Big Data
Quality Assurance

Roger Clarke

Xamax Consultancy, Canberra

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

Redefining R&D Needs for Australian Cyber Security
Australian Centre for Cyber Security (ACCS)
16 November 2015

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

Vroom, Vroom
The 'Hype' Factor in Big Data

- Volume
- Velocity
- Variety

Vroom, Vroom
The 'Hype' Factor in Big Data

- Volume
- Velocity
- Variety
- Value
- **Veracity**

XAMAX Consultancy Pty Ltd

Laney 2001, Livingston 2013

Opportunities in the Security Area

- Network traffic
- Open Source Intelligence
- Social media postings – public
- Social media postings – organisation-internal

...

- Streams of data from eObjects
  (the datafication of things and people)

XAMAX Consultancy Pty Ltd

Use Categories for Big Data Analytics

- **Population Focus**
  - Hypothesis Testing
  - Population Inferencing
  - Profile Construction
- **Individual Focus**
  - Outlier Discovery
  - Inferencing about Individuals
    - Inconsistencies
    - Non/-conformance with a profile

XAMAX Consultancy Pty Ltd

Data

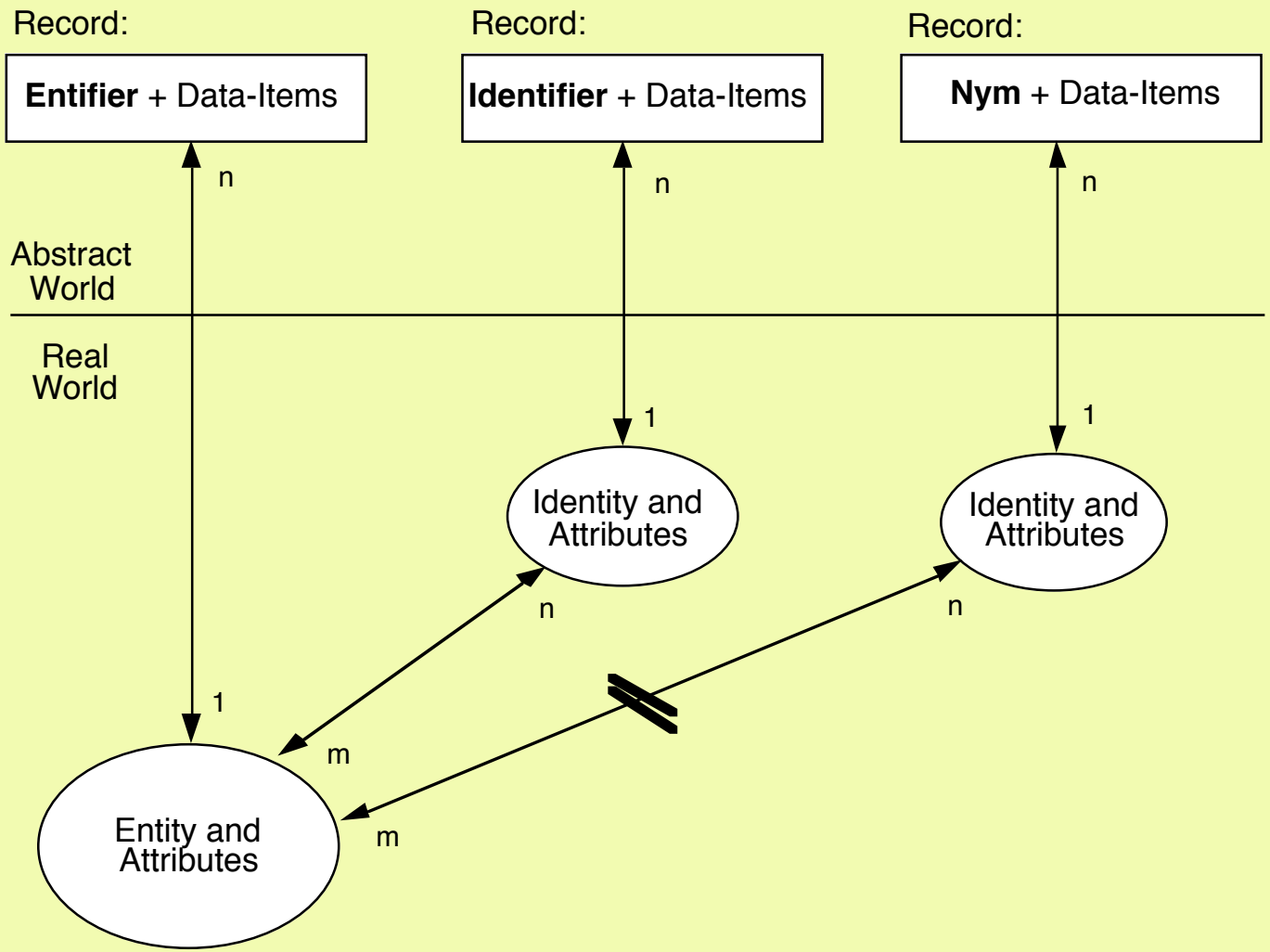
A symbol, sign or measure that is accessible to a person or an artefact

