Risk Management for Big Data Projects

Roger Clarke

Xamax Consultancy, Canberra

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

March 2015

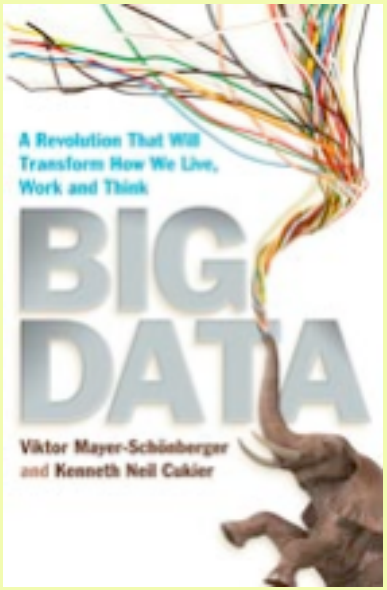
http://www.rogerclarke.com/EC/BDRM {.html, .ppt}

"[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

XAMAX Consultancy Pty Ltd



"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

XAMAX Consultancy Pty Ltd

Risk Management for Big Data Projects

Roger Clarke

Xamax Consultancy, Canberra

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

March 2015

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

How 'Big Data' Came To Be

Data Capture Developments

- Bar-Code Scanning
- Toll-Road Monitoring
- Payment, Ticketing
- eComms, Web-Access
- Social Media
- 'Wellness Data'
- Environmental Sensors

Storage Developments

- Disk (Speed, Capacity)
- Solid-State (Cost)

Economic Developments

- Data Retention now much cheaper than Data Destruction

Government Open Data Initiatives

- data.gov.au — total 5,298 datasets
- data.nsw.gov.au — incl. 3,843 spatial datasets at LPI

