Ethical pitfalls of Big Data
ENBIS 2016

Joanna Berry
Durham University
Shirley Coleman
Newcastle University

Durham University Business School

Newcastle University

Big Data – increasingly part of the global consciousness

- Define Data
- Are we in a new age of instability?
  - Increasing velocity, variety and volume of data creates insecurity – are we losing control?
  - Is Google's algorithm more important and influential than government policy?
- Big Data is not binary – neither "good" nor "bad"
- Practical problem remains re: storage and compression
- Philosophical problem has become re: ethics
  - Job automation
  - AI
  - Technology's impact on the environment and the poor
  - Privacy (RT @IWantPrivacy)

Velocity, Variety, Volume…

- Volume - the size of the data, such as terabyte- or petabyte-sized databases
- Velocity – the time intervals between data arriving and the time when analysis can be produced and shown; this is often in batch or near-real-time in the big data environment
- Variety - the different data types available. These can include structured data from databases and tables and unstructured data from videos, images, audio, SMS, and social media

And what about Veracity?

- Veracity?
  - The danger of data fumes: data available for collection have been shaped by the affordances of the apps in question
    - Determined by a small, homogenous group of developers
    - Datasets available from social media platforms are inherently exclusionary
  - “the very limits of knowledge are set through the data infrastructure of private corporations”

And what about Value?

- Value?
  - How useful is analysis of big data when data sets are inherently exclusionary, because of the populations represented as well as the methodologies used to harness them?
  - Should policy that impacts one customer base be based on the ("big") data of the few?
  - Is Big Data a resource to be consumed or a force to be controlled?

A sixth V?

- Voyeurism
- Google, WhatsApp, Bing, Facebook…can all see
  - Who you are
  - What you earn, what you spend
  - Who you meet and where
  - Who you are thinking of meeting
  - Who might be thinking of meeting you
  - Where you are thinking of going….
- Mosaicked deidentification
- Differential Privacy (a framework for formalising privacy in statistical databases)
- Unicity (uniqueness – 4 spatio-temporal points are enough to uniquely identify 95% of 1.5m people in a database)

The Power of the Password

- Collaborative filtering
- If you liked that you will like this
- Manipulation?
- Steering you to books, films, cultural experiences, life partners….
- The power of the algorithm

The power of the password

- Some (encrypted) data can only be accessed through use of password
- Privacy is hugely dependent upon the individual's desire to be private
- Password as 'PASSWORD" or '1234" or birthdays, names of spouses/children….
- Does privacy matter? Isn't big data already anonymous?

Anonymity

- Even if you want to be anonymous
  - can you be?
    - "raw data is both an oxymoron and a bad idea: to the contrary, data should be cooked with care" (Bowker 2005, p183 – 184)
  - should you be?
    - Is participation in the big data project the responsibility of all good citizens?
    - Why be concerned about releasing individual data if it can help many others?

Cultural differences

- High context vs low context
- Data – big or otherwise – does not include and is not inclusive of high context communicators
- Data sets are not neutral – they require active interpretation by individuals who have their own ways of seeing
- The revolutions were tweeted
- And what of the strength of weak ties in an age of increasingly weak ties?

Ethics

- These issues are only as strong as the ethics of the individual
- Varying with culture, generation, age…
  - Imbalance of exchange of data
  - Companies “rob” data from individuals
  - We don’t realise this
  - Even if we do, we often don’t care, or know why we should

The Last Guardians

- Of data protection
- Of individuality and the self
- Dataism, anyone?
  - The pre-eminence of information and algorithms that can replace/predict/foresee human instincts (eg Tinder)
  - Not quite The Singularity BUT – all computers communicate with each other so everything is trackable
- Techno-humanism
  - The singularity
  - From an internet of things to an internet of minds?

Ethical pitfalls are therefore…

- Obviously – using individual data without individuals permission – but what if they don't know they can be identified?
- Storing data about anyone
- Should concerns about the privacy of individual data be sacrificed for the greater good?
- How valuable is 'big data' gathered from a prescribed data set (only users of Facebook/Twitter/RBS/UCU)?
- How valuable is policy when based on 'useful' statistics?

Literature

- Harari Y. N. (2016) Homo Deus: A Brief History of Tomorrow; Harvill Secker
- Boyd, D; Crawford, K; (2012) Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon; *Information Communication and Society, vol 15, 2012, issue 5 pp* 662-679
- Crawford, K: Miltner; Gray, M. L.; (2014) Critiquing Big Data: Politics, Ethics, Epistemology, International Journal of Communication 8 pp 1663–1672
- Bowker, G. C. (2005) Memory Practices in the Sciences. MIT Press, Cambridge, Massachusetts.
- boyd, d. and Marwick, A. (2011) 'Social Privacy in Networked Publics: Teens' Attitudes, Practices, and Strategies,' paper given at Oxford Internet Institute. [online] Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1925128 (28 September 2011).
- Anderson, C. (2008) 'The End of Theory, Will the Data Deluge Makes the Scientific Method Obsolete?', Edge. [online] Available at: http://www.edge.org/3rd_culture/anderson08/anderson08_index.html (25 July 2011).
- Bollier, D. (2010) 'The promise and peril of big data', [Online] Available at: http:// www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_ Promise_and_Peril_of_Big_Data.pdf (11 July 2011).
- Gitelman, L. (2011) Notes for the upcoming collection 'Raw Data' is an Oxymoron, [online] Available at: https://files.nyu.edu/lg91/public/ (23 July 2011).
- Granovetter, M. S. (1973) 'The Strength of Weak Ties,' American Journal of Sociology vol. 78, no. 6, pp. 1360-80.
- Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., & boyd, d. (2011). 'The Revolutions Were Tweeted: Information Flows during the 2011 Tunisian and Egyptian Revolutions,' International Journal of Communications vol. 5, Feature 1375–1405.
- Meeder, B., Tam, J., Gage Kelley, P., & Faith Cranor, L. (2010) 'RT @IWantPrivacy: Widespread Violation of Privacy Settings in the Twitter Social Network', Paper presented at Web 2.0 Security and Privacy, W2SP 2011, Oakland, CA.
- Shamma, D.A., Kennedy, L., and Churchill, E.F.. (2010). 'Tweetgeist: Can the Twitter Timeline Reveal the Structure of Broadcast Events?,' CSCW 2010.
- Zimmer, M. (2008) 'More on the 'Anonymity' of the Facebook Dataset – It's Harvard College', MichaelZimmer.org Blog, [online] Available at: http://www.michaelzimmer.org/2008/01/03/more-on-the-anonymity-of-the-facebook- dataset-its-harvard-college/ (20 June 2011).
- L. Sweeney, Simple Demographics Often Identify People Uniquely. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.

Ethical issue of insight