Big Data Prophylactics
How Negative Impacts Can, and Cannot, be Avoided

Roger Clarke

Xamax Consultancy Pty Ltd, Canberra

Visiting Professor, UNSW Law

Visiting Professor, Computer Science, ANU

http://www.rogerclarke.com/DV/KS16 {.html, .pdf}

IFIP Summer School on Privacy & Identity Management

Karlstad SE – 22 August 2016

Vroom, Vroom
Beyond the 'Hype' Factor in Big Data

- Volume
- Velocity
- Variety

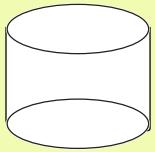
- Value

- **Visibility – Transparency**
- **Veracity – Quality of Data, Inferences**
- **Validity – Effective Outcomes**

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
- 'Fast Data', i.e. streaming

Big Data Analytics

Techniques to draw inferences

XAMAX Consultancy Pty Ltd

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'

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

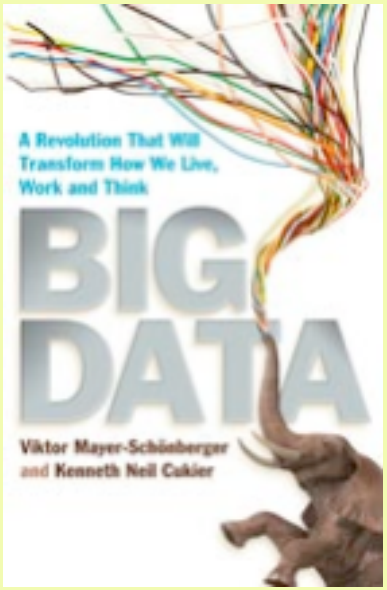
"[F]aced with massive data,
[the old] approach to science
-- hypothesize, model, test -- is ... obsolete.

"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

> 800 Google citations, as at July 2016

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

> 1,800 Google citations, as at July 2016

XAMAX Consultancy Pty Ltd

2. People Need Defences Against Big Data

Big Data Analytics

Techniques for analysing 'Big Data'

Big Data Prophylactics

Safeguards for an entity
against potentially harmful acts on it
by another entity

XAMAX Consultancy Pty Ltd

Reasonable / Naïve Public Expectations

- An organisation that causes harm to a person is financially responsible for the consequences
- Criminal sanctions apply to irresponsible acts

Reasonable / Naïve Public Expectations

- An organisation that causes harm to a person is financially responsible for the consequences
- Criminal sanctions apply to irresponsible acts

- Organisations exercise care undertaking acts that have potentially negative consequences for people
- Organisations exercise particular care when undertaking novel/innovative/experimental acts

Reasonable / Naïve Public Expectations

- An organisation that causes harm to a person is financially responsible for the consequences
- Criminal sanctions apply to irresponsible acts

- Organisations exercise care undertaking acts that have potentially negative consequences for people
- Organisations exercise particular care when undertaking novel/innovative/experimental acts

- **To avoid liabilities, organisations undertake some form of <u>Evaluative Process</u>, prior to performing acts**

XAMAX Consultancy Pty Ltd

Alternative Approaches to Evaluation

	Mainly Quantitative and Financial Data	Mainly Qualitative Data
Mainly the Sponsor's Perspective	Discounted Cash Flow Investment Analysis Financial Sensitivity Analysis Financial Risk Assessment	Internal Cost-Benefit Analysis Risk Assessment
Multiple Stakeholder Perspectives	External or Economic Cost-Benefit Analysis (CBA) Economic Feasibility Assessment	Cost, Benefit and Risk Assessment (COBRA) Economic, Social and Environmental Impact Assessment

Evaluation Technique Example No. 1
Business Case Preparation

- There are many variants, some disciplined and formalised, most pragmatic and informal
- Typically:
  - Spreadsheets, often primarily financial data
  - Cost-Benefit Analysis, but internal-only
- The focus is on:
  - Payback / Return on Investment
  - Alignment with corporate strategy
- **But all are designed to support the proposal!**

Humphrey W.S. (2000) 'Justifying a Process Improvement Proposal' SEI Interactive, March 2000, at http://northhorizons.com/Reference%2520Materials/5%2520Justifying%2520a%2520PIP.pdf

