ITEC874 — Big Data Technologies

Week 03 Lecture 1: Big Data and Society

Diego Mollá

ITEC874 2019H2

Abstract

In this lecture we will explore some of the key impacts of Big Data in Society, with special attention to the ethical implications of processing and managing Big Data.

Update August 8, 2019

Contents

Reading

- Danah Boyd & Kate Crawford. Critical Questions for Big Data. *Information, Communication & Society*, Vol 15, No 5, June 2012, pp.662–679.

Tentative Outline of ITEC874 — Weeks 1 to 6

1. Introduction
2. Organising Big Data
3. *Big Data and Society*
4. Indexing Big Data
5. Searching Big Data
6. Processing Big Data

Tentative Outline of ITEC874 — Weeks 7 to 12

7. Industry Talk: TBA
8. Analysing Big Data
9. Analysing Unstructured Data
10. Visualising Big Data
11. Analysing Streaming Data
12. Industry Talk: TBA

1 Big Data and Society

Impact of Technology in Society

Kranzberg 1986, p545, **https://doi.org/10.2307/3105385**
"Technology is neither good nor bad; nor is it neutral ... technology's interaction with the social echology is such that technical developments frequently have environmental, social, and human consequences that go far beyond the immediate purposes of the technical devices and practices themselves."

https://www.findagrave.com/memorial/85027064

2 Critical Questions for Big Data

Critical Questions for Big Data
Danah Boyd & Kate Crawford. *Information, Communication & Society*, Vol 15, No 5, June 2012, pp.662–679.

Abstract
"...In this article, we offer six provocations to spark conversations about the issues of Big Data: a cultural, technological, and scholarly phenomenon that rests on the interplay of technology, analysis, and mythology that provokes extensive utopian and dystopian rhetoric."

1. Big Data changes the definition of knowledge.
2. Claims to objectivity and accuracy are misleading.
3. Bigger data are not always better data.
4. Taken out of context, Big Data loses its meaning.
5. Just because it is accessible does not make it ethical.
6. Limited access to Big Data creates new digital divides.

1. Big Data changes the definition of knowledge

- Big Data not only refers to the data and how to manage it, but also to a computational turn in thought and research.
- Big Data creates a radical shift in how we think about research.
- The specialized tools of Big Data have their own inbuilt limitations and restrictions.

Example: Twitter, Facebook

- Twitter and Facebook are examples of Big Data sources that offer very poor archiving and search functions.
- Consequently, researchers are much more likely to focus on something in the present or immediate past.

2. Claims to objectivity and accuracy are misleading

- Big Data offers (the humanistic disciplines) a new way to claim the status of quantitative science and objective method.
- However, the interpretation of data is subjective.
- The data cleaning process may also be subjective.
  - Making decisions about what attributes and variables to keep.
- The process of gathering data may introduce bias.
- Big Data enables the practice of apophenia.
  - Seeing patterns where none actually exist.

Example
Leinweber (2007) *http://joi.iijournals.com/content/16/1/15* observed a strong correlation between the changes in the S&P 500 stock index and butter production in Bangladesh.

Leinweber, D. (2007). Stupid data miner tricks: overfitting the S&P 500, *The Journal of Investing*, 16(1):15–22.

3. Bigger data are not always better data

- Just because we have larger quantities of data does not mean that methodological issues are no longer relevant.
- How the data is sampled is still important.

Example: Twitter

- Twitter users are a particular sub-set of "all people".
- A Twitter user does not necessarily represent a "person".
  - Corporate accounts.
  - People with multiple Twitter accounts.

  - Bots.

- Some frequent Twitter users are "listeners" who never post.
- Twitter removes posts with particular topics, e.g. porn.
- Twitter makes only a fraction of the data public.
- Tweets from protected accounts are excluded.

4. Taken out of context, Big Data loses its meaning

- Social graphs collected through social media may be different from those produced through traditional methods (surveys, interviews, etc).
- Two types of social networks derived from social media:

  Articulated networks built by users through their "friends" and "follower" lists.

  Behavioural networks built by examining behaviour patterns such as replies, tagged in the same image, etc.

- These are different from traditional personal networks.

5. Just because it is accessible does not make it ethical

- It is possible to *de-anonymise* parts of anonymised data sets.
- Subjects post without knowing that their posts may be used by people that are not their intended audience.
- Subjects might not even know that their posts are used for research.
- In order to act ethically, researchers should reflect on the importance of accountability.
  - "Accountability requires rigorous thinking about the ramifications of Big Data, rather than assuming that ethics boards will necessarily do the work of ensuring that people are protected."
- There is a difference between "being in public" and "being public" (i.e. actively seeking attention).

6. Limited access to Big Data creates new digital divides

- Only some companies have access to really large social data.
- Those with money — or those inside these companies — can produce a different type of research than those outside.
- Those without access can neither reproduce nor evaluate the methodological claims of those who have privileged access.
- Some computational skills are needed to access, prepare and analyse the data.
- Most with computational skills are male and these are those who will ask the questions.
  - There is gender-based bias in the kinds of questions being asked.
- Questions difficult for companies who control access to the data may never be asked.

3 Case Studies

3.1 The Netflix Challenge

The Netflix Challenge

- Netflix established the Netflix Prize for the best collaborative filtering algorithm to predict user ratings for films.
  - *https://www.netflixprize.com/*
  - $1 million was awarded to the winner on September 2009.
- Netflix released a large collection of training data:
  - Over 100 million ratings.
  - Nearly 480 thousand users.
  - Over 17 thousand movies.
- There was no user or information about the films, only anonymised IDs.
- But researchers were able to *de-anonymise* (identify) individual users from the data set.
- Several Netflix users filed a class action lawsuit against Netflix for releasing the data.

De-anonymising users from the Netflix training data

- The NetFlix training data did not contain information about the users or the movies.
- But every user and movie were identified with a unique ID.
- So it is possible to determine patterns and connections between users and movies.
- Researchers cross-referenced this information with information from the Internet Movie Database (IMDB).

Narayaman & Shmatikov, 2008 https://arxiv.org/abs/cs/0610105

"Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information."

3.2 Cambridge Analytica

Cambridge Analytica

- Cambridge Analytica (CA) was a British political consulting firm.
- They exploited a loophole in Facebook's mechanism to allow apps to access personal data of Facebook's users.
- CA developed and app that eventually extracted personal data from 87 million users.
- They allegedly used this information to influence electoral results.
- Results of several elections have been said to have been affected by CA's involvement, including:
  - 2016 Brexit referendum.
  - 2016 US presidential elections.

3.3 Bias in Data and Racist AI

Bias in Data

- Machine learning predictors are trained to minimise the prediction error in training data.
- If the data show bias, this bias is transferred to the predictors.
- Examples of AI applications that have been shown to be biased:
  - Image searches for particular terms
    https://www.independent.co.uk/life-style/gadgets-and-tech/news/bing-image-search-microsoft-jews-racist-hitler-nazis-a8579596.html
  - Chat bots
    https://www.independent.co.uk/life-style/gadgets-and-tech/news/tay-tweets-microsoft-ai-chatbot-posts-racist-messages-about-loving-hitler-and-hating-jews-a6949926.html

Take-home Messages

- Critical questions for Big Data.
- Case studies.

What's Next

Week 4

- Indexing Big Data — R-Tree
  - A central topic for assignment 2
- Submit assignment 1 by Sunday 25 August
- Remember academic honesty
  - Submit your own work
  - Don't copy from others
  - Don't collaborate