XAMAX Consultancy Pty Ltd

Vroom, Vroom
The 'Hype' Factor in Big Data

- **Volume**
- **Velocity**
- **Variety**

- Value

XAMAX Consultancy Pty Ltd

Vroom, Vroom
The 'Hype' Factor in Big Data

- Volume
- Velocity
- Variety

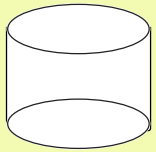
- Value

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

XAMAX Consultancy Pty Ltd

Laney 2001, Livingston 2013

Working Definitions



Big Data

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

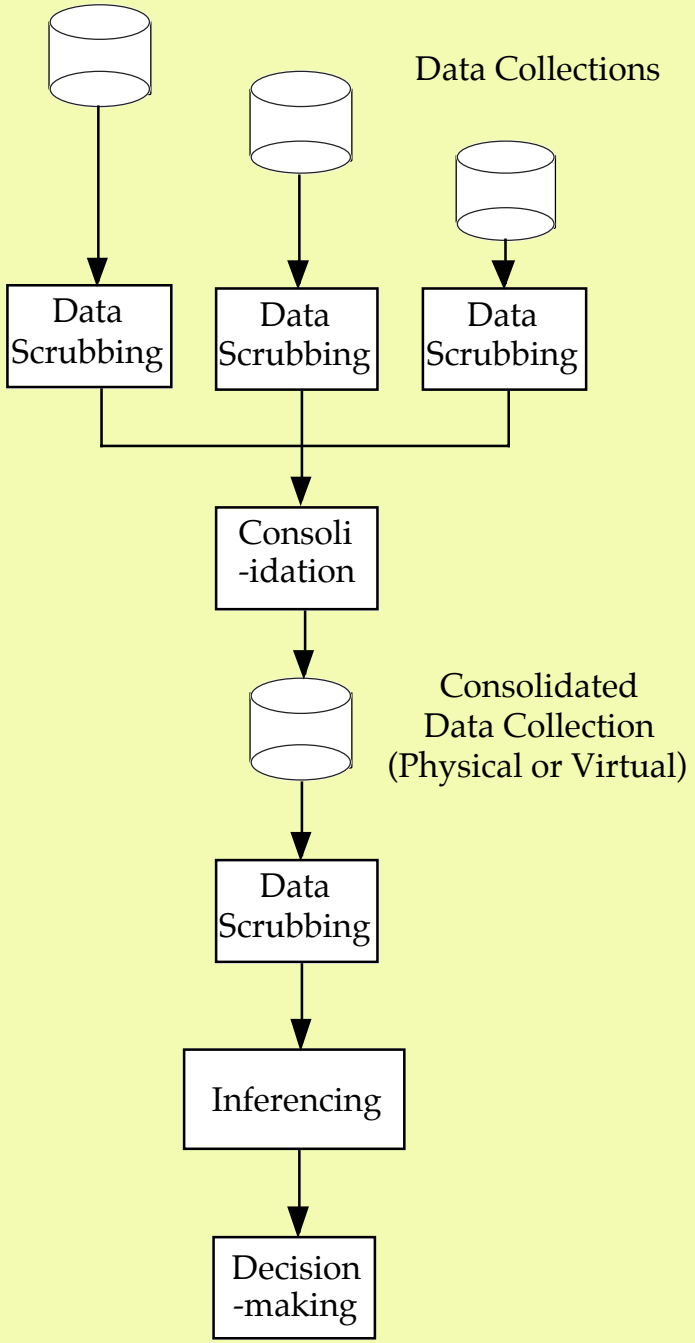
Big Data Analytics

Techniques for analysing 'Big Data'

XAMAX Consultancy Pty Ltd

Big Data & Big Data Analytics

Process View



Working Definitions
The Third Element

Mythology

”[There is a] 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)

Working Definitions
The Third Element

Mythology

”[There is a] 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”

e.g. the ‘Beers and Diapers’ Correlation

‘If it happened, it didn’t happen like that’

XAMAX Consultancy Pty Ltd

boyd & Crawford (2012, p.663)
http://www.dssresources.com/newsletters/66.php

Data Categories for Big Data Analytics

- Geo-Physical Data
- Geo-Spatial Data

...

- Personal Data acquired by Govt Agencies

- Social Media Content

- Biochemical Data
- Epidemiological Data

...

- Pharmaceutical and Medical Services Data
- Personal Health Care Data

- Personal 'Wellness Data'

XAMAX Consultancy Pty Ltd

Use Categories for Big Data Analytics

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

XAMAX Consultancy Pty Ltd

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

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

Use Categories for Big Data Analytics

P Hypothesis Testing

Evaluate whether propositions are supported by available data
Propositions may be predictions from theory, heuristics, hunches

P Population Inferencing

Draw inferences about the entire population or sub-populations, in particular correlations among particular attributes

P Profile Construction

Identify key characteristics of a category, e.g. attributes and behaviours of 'drug mules' may exhibit statistical consistencies

---

I Inferencing about Individuals

Inconsistent information or behaviour
Patterns associated with a previously computed profile

I Outlier Discovery

Find valuable needle in large haystack (flex-point, quantum shift)