Example 2
Risk Assessment (RA)

Analyse

(1) Define the Objectives and Constraints

(2) Identify the relevant Stakeholders, Assets, Values and categories of Harm

(3) Analyse Threats and Vulnerabilities

(4) Identify existing Safeguards

(5) Identify and Prioritise the Residual Risks

http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf
http://www.rogerclarke.com/EC/PBAR.html#PTC

Example 2
Risk Assessment (RA) and Risk Mngt Planning

Analyse

(1) Define the Objectives and Constraints

(2) Identify the relevant Stakeholders, Assets, Values and categories of Harm

(3) Analyse Threats and Vulnerabilities

(4) Identify existing Safeguards

(5) Identify and Prioritise the Residual Risks

Design

(1) Postulate alternative Designs

(2) Evaluate the alternatives against the Objectives and Constraints

(3) Select a Design (or adapt / refine the alternatives to achieve an acceptable Design)

Do

(1) Plan the implementation

(2) Implement

(3) Review the implementation

XAMAX Consultancy Pty Ltd

http://csrc.nist.gov/publications/nistpubs/800-39/SP800-39-final.pdf

Big Data Risk Factors – 1

Use Categories need to be Distinguished!

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

Bug Data Risk Factors – 2

Data Quality
(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
(Assessable only at time of use)

I1 Theoretical Relevance

I2 Practical Relevance

I3 Currency

I4 Completeness

I5 Controls

I6 Auditability

Data Scrubbing / 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**
  - **Checks against Reference Data – ??**
  - Internal Checks
  - Inter-Collection Checks
  - Algorithmic / Rule-Based Checks
- **Its Implications**
  - Better Quality and More Reliable Inferences
  - **Worse Quality and Less Reliable Inferences**

XAMAX Consultancy Pty Ltd

Big Data Risk Factors – 3
Decision Quality

1. Appropriateness of the Inferencing Technique
2. Data Meaning
3. Data Relevance
4. Transparency
   - Process
   - Criteria

http://www.rogerclarke.com/EC/BDQF.html#DeQF

Transparency

- **Accountability** depends on clarity
  about the Decision Process
  and the Decision Criteria
- **But Transparency is compromised or absent**:
  - Manual decisions – Often poorly-documented
  - Algorithmic languages
    Process & criteria explicit, or at least extractable
  - Rule-based 'Expert Systems' software
    Process implicit; Criteria implicit
  - **Empirical software**
    (neural nets, machine learning)
    Process implicit; Criteria inscrutable!

XAMAX Consultancy Pty Ltd

Organisational Risks

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

Personal Risks
Implications for Individuals

- Outlier Discovery
- Inferencing about the Individual based on a digital persona

- "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

Personal Risks
Implications for Organisations

Discrimination

'Unfair' Discrimination

Breaches of Trust

- Data Re-Purposing
- Data Consolidation
- Data Disclosure

Morale

Active Obfuscation, Falsification

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

Example 3
Privacy Impact Assessment

Privacy impact assessment (PIA) is:

- a systematic process, which ...
- identifies and evaluates ...
- from the perspectives of all stakeholders ...
- the potential effects on privacy of ...
- a project, initiative or proposed system or scheme
- and which includes a search for ways to avoid or mitigate negative privacy impacts