- **Empirical Data** represents a real-world phenomenon
  Synthetic Data does not
- **Quantitative Data on a Ratio Scale**
  is suitable for powerful statistical techniques
  Quantitative Data on Ordinal and Cardinal Scales
  is suitable for less powerful techniques
  Qualitative Data on a Nominal scale
  is subject to limited analytical processes

Data Collection

- **Data Collection is:**
  - for a purpose
  - selective
- **Data Collection processes** are constrained by cost, which inevitably compromises the quality of the data
- **Data may be compressed** at or after collection, e.g. through sampling, averaging and filtering of outliers

Data needs to be Associated with (Id)Entities



XAMAX Consultancy Pty Ltd

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

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 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

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

Key Decision Quality Factors

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

XAMAX Consultancy Pty Ltd

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

Transparency & Accountability: The Quality of Published Research Data

- Of 18 microarray studies, only 2 were fully reproducible using the archived data [27]
- Of 19 papers in population genetics,
  - 30% of analyses could not be reproduced
  - 35% of datasets were incorrectly or insufficiently described [9]
- Of 100 datasets in nonmolecular biology, 56% were incomplete, and 64% were archived such that reuse was much impaired

XAMAX Consultancy Pty Ltd

Roche DG, Kruuk LEB, Lanfear R, Binning SA (2015) Public Data Archiving in Ecology and Evolution: How Well Are We Doing? PLoS Biol 13(11), **10 Nov 2015**

Transparency & Accountability: The State of Play with Research Publications

- " ... reproducibility of much research has become questionable, if not impossible [because it is] shrouded by the opaque use of computers"
- "... new 'scopes' ... see new patterns in ... data ..."
- **Empirical articles need to be supported by:**
  - **the data, defined in open formats**
  - **the data model**
  - **the code**

XAMAX Consultancy Pty Ltd

Marwick B. (2015) 'How computers broke science – and what we can do to fix it' The Conversation, **9 November 2015**, at https://theconversation.com/how-computers-broke-science-and-what-we-can-do-to-fix-it-49938