XAMAX Consultancy Pty Ltd

Risk Management for Big Data Projects

Agenda

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

XAMAX Consultancy Pty Ltd

Data

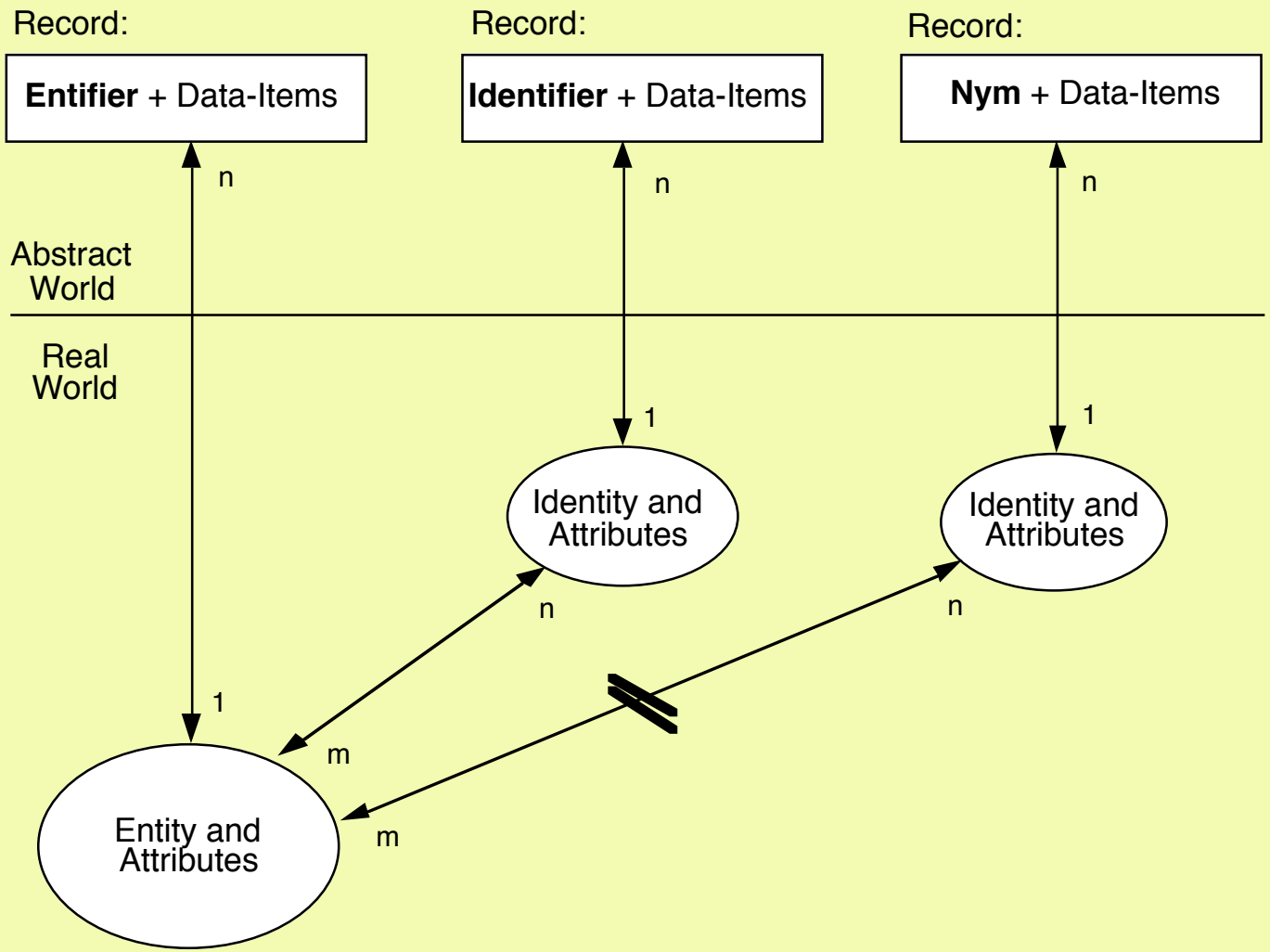
A symbol, sign or measure that is accessible to a person or an artefact

- **Empirical Data** represents 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 Collection is selective and for a purpose**
- **Data may be compressed** at or after the time of collection, e.g. through sampling, averaging and filtering of outliers

The Association of Data with (Id)Entities



XAMAX Consultancy Pty Ltd

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. a selection among a number of alternatives

- **Knowledge** is the matrix of impressions within which a sentient being situates new Information

- **Wisdom** is the capacity to exercise judgement by selecting and applying Decision Criteria to Knowledge combined with new Information

XAMAX Consultancy Pty Ltd

Risk Management for Big Data Projects

Agenda

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

XAMAX Consultancy Pty Ltd

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. '± 1 degree Celsius'

Precision

The **level of detail** at which the data is captured

Reflects the domain on which the data-item is defined

e.g. 'whole numbers of degrees Celsius'

e.g. 'multiples of 5', 'integers', 'n digits after the decimal point'

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

Timeliness

Up-to-Dateness

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

Currency or Period of Applicability

- The date after which a marriage or a licence is applicable
- When the data-item was captured or last authenticated
- The period during which an income-figure was earned
- The period over which an average was computed

  Particularly 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 Falls Over Time

Data Integrity deteriorates, as a result of:

- Storage Medium Degradation
- Loss of Context
- Changes in Context
- Changes in Business Processes
- Loss of Associated (Meta)Data, ...