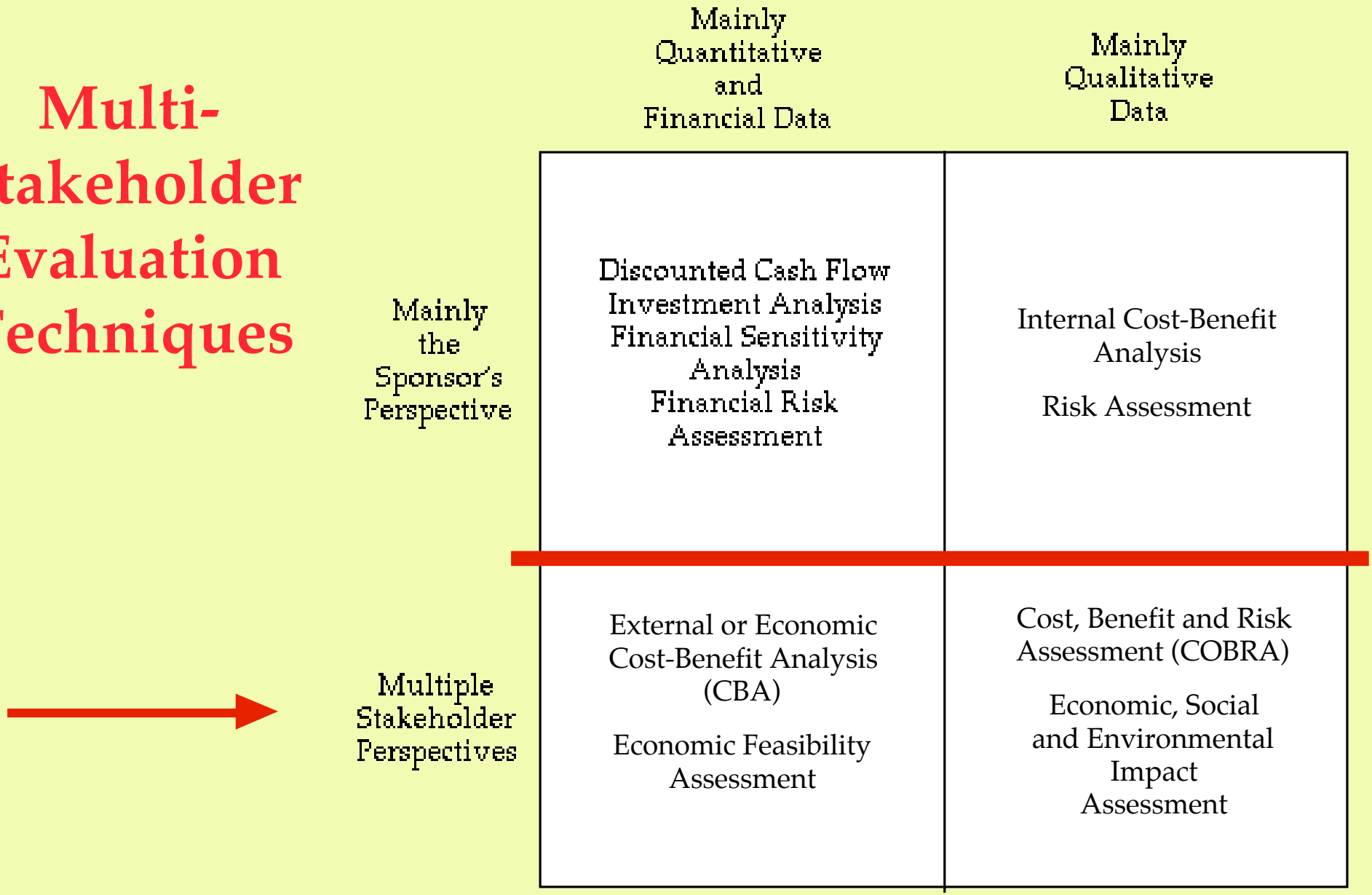
Embedment of PIA in Risk Assessment?

- The evolution of PIAs needs to be seen within the context of larger **trends in advanced industrial societies to manage 'risk'** and to impose the burden of proof for the harmlessness of a new technology, process, service or product on its promoters (Raab 2004)
- Ontario's 1999 Guidelines were revised this way in 2007
- Wright et al. (2014) showed it was feasible but difficult
- The problem comes down to imbalance of power: **head-on clashes between organisational and other interests are resolved in the organisation's favour**

XAMAX Consultancy Pty Ltd

Wright D., Wadhwa K., Lagazio M., Raab C. & Charikane E. (2014) 'Integrating privacy impact assessment in risk management' Int'l Data Privacy Law 4, 2 (May 2014) 155-170

Multi-Stakeholder Evaluation Techniques



http://www.rogerclarke.com/EC/PETsBusCase.html#BC (2008)

Reasonable / Naïve Public Expectations Have Not Been, and Will Not Be, Fulfilled

- Organisations don't undertake evaluation processes that reflect multiple Stakeholders' interests
- So the requirement has to be imposed from without