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

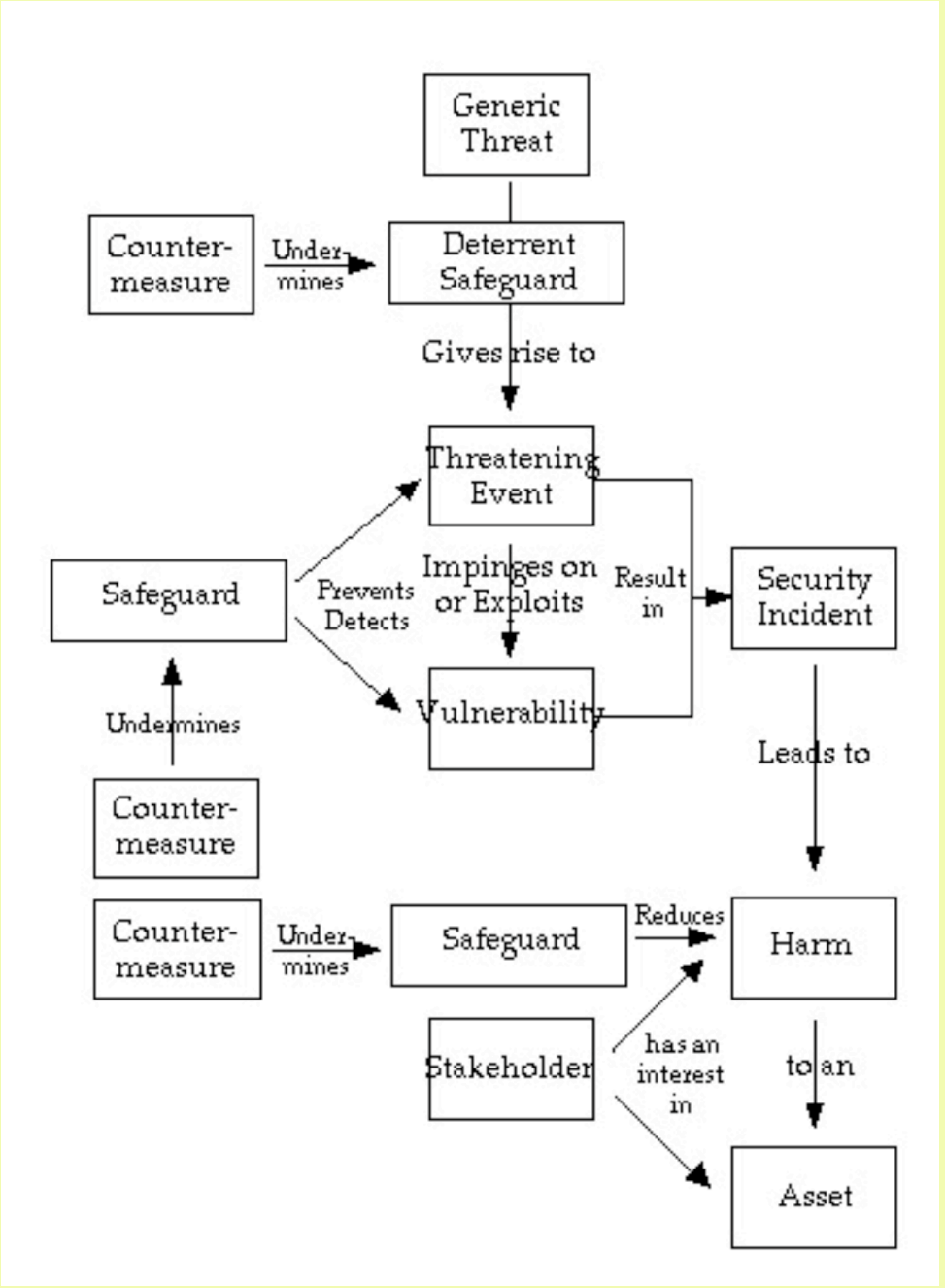
- **Problems It Tries to Address**
  - Missing Data
  - Low and/or Degraded Data Quality
  - Failed and Spurious Record-Matches
  - Differing Definitions, Domains, Applicable Dates
- **How It Works**
  - Internal Checks
  - Inter-Collection Checks
  - Algorithmic / Rule-Based Checks
  - **Checks against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

XAMAX Consultancy Pty Ltd

Big Data Applied to Security

cf.