XAMAX Consultancy Pty Ltd

George Brandis explains metadata
#auspol

XAMAX Consultancy Pty Ltd

http://www.rogerclarke.com/DV/DRPS.html#CP
https://www.privacy.org.au/Papers/
PJCIS-DataRet-Supp-150131.pdf

Data Quality Falls Over Time

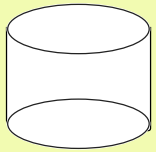
Data Integrity deteriorates, as a result of:

- Storage Medium Degradation
- Loss of Context
- Change of Context
- Changes in Business Processes
- Loss of Associated (Meta)Data, e.g.
  - Provenance of the data
  - The Scale against which it was measured
  - Valid Domain-Values when it was recorded
  - Contextual Information to enable interpretation

Measures are necessary to sustain Data Integrity

XAMAX Consultancy Pty Ltd

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'

XAMAX Consultancy Pty Ltd

Key Decision Quality Factors

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

Appropriateness of the Inferencing Technique

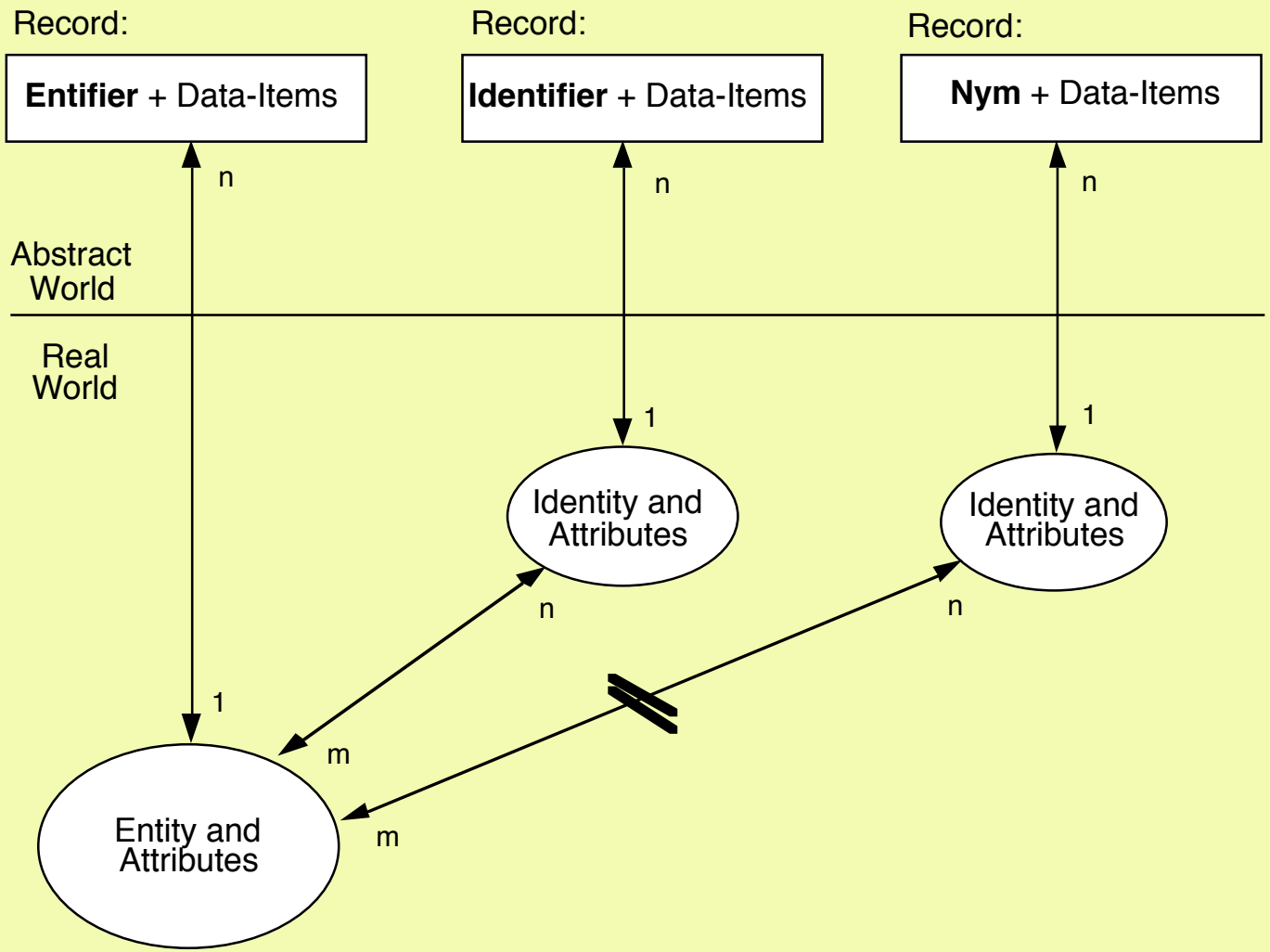
- …

XAMAX Consultancy Pty Ltd

What Does 'Data Meaning' Mean?

- **Syntactics**
  - The relationships among data-items
  - The values that a data-item may contain
  - The formats in which the values are expressed
- **Semantics**
  - The particular real-world attribute that the data-item is intended to represent
  - The particular state of the real-world attribute that the content of the data-item is intended to represent

The Identity Model



XAMAX Consultancy Pty Ltd

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

What Does 'Data Meaning' Mean?

- **Syntactics**
  - The relationships among data-items
  - The values that a data-item may contain
  - The formats in which the values are expressed
- **Semantics**
  - The particular real-world attribute that the data-item is intended to represent
  - The particular state of the real-world attribute that the content of the data-item is intended to represent
- **Pragmatics**
  - The inferences that people may draw from particular data-items and the particular values they contain

Key Decision Quality Factors

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

Data Relevance

- **<u>The Category of Decision</u>**

  Could the Data-Item make a difference?

  & Do applicable law, policy and practice allow the Data-Item to make a difference?

- **<u>The Particular Decision</u>**

  Could the value that the Data-Item adopts in the particular instance make a difference?

