General

DO's:
Governance, Expertise, Compliance

2. Data Acquisition

DO's:
The Problem Domain, The Data Sources, Data Merger, Data Scrubbing, Identity Protection, Data Security

DON'Ts:
Identifier Compatibility, Content Compatibility

3. Data Analysis

DO's:
Expertise, The Nature of the Tools, The Nature of the Data Processed by the Tools, The Suitability of the Tools and the Data

DON'Ts:
Inappropriate Data, Humanly-Understandable Rationale

4. Use of the Inferences

DO's:
The Impacts, Evaluation, Reality Testing, Safeguards, Proportionality, Contestability, Breathing Space, Post-Implementation Review

DON'Ts:
Humanly-Understandable Rationale, Precipitate Actions, Automated Decision-Making

Computer Law & Security Review 34, 3 (May-Jun 2018)
https://doi.org/10.1016/j.clsr.2017.11.002
PrePrint at http://www.rogerclarke.com/EC/GDA.html

2. Data Acquisition

2.1 The Problem Domain

Understand the real-world systems about which inferences are drawn, to which data analytics are applied

2.2 The Data Sources

Understand each source of data, including:

a. the data's provenance
b. the purposes for which the data was created
c. the meaning of each data-item at time of creation
d. the data quality at the time of creation
e. data quality and information quality at time of use

4. Uses of the Inferences

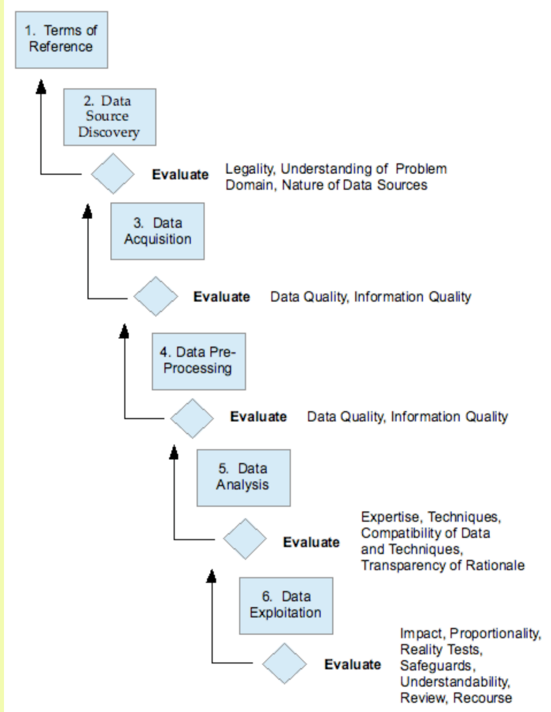
4.9 Humanly-Understandable Rationale

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

4.11 Automated Decision-Making

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

An Adapted Business Process ... Articulated



Instantiations

- For each Use Category (as per Slide 5)
- Embeddedness in a corporate framework (e.g. standalone project, or constrained by corporate policies and practices, standards)
- Ground-breaking vs. novel project
- Degree of team-expertise and -experience

Demonstration via Case Study

Centrelink's Online Compliance Intervention (OCI) System

- Implicit assumption that declared annual income could be divided by 26 to infer income for each fortnight of that year
- Abandonment of checks with employers, transferring those costs to the recipients
- Automation of debt-raising
- Automated referral to debt collectors
- Leap in case-load by more than 30-fold, hence most complaints were ignored

http://www.rogerclarke.com/DV/CRD17.html

Conclusions

- Conventional business processes for data analytics lack three important features
- On the basis of established theories, plus prior research into risk assessment of data analytics projects, an adapted business process model was proposed, to make good those deficiencies
- A recent case was considered in the light of the adapted model

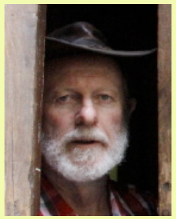
Implications for Practice

- Data analytics projects need to be intercepted before they are applied
- Company directors and executives must manage direct organisational risks
- Risks to the public may be publicised and may snowball, resulting in reputational, compliance and diversion risks
- QA, RA and RM need to be applied, but also IA and IM

Implications for Research

- Instantiation is needed
- Articulation may be needed
- Case studies are needed of applications of the adapted business process
- Commercial, strategic, ethical, legal and political factors give rise to barriers to such research
- Quality and risk factors should be considered far earlier in the technology life-cycle

Towards Responsible Data Analytics: A Process Approach



Roger Clarke
Xamax Consultancy Pty Ltd, Canberra, RSCS ANU, UNSW Law

Kerry Taylor
RSCS ANU, Canberra

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

Bled eConference – 19 June 2018