Option 1: Regulatory Action

Regulatory Forms

Forms: / Actors:	**Formal Regulation** ('Government')	**Co-Regulation**	**Industry Self-Regulation**	**Organisational Self-Regulation** ('Governance')
The State	**Determines What and How**	**Negotiates What and How**	Influences What	Has Limited Influence
Industry Assocn	Influences What and How	**Negotiates What and How**	**Determines What and How**	Influences What and How
Corporations	Contribute to Industry Assocn	Contribute to Industry Assocn	Contribute to Industry Assocn	**Determine What and How**
Other Stakeholders	May or May Not Have Some Influence	May or May Not Have Some Influence	May or May Not Have Some Influence	May or May Not Have Some Influence
	Statutes & Delegated Legislation	Statutory Codes & Standards	Industry Codes & Standards	Customer Charters

Case 1: Industry and Professional Codes

UNSD (1985) 'Declaration of Professional Ethics'
United Nations Statistical Division, August 1985, at http://unstats.un.org/unsd/dnss/docViewer.aspx?docID=93#start

ASA (2016) 'Ethical Guidelines for Statistical Practice'
American Statistical Association, April 2016, at http://ww2.amstat.org/about/pdfs/EthicalGuidelines.pdf

DSA (2016) 'Data Science Code Of Professional Conduct'
Data Science Association, undated but apparently of 2016, at http://www.datascienceassn.org/sites/default/files/datasciencecodeofprofessionalconduct.pdf

UKCO (2016) 'Data Science Ethical Framework'
UK Cabinet Office, v.1, 19 May 2016, at https://www.gov.uk/government/publications/data-science-ethical-framework

Case 2: PIAs as a Control Mechanism
The Five-Factor Test

1. Is there evidence of a PIA process being **performed**?
2. Were advocacy organisations **aware** of that process?
3. Did the project sponsor(s) **engage** with advocacy organisations?
4. Was the PIA Report **published** on completion?
5. Were advocacy organisations' views appropriately **reflected** in the PIA Report?

   However, it was known that there was a low incidence of published Reports. Hence:

6. Did the PIA Report come to light later, e.g. as a result of an FoI request by the media?

PIAs don't operate as a Control Mechanism over Australian National Security Initiatives

AGD

- **Passed** the 5-factor test **2/36**
- Engagement with advocacy organisations 3/36 (but their views were ignored)
- Secret (hence flawed) PIA processes 10/36

Other Agencies

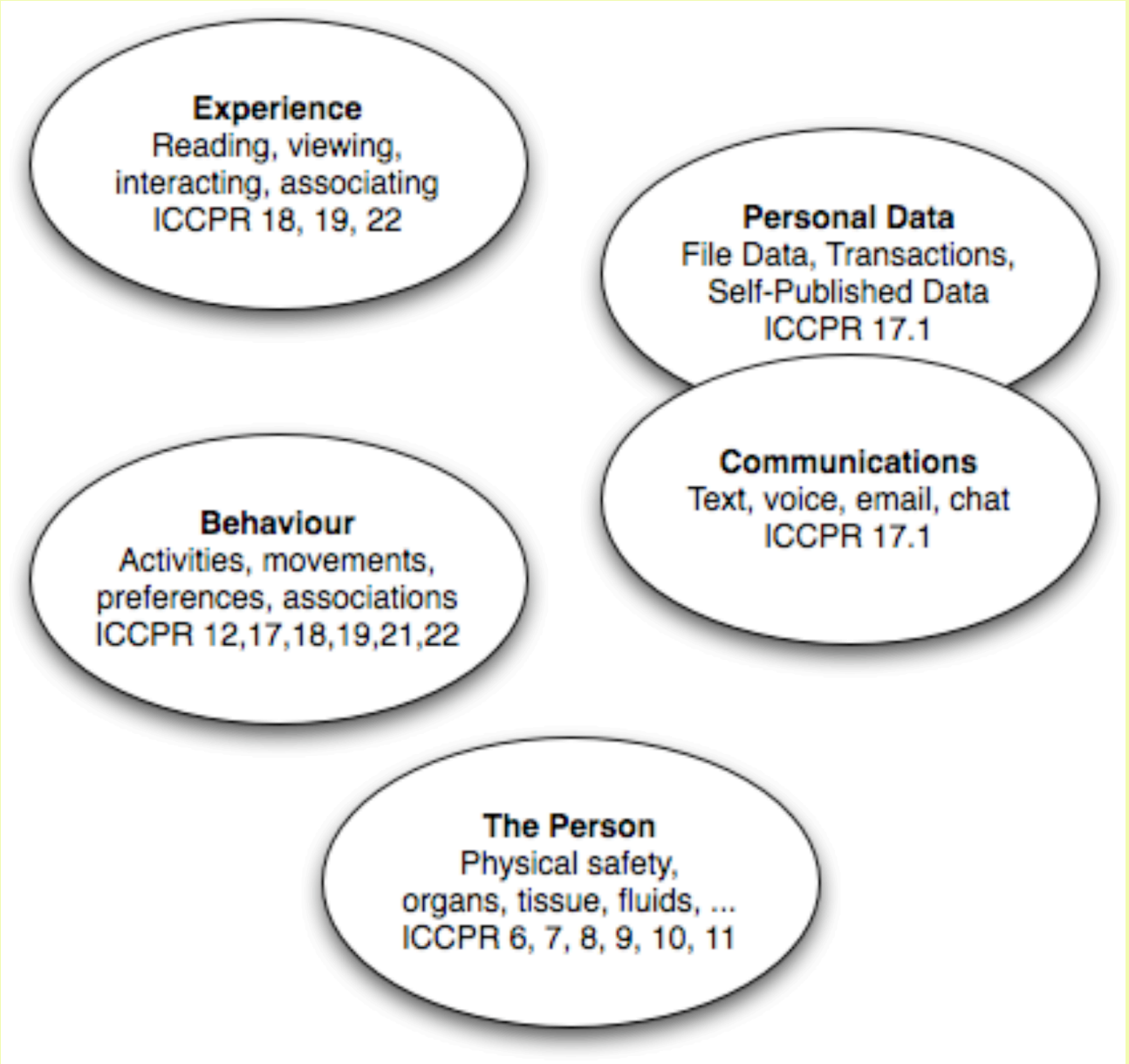
- **Passed** the 5-factor test **1/36**
- Engagement with advocacy organisations 5/36