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

XAMAX Consultancy Pty Ltd

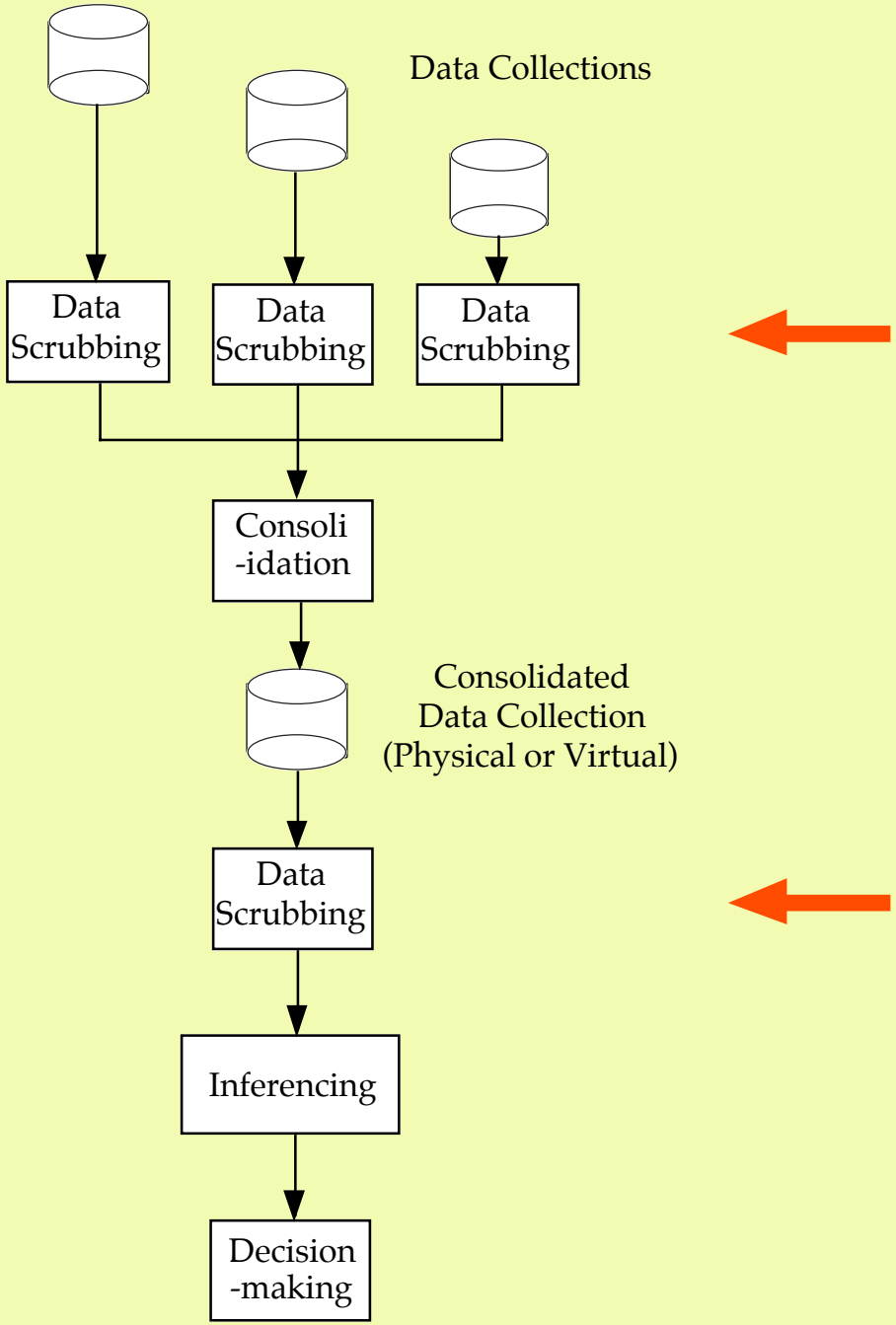
Transparency

- Accountability requires 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

Big Data & Big Data Analytics

Process View



Data Scrubbing / Cleaning / Cleansing

- **Problems It Tries to Address**
  - Missing Data
  - Low and/or Degraded Data Quality
  - Failed and Spurious Record-Matches
  - Differing Definitions, Domains, Applicable Dates

XAMAX Consultancy Pty Ltd

Data Scrubbing / Cleaning / Cleansing

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

XAMAX Consultancy Pty Ltd

Data Scrubbing / Cleaning / Cleansing

- **Problems It Tries to Address**
  - Missing Data
  - Low and/or Degraded Data Quality
  - Failed and Spurious Record-Matches
  - Differing Definitions, Domains, Applicable Dates
- **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**

XAMAX Consultancy Pty Ltd

Risk Management for Big Data Projects

Agenda

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

XAMAX Consultancy Pty Ltd

Summary: Quality Factors in Big Data Inferences

- Data Quality in each data collection:
  - Accuracy, Precision, Timeliness, Completeness
- Data Meaning Ambiguities
- Data Scrubbing Quality
- Data Consolidation Logic Quality
  - esp. Data Compatibility Issues
- Inferencing Process Quality
- Decision Process Quality:
  - Relevance, Meaning, Transparency

Additional Factors Resulting in Bad Decisions

Low-Grade Correlations

- Complex realities, high diversity, complex questions

Assumption of Causality

- Inferencing Techniques seldom discover causality
- In complex realities, there is no single 'cause', or 'primary cause', or even 'proximate cause'

Inadequate Models

- Important Independent, Moderating, Confounding Variables may be missing from the model
- There may not be a Model
- *And Big Data Devotees recommend you not have one!??*

Organisational Risks – Internal

Security Considerations

- More Copies lie around
- Consolidation creates Honeypots
- Honeypots attract Attackers
- Attacks succeed

Resource Misallocation

- Negative impacts on ROI

XAMAX Consultancy Pty Ltd

Personal Risks

- Outlier Discovery
- **Inferencing about Individuals**
  - **Targetted Advertising**

"Darling, I thought you'd stopped gambling.

"So how come so many gambling ads
pop up in your browser window?"

