The Ideology of Big Data Analytics (core assumptions & techniques)

Graham Greenleaf
Professor of Law & Information Systems, UNSW Australia

Information Environmentalism Conference, Fremantle WA
2-5 February 2016

Big data analytics

- *Big data analytics* is
  - the processing of very large data sets, containing data varying widely in types, structures and sources;
  - requiring significant hardware (eg parallel processing);
  - by the use of new programming and database methods ('analytics');
  - in order to obtain correlations between data, and to take actions based on them.
- Prefer 'big data analytics' (BDA) to 'big data' or 'data analytics'– it stresses both components

Aims

1. Can an ideology be identified in the aims, assumptions and techniques of the proponents and practitioners of BDA?
2. What are the implications of that ideology when BDA are applied to personal data? (privacy & other interests)
3. How is BDA ideology reflecting in legal 'reforms' ?
   - Example: new Japanese law, contrast EU GDPR

Scope – People

1 These arguments are only fully relevant to:
- Data which is directly or indirectly about individuals *(conventional 'personal data')*;
- OR Data which can be used to impute/predict characteristics to/of people, individually or collectively *(broader than current data privacy laws);*
- AND can be used to make decisions affecting them as individuals, whether or not their identity is known.

2 Many BDA applications don't have such effects
- May be valuable, or methodologically suspect, but they raise *different* issues

'Ideology'?

1. Dictionary: 'a system of ideas and ideals, especially one which forms the basis of economic or political theory and policy.'
2. I argue BDA has common elements of scientific method, knowledge (epistemology), what exists (ontology), what is worth valuing (ethics), and consequent approaches to decision-making, surveillance, law and social relationships (economics & politics)
3. Barocas & Nissenbaum: a *paradigm*, rather than a particular set of tools and practices; 'a way of thinking about knowledge through data and a framework for supporting decision-making'; 'a belief in the power of finely observed patterns … drawn inductively from massive datasets.'

'Ideology' – Caveats

1. Any description of an ideology is a caricature, emphasizing what is seen as characteristic
2. Most adherents do not embrace every aspect of an ideology, even though the 'family resemblance' fits
3. This draft is based on too few sources
4. Aim is to find a logical order of exposition, not to critique here
5. Criticisms and suggestions are invited

Elements of Big Data Analytics ideology

I have put under 6 headings many distinct assumptions and techniques characteristic of BDA:

I. *Forget causation, follow the correlations*
II. *Only a datafied world is understandable and valuable*
III. *Laws (eg data privacy) may need adjustment to fit BDA*
IV. *Collect, use, combine & retain all possible data – anonymise minimally*
V. *Lower data quality & retrieval are justified by results*
VI. *Transparency is not necessary – to developers, users or data subjects*

(I) Forget causation, follow the correlations

1. Correlations are sufficient for prediction/actions
   - Correlation is a sufficient proxy for causation: 'Sophisticated computational analyses can now identify the optimal proxy' (MS&C) without hypotheses
   - Correlation enables prediction (MS&C) – using combinations of data mining and machine learning (ML)
   - Spurious correlations can be avoided (MS&C)
   - 'Petabytes allow us to say "Correlation is enough"' (Anderson)
2. Correlation-based argument is now necessary & desirable
   - Causation is over-rated (Anderson): "knocked off its pedestal as the primary foundation of meaning"; rarely needed or desirable (MS&C)
   - The 'end of theory' (Anderson) does causality exist? (MS&C) - prior hypotheses, and deductive/inductive reasoning discounted;
   - Influence of increased ease of collection/storage of massive data.
   - Human expertise is needed less, and less affordable (MS&C)

(II) Only a datafied world is understandable and valuable

3. 'Datafication' is an epistemology
   - The world is most usefully viewed as data; datafication is its conversion to 'a quantified format' allowing analysis (M&S)
   - **All experience** can be and should be datafied; a 'great infrastructure project' (MS&C ch.5)
   - Precursors in both (since 1800) the ubiquity of accounting and the 'probabilistic revolution' of statistics (Ambrose)
4. 'Datafication' is an ethic
   - All datafication creates value ('**new oil**')
   - **Surveillance is innovation** and creates value (contra: J Cohen)
   - '**Data's value** [is] all the possible ways it can be employed in the future' (MS&C 103)
   - Data's worth is primarily its **re-use** value, not the purpose of its collection, or value in organisational operations.

(III) Laws may need adjustment to fit Big Data

These steps are necessary here or elements 7-14 in BDA ideology will repeatedly conflict with data privacy (+other) laws

5. Data privacy laws are only about notice & consent, and anonymisation
   - 'the three *core strategies* long used to ensure privacy … have lost much of their effectiveness' (a 'Maginot line') (MS&C)
   - *Big Data ideology starts with a misrepresentation of existing data privacy law*, based on US assumptions
   - Most of the 108 other jurisdictions with data privacy laws have to be regarded as irrelevant, for this step to succeed.