Security Risks arising from Big Data Applications



http://www.rogerclarke.com/EC/SSACS.html#App1

XAMAX Consultancy Pty Ltd

Contexts of Quality Issues

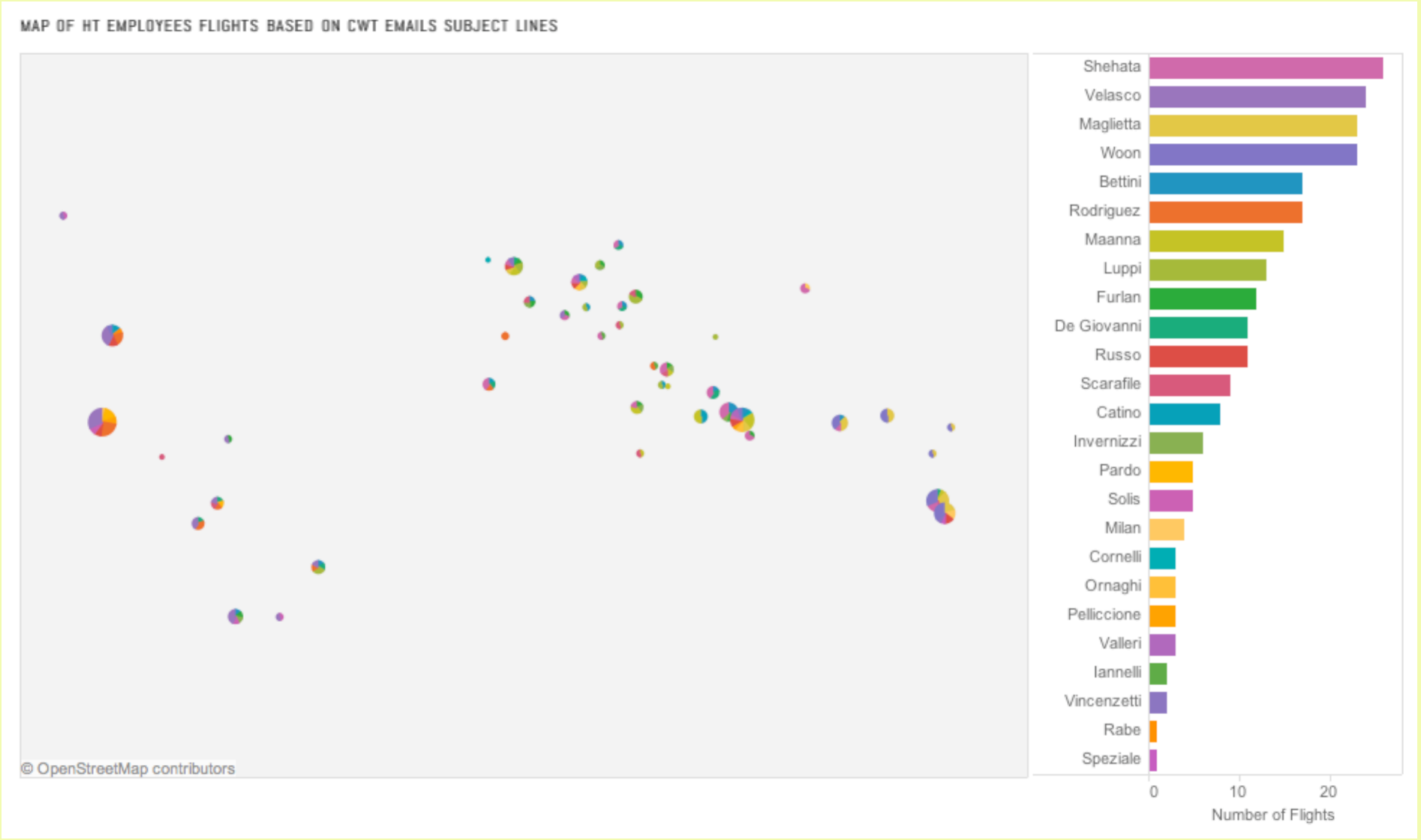
- A Single Data-Collection
- A Consolidated / Multi-Source Data-Collection
- Scrubbing / Cleaning / Cleansing
- Inferencing Processes
- Decision-Making and Action

Organisational Risks – Internal

Security Considerations

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

XAMAX Consultancy Pty Ltd



XAMAX Consultancy Pty Ltd

Hacking Team frequent flyers and locations they visit
http://labs.rs/en/metadata/
29 October 2015

Organisational Risks – Internal

Security Considerations

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

Resource Misallocation

- Negative impacts on ROI or Public Policy outcomes
- Opportunity Costs

XAMAX Consultancy Pty Ltd

Scenario – Insider Detection

The Minister gives terse instructions about whistleblowers
(Brutus, Judas Iscariot, Macbeth, Manning, Snowden, ...)

The agency:

- Increases intrusiveness and frequency of employee vetting
- Lowers the threshold for positive vetting
- Exercises its powers to gain access to and consolidate:
  - border movements • credit history • court records
  - LEA persons-of-interest lists • financial tracking alerts
  - all internal communications • social media postings
- Applies big data analytics to the consolidated database

Personal Risks
Implications for Organisations

- Outlier Discovery
- Inferencing about Individuals

- "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
Implications for Organisations

Discrimination

'Unfair' Discrimination

Breaches of Trust

- Data Re-Purposing
- Data Consolidation
- Data Disclosure

Morale

Active Obfuscation, Falsification

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

Risk Management for Big Data Projects

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

Research Opportunities

- Indicators and Contra-Indicators
  for particular Data Analytic Techniques
- Scenario Analyses ==>> Case Studies
- Data Scrubbing against external reference-points
- Quality Audit Techniques
  for Data Scrubbing and for Inferencing
- Transparency Mechanisms for rule-based,
  neural-net and machine-learning analytics
- Integration into QA, TRA and SRMP processes
- Cognitive Load Management incl. anomaly
  definition, filtering, clustering, prioritisation

XAMAX Consultancy Pty Ltd

Big Data
Quality Assurance

Roger Clarke

Xamax Consultancy, Canberra

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

Redefining R&D Needs for Australian Cyber Security
Australian Centre for Cyber Security (ACCS)
16 November 2015

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