Personal Risks

I Outlier Discovery

I Inferencing about Individuals

- Targetted Advertising
- **Tax/Welfare Fraud Control**

- "A predermined model of infraction"
  "Probabilistic Cause cf. Probable Cause"
- Non-Human Accuser, Unclear Accusation, Reversed Onus of Proof, Unchallengeable
- Inconvenience, Harm borne by the Individual

XAMAX Consultancy Pty Ltd

Personal Risks

- P Hypothesis Testing
- P Population Inferencing
- P Profile Construction

Anonymisation, Non-Reidentifiability is vital

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

Slee 2011, DHHS 2012, **UKICO 2012**
https://www.oic.qld.gov.au/guidelines/for-government/guidelines-privacy-principles/applying-the-privacy-principles/dataset-publication-and-risk-assessment

XAMAX Consultancy Pty Ltd

Personal Risks
with Implications for Organisations

Breaches of Trust

- Data Re-Purposing
- Data Consolidation
- Data Disclosure

Discrimination

'Unfair' Discrimination

XAMAX Consultancy Pty Ltd

Organisational Risks – External

- **Public Civil Actions**, e.g. in Negligence
- **Prosecution** / Regulatory Civil Actions:
  - Against the Organisation
  - Against Directors

XAMAX Consultancy Pty Ltd

Organisational Risks – External

- **Public Civil Actions**, e.g. in Negligence
- **Prosecution** / Regulatory Civil Actions:
  - Against the Organisation
  - Against Directors

- **Public Disquiet** / Complaints / Customer Retention / Brand-Value
- **Media Coverage** / Harm to Reputation
- **Active Obfuscation and Falsification**

XAMAX Consultancy Pty Ltd

Risk Management for Big Data Projects

Agenda

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

XAMAX Consultancy Pty Ltd

Risk Management in SFIA

Business strategy and planning	Research	RSCH	3 4 5 6
	Innovation	INOV	5 6
	Business process improvement	BPRE	5 6 7
	Enterprise & business architecture development	STPL	5 6 7
	Business risk management	BURM	4 5 6 7
	Sustainability strategy	SUST	5 6

- A key element of Bus Strategy and Planning
- Strongly aligned with SFIA Level 5
- "under broad direction ...
  "often self-initiated ...
  "fully accountable for meeting objectives ..."

XAMAX Consultancy Pty Ltd

https://www.acs.org.au/sfia-certification/mysfia/about-sfia

Risk Assessment / Risk Management

- ISO 31000/10 – Risk Mngt Process Standards
- **ISO 27005** etc. – Information Security Risk Mngt
- **NIST SP 800-30** – Risk Mngt Guide for IT Systems
- ISACA, COBIT, etc.

Generic Strategies:

- Exploitation
- Avoidance
- Removal
- Amelioration
- Sharing
- Acceptance

Risk Management

- **Assess Risk**
  - Objectives and Constraints
  - Stakeholders, Assets, Values, Harm
  - Threats, Vulnerabilities, Combinations
  - Existing Safeguards
  - Residual Risks
  - Priorities
- **Design More / Better Safeguards**
- **Implement**
- **Review and Revise**

XAMAX Consultancy Pty Ltd

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"
  i.e. mostly limited to syntactic aspects

Risk Management for Big Data Projects

1. Frameworks
2. Data Consolidation
3. Effective Anonymisation
4. Data Scrubbing
5. Decision-Making

XAMAX Consultancy Pty Ltd

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**
- If you haven't got one, get one

Risk Management for Big Data Projects

2. **Data Consolidation**

Don't consolidate data collections **unless**:

- they satisfy **threshold data quality tests**
- their purposes, their quality and the meanings of relevant data-items satisfy **threshold compatibility tests**
- relevant legal, moral and public policy constraints are respected

XAMAX Consultancy Pty Ltd

Risk Management for Big Data Projects

3. **Effective Anonymisation**

- Where sensitive data is involved, particularly personal data, apply anonymisation techniques, and ensure the data is not re-identifiable

XAMAX Consultancy Pty Ltd

Risk Management for Big Data Projects

4. **Data Scrubbing**

- Undertake cleansing within the context of **the organisation's data quality framework**
- Use external reference-points, not just internal consistency checks
- Audit accuracy and effectiveness
- **Don't** use the results for decision-making **unless** the audits demonstrate that **the results satisfy threshold tests**

Risk Management for Big Data Projects

5. **Decision-Making**

- **Don't rely on inferencing mechanisms, unless** their applicability to the data has been independently reviewed and found to be suitable
- **Check relevance, meaning, and transparency**
- **Audit the results**, testing against known instances
- **Conduct outcome assessment** through transparency arrangements, complaints

Risk Management for Big Data Projects

Roger Clarke

Xamax Consultancy, Canberra

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

March 2015

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