Laws may need adjustment …

6. De-identification removes other privacy problems
   - Effective (acceptable) anonymisation is possible – contra Ohm
   - There are degrees of personal data (redefinition of PI) – see Japanese example
   - Effects on groups through analytics are not effects on individuals
     - Contra: Non-identified individuals are still 'reachable' (B&N 2014)
7. Data privacy laws must not expand beyond 'identifiability'
   - Current laws (in all jurisdictions, even EU) give scope for BDA because PI is limited to 'identifiability'
   - Extension to data having 'individual effects' would endanger BDA
   - Subsidiary argument is that laws should be 'harm-based', and those whose data is 'merely' used are not harmed.

Discussed further in conclusion on Japan and EU

(IV) Collect, use, combine & retain all possible data – anonymise minimally

8. Data collection/combination/retention should be maximised ('big')
   - **Reasons** for collecting data are irrelevant (correlations unknown)
   - **Heterogenous** data sets should be combined (correlations result)
   - Big Data entails **borderless** data – correlations cross borders
   - **Duration** of potential 'analytic' uses are unpredictable (MS&C)
9. Most data collection is unproblematic

These collection situations & methods are acceptable:

- from the Internet-of-things (and all forms of 'digital exhaust')
- accessible / public data (contra b&C 2012)
- from workplace, classroom and other monitoring of behaviours
- from all aspects of datafication: 'Many of the inherent limitations on the collection of data no longer exist' (MS&C 101)

(V) Lower data quality & retrieval are justified by results

Only some new Internet businesses can insist on homogenous data; the rest have to justify merging vast heterogenous data sets outside their creation.

10. Compromised data quality is acceptable: 'More trumps better'
    - Comprehensive data is best : Bias against samples (MS&C)
    - Data scrubbing is necessary but unproblematic (contra b&C; B&H)); heterogenous data sets can/must be 'normalised' to create value
    - Few studies even question quality of Big Data - assumed (Clarke).
11. Techniques reducing search/retrieval precision are acceptable
    - Only minimal data structures can be achieved; scrubbing is too costly
    - Sophisticated structured searching is therefore impossible
    - 'Messyness' is OK: relax standards for allowable errors (MS&C)
    - 'it is more productive to tolerate error than to work at preventing it' (MS&C); 'Lossy' search results should be tolerated (MS&C 46)

(VI) Transparency is not necessary

The lack of transparency applies to developers, users and data subjects

12. Programming of analytics is value-neutral
    - Machine-learning (ML) = predictions from BD, based on known properties learned from training data; Data mining (DM) = discovery of unknown properties in BD. DM can provide training data for ML.
    - ML & DM give appearance of neutral delegation; Programs based on ML have a weak authorship nexus.
    - ML developers don't necessarily understand *why* they work
13. Programming does not require causal explanation generation
    - ML/DM is not rule-based AI: Procedural programs can give a single explanation of how a result was reached; Declarative programs can give alternative explanations of how to reach a result. Both involve explicit programming of 'rules', and implicitly of explanations.
    - ML programs cannot give meaningful explanations, in a causal sense. They may be able to provide feedback of correlations used.

Transparency is not necessary

14. Usage does not require transparency

- Organisations can and do use data analytics without understanding why they generate results.
- Individuals can be affected by BDA without knowing either that this occurred; and if they do the user cannot explain the result, other than that they 'did not meet the desired profile'.
- BDA proponents can/must justify both transparency gaps by saying that correlations are sufficient justification (ie step 1/14).

Recap: 14 characteristics of BDA

1 Correlations are sufficient for prediction/actions
2 Correlation-based argument is now necessary & desirable
3 Datafication' is an epistemology
4 'Datafication' is an ethic
5 Data privacy laws are only about notice & consent, and anonymisation
6 De-identification removes other privacy problems
7 Data privacy laws must not expand beyond 'identifiability'
8 Data collection/combination/ retention should be maximised ('big')
9 Most data collection is unproblematic
10 Compromised data quality is acceptable: 'More trumps better'
11 Techniques reducing search/ retrieval precision are acceptable
12 Programming of analytics is value-neutral
12. Programming does not require causal explanation generation
14. Usage does not require transparency

Suggestions

1 Uses of BDA in anything close to the ideological way suggested can easily result in opaque adverse consequences to individuals

2 Many BDA practices are not consistent with existing privacy laws, if enforced

3 Some dangerous BDA practices operate outside the scope of existing laws

Adjusting data privacy laws to BD: Japan's revised PIPA law (2015)

- Defines 'anonymous processed information' (API) as 'information related to an individual that was obtained by processing personal information [PI] such that a specific individual cannot be identified, and so that such PI cannot be restored…'
- Then *mandates* the 2 methods to achieve this:
  i. ID codes/biometrics must be replaced by other random codes
  ii. Descriptive information must be randomly replaced.
- PIPC (DPA) Rules can also give instructions to achieve this
- Required methods will often not deliver anonymity
  - But if businesses comply with the Rules, law is satisfied.