XAMAX Consultancy Pty Ltd

Case 3: The EC's GDPR
Data Protection Impact Assessment ('DPIA')

- The Trigger (Art. 35.1-35.6):
  Only '**high risks** to the rights and freedoms of data subjects' ...
- 'An assessment of the **impact of** the envisaged processing **operations [only] on the protection of personal data**' (35.1). So:
  - **not driven by social values**, and will be interpreted as a mere Data Protection Law Compliance Assessment
  - **not all five dimensions, and not even data privacy**, but merely the sub-set that is subject to data protection
- **Seeking civil society's views is optional, and there is no requirement that they be reflected in the design** (35.9)
- **Exemption** for authorised programs (35.10)
- **Feature implementation is optional**, ditto review (35.7(d), 35.11)

The Multiple Dimensions of 'Personal Space'



http://www.rogerclarke.com/DV/Intro.html (1997)
Finn, Wright & Friedewald (2013)
Wright & Raab (2014) Koops, Newell et al. (2016)

A DPIA isn't a PIA

(1) It's merely a Privacy Law Compliance Audit

(2) There's no need to do anything afterwards

- "a methodical ...
- and independent ...
- assurance process ...
- to elicit evidence ...
- **to establish whether practices conform with [insert the legal authority/ies] ...**
- **to identify deficiencies** and ...
- to indicate how deficiencies will be eliminated"

Case 4: The Precautionary Principle

Strong / Legal Form:

"When human activities may lead to morally unacceptable harm that is scientifically plausible but uncertain, **actions shall be taken to avoid or diminish that [potential] harm**"

http://unesdoc.unesco.org/images/0013/001395/139578e.pdf

Moderate / Moral Form:

'If an action or policy is suspected of causing harm, and scientific consensus that it is not harmful is lacking, **the burden of proof ... falls on those taking [the] action**'

After https://en.wikipedia.org/wiki/Precautionary_principle

The Precautionary Principle in Australian Environmental Law

If:

(1) a threat of serious or irreversible environmental damage exists; &

(2) there is scientific uncertainty as to the extent of possible damage

Then:

A. precautionary measures may be imposed by the court to avert the anticipated threat, but such measures must be appropriate and proportionate