Adjusting data privacy laws to BD: Japan

- Obligations of businesses handling API:
  - Security of deleted PI, and method of creating API
  - Publicly announce whether they will create API, and if they will give it to 3rd Ps (+ if so, must inform them)
  - Anyone processing API (i) must not re-identify it; (ii) have same security + disclosure obligations.
- API can therefore be used, disclosed, transferred + retained contrary to normal rules for PI
  - But separate 'API rules' aim to stop API re-identification
  - API will still affect individual interests, without being PI
  - Will weak anonymisation really prevent re-identification?

Adjusting BD to data privacy law: EU General Data Protection Regulation (2016)

- GDPR retains 'identifiable' as the core of PI
  - Test of identifiability must take account of 'all the means reasonably likely to be used' (including available technologies and cost), by controller or any 3rd P, to re-identify.
  - If anonymous, not PI. No positive category like 'API'; opposite approach of 'pseudonymous = PI, unless it fails the identifiability test'.
  - A broad approach to when online identifiers (eg RFID, IP addresses) may constitute PI.
  - But EU has not extended GDPR beyond 'identifiable' – plenty of leeway for some BD.

Adjusting BD to data privacy law: GDPR

Some GDPR barriers to Big Data ideology:

1. Essentially a single law and enforcement system across the EU
2. Tougher penalties including up to 4% of global turnover (No 'Maginot Line'?)
3. Collection requires explicit purposes, and further processing is limited to compatible purposes
4. Retention for no longer than these purposes
5. Additional limits on sensitive data categories, including strict limits on automated decisions
6. Various rights of rectification, erasure etc
7. Right not to be subjected to decisions based solely on automated processing, which produces legal effects or 'similarly significant affects him or her'.
8. Cross-border transfers limited to where 3rd P country provides 'adequate' protection (now 'essentially equivalent' to EU)

BUT: Processing can now be justified by the legitimate interests of a 3rd party recipient, consistent with the data subject's expectations. (B&B)

Cited sources for Big Data ideology

Proponents of Big Data analytics:

1 C.Anderson 'The end of theory: The Data Deluge Makes the Scientific Method Obsolete' (2008) 16(7) Wired

2 [MS&C] V.Mayer-Schönberger & K.Cukier *Big Data* (Mariner Books, 2013)

More critical:

1 [b&c] d. boyd and K.Crawford, 'Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon' (2012) *15 Information, Communication & Society* 662, 663

2 [B&H] M Burdon & P Harpur (2014) 'Re-conceptualising privacy and discrimination in an age of talent analytics (2014) 37(2) UNSWLJ 679-712

3 J Cohen, 'The Surveillance-Innovation Complex: The Irony of the Participatory Turn' in D.Barney et al (eds), *The Participatory Condition* (U. Minn. Press, 2015)

4 [B&N] S Barocas & H Nissenbaum 'Big Data's End Run around Anonymity and Consent' in Lane, Stodden, Bender & Nissenbaum (Eds)

5 Clarke, R 'Quality factors in Big Data and Big Data Analytics' 2014

Other References

1 M.L Ambrose 'Lessons from the avalanche of numbers: Big Data in historical context' I/S: A Journal of Law and Policy for the Information Society, 2014-5

2 [B&B] C Bennett and R Bayley 'Privacy Protection in the Era of 'Big Data': Regulatory challenges and social assessments' (2015)

3 D.Bollier (Rapporteur) *The Promise and Peril of Big Data* (Aspen Institute, 2010)

4 P. Leonard '"Big Data" business: Evolving business models and privacy regulation' (2013)

5 O.Tene and J.Polonetsky, 'Big Data for All: Privacy and User Control in the Age of Analytics', (2013) 11 Nw. J. Tech. & Intell. Prop. 239

6 L.Bennett Moses & J.Chan 'Using big data for legal and law enforcement decisions: Testing the new tools'(2014) 37(2) UNSWLJ 643-678

7 G Greenleaf ' Abandon All Hope?' (Foreword) 37(2) *UNSW Law Journal* 636-642

8 G Greenleaf, G 'Japan: Toward international standards – except for 'big data' (2015) 135 *Privacy Laws & Business International Report,* 12-14

9 Lane, Stodden, Bender & Nissenbaum (Eds) *Privacy, Big Data and the Public Good*, CUP 2014

Acknowledgments: *Valuable comments from* Roger Clarke; Mark Burdon