XAMAX Consultancy Pty Ltd

Telstra Corporation Limited v Hornsby Shire Council [2006] NSWLEC 133 (24 March 2006), esp. paras. 113-183
http://www.austlii.edu.au/au/cases/nsw/NSWLEC/2006/133.html

The Precautionary Principle

Strong / Legal Form (in some environment laws only):

"When human activities may lead to morally unacceptable harm that is scientifically plausible but uncertain, **actions shall be taken to avoid or diminish that [potential] harm**"

http://unesdoc.unesco.org/images/0013/001395/139578e.pdf

<u>Moderate / Moral Form (much-discussed, seldom imposed)</u>:

'If an action or policy is suspected of causing harm, and scientific consensus that it is not harmful is lacking, **the burden of proof ... falls on those taking [the] action**'

After https://en.wikipedia.org/wiki/Precautionary_principle

Reasonable / Naïve Public Expectations Have Not Been, and Will Not Be, Fulfilled

- Organisations don't undertake evaluation processes that reflect multiple Stakeholders' interests
- So the requirement has to be imposed from without

Option 1: Regulatory Action – is largely a failure

Option 2: Public Activism

XAMAX Consultancy Pty Ltd

4. Public Activism

Civil Disobedience

- Obfuscation of data, traffic, location, identity
- Falsification of data, traffic, location, identity

Public Pressure

- Organisation, Coordination, Targetting
- Channels
  - Media (in decline)
  - Social Media (unprecedented scope)

Davies S. (2014) 'Ideas for Change: Campaign principles that shift the world' The Privacy Surgeon, December 2014 http://www.privacysurgeon.org/resources/ideas-for-change/

5. Public Activism
A Specific Proposition

Civil Society Standards

The Politics of Standards

- Institutionalisation and Scale
- Influence determined meritocratically
- From Volunteer Professionals

  To Corporations, Government Agencies, Industry Associations

- **Consumers / Citizens / Reps / Advocates ?**
- Unfluence muted and even nil, due to:
  - Dominance of Meritocracy
  - Dominance of Corporate Power
  - Limited Resources for Analysis, Travel

Civil Society should create its own Standards

- **An alternative voice** to the documents published by and for industry and governments
- **Counter-balance** to the power of industry and governments
- **Antidote to civil society's exclusion / weak voice** in industry standards processes

- **Public Expectations**:
  - Articulated
  - Communicated
  - Available in Advance
- **Benchmarks**:
  - Established
  - Applied by Civil Society
  - Applied by Others
- Avoidance of public harm from badly conceived projects
- Avoidance of public and private investment failures

Subject-Matter for Civil Society Standards

Meta-Principles for the Evaluation of Initiatives

Processes

- Generic, for the evaluation of initiatives
- Specific, for the evaluation of categories of initiatives
- Quality Assurance
- Audit of Processes, of Outcomes
- Consultation / Engagement
- Complaints-Handling

Checklists

- Mitigation Measures
- Controls

Evaluation Meta-Principles

Pre-Conditions

1. Evaluation
2. Consultation
3. Transparency
4. Justification

Design

5. Proportionality
6. Mitigation
7. Controls

Post-Condition

8. Audit

Summary

- Big Data is more ideological than analytical
- People need defences against Big Data
- Organisations' evaluation techniques fail the test
- Multi-Stakeholder Evaluation must be imposed
- Not only is 'market failure' evident,
  but 'regulatory failure' is as well
- Hence public activism is needed
- Civil Society Standards are proposed

XAMAX Consultancy Pty Ltd

Research Opportunities

Clearly Distinguished from Advocacy!

- Beyond Scenarios to Deep Case Studies
- Audit of Data Collections
- Audit of Merged Data Collections
- Audit of Big Data Analytics Process and Outputs
- Evaluation of Data Analytics Outcomes
- Evaluation of Codes and Guidelines against Normative Standards
- Critical Theory Research into the exercise of power

Big Data Prophylactics
How Negative Impacts Can, and Cannot, be Avoided

Roger Clarke

Xamax Consultancy Pty Ltd, Canberra

Visiting Professor, UNSW Law

Visiting Professor, Computer Science, ANU

http://www.rogerclarke.com/DV/KS16 {.html, .pdf}

IFIP Summer School on Privacy & Identity Management

Karlstad SE – 22 August 2016