COMP20008
Elements of Data Processing

Semester 1 2020

Lecture 21: Social and Ethical Implications of Big Data Analytics

Plan for today

- Big data analytics
  - Issues
  - Stakeholders
  - Processes & implications
  - 10 simple rules for responsible and ethical big data research

- Acknowledgements
  - Many of these slides have been adapted from materials developed by Ida Asadi Someh

Question- what is the story?

	Top 3 silicon valley (2014)	Top 3 auto (car) makers (1990)
Revenue	$247 billion	$250 billion
Number of employees	137000	1.2 million employee
Market capitalisation	$1.09 trillion	$36 billion

(Zuboff 2015)

What is Big Data Analytics? What's different now?

- The ability to collect, store, and process increasingly large and complex data sets from a variety of sources, into competitive advantage (LaValle and Lesser 2013)

- Big data management capabilities
  - Volume, Variety and Velocity (3Vs) + Veracity (4Vs) + Value (5Vs)
- Algorithms to process big data
  - Advanced statistical and computational techniques to process large, unstructured and fast data

how we able to use data and achieve value

the amount of data

lots of unstructured data (various form)

speed of data coming in

how reliable the data to be is representative of the population?

Is this a sufficient definition?

Creepy uses of data in the media

- Target exposing a teen girl's pregnancy
  - *Father: "My daughter got this in the mail!" he said. "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?"*

https://www.businessinsider.com.au/the-incredible-story-of-how-target-exposed-a-teen-girls-pregnancy-2012-2
https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&_r=1&hp

- Facebook's 2012 secret mood experiment
  - how people react to an emotional contagion process
  - filtered users' news feeds – the flow of comments, videos, pictures and web links posted by other people
  - 689,000 users affected
  - https://www.theguardian.com/technology/2014/jun/29/facebook-users-emotions-news-feeds

- Cambridge Analytica (we have discussed previously ..)

Consequences of Big Data Analytics

- Positive consequences
  - Tracking criminals, higher product margins, new business models, improved healthcare and …
- Negative consequences:
  - Misuse of personal information, breaching privacy, profiling of individuals, discrimination and …
- Where is the boundary? What is the balance?
  - There is no agreement on what is ethical and what is not!

eg big data used to predict someone's possibility to commit a crime

bad → discrimination, racism

↳ good accurate predict

(Markus and Topi 2015; Newell and Marabelli 2015)

BDA: Not only about Technology

- Technological view (3Vs) does not help to understand unethical use
  - Technology is neutral in nature
  - It does not consider the underlying processes that are enabled
  - It does not consider the stakeholders that are involved and influenced
  - 3Vs do not consider either people or process
  - A new social phenomenon, a new market economy

- We need further (non technical) perspectives on big data analytics
  - It is not OK to exclude social from the definition

Stakeholder View on BDA

The exchange of data is characterised by its 4Vs.

Unequal exchanges between stakeholders

individual has less power

organization has lots of power

BDA from Social Perspective

- Interactions among stakeholders

- Data is contributed, collected, extracted, exchanged, sold, shared, and processed for the purpose of predicting and modifying human behaviour in the production of economic or social value.

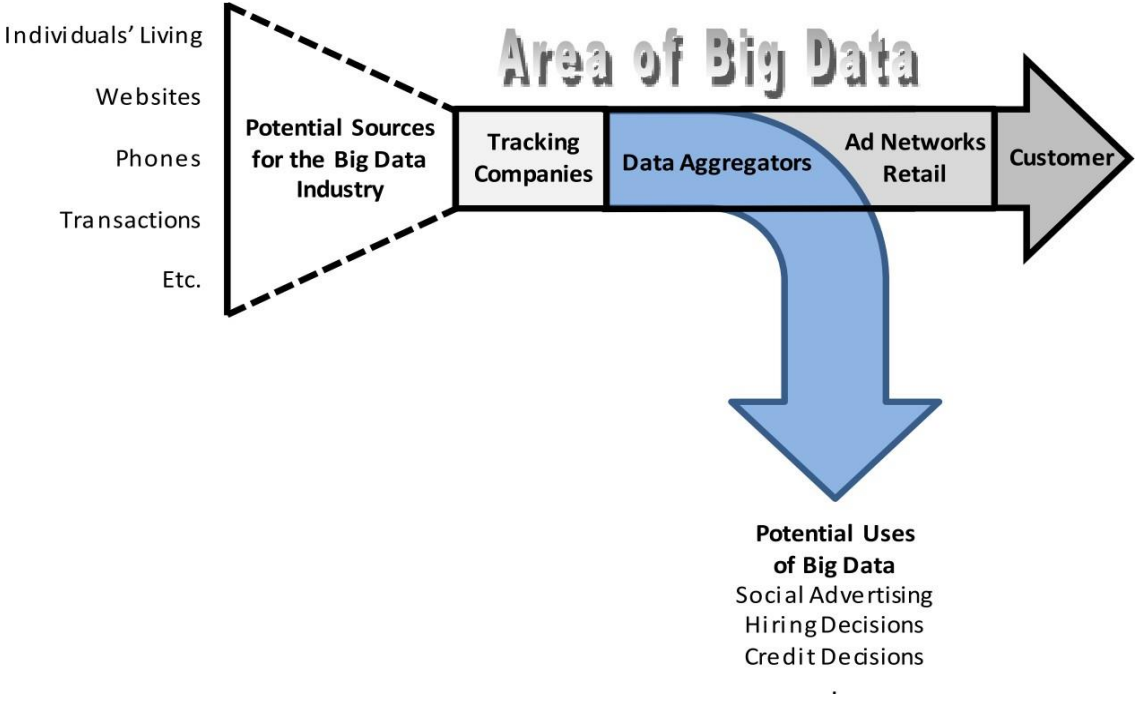
- BDA involves several processes, discussed next

Process 1: Data Extraction

- Data *extraction*, not data *collection* → pull data out rather than simply collect it
  - Google Streetview ("single greatest breach in the history of privacy")
    - https://www.theguardian.com/technology/2010/may/15/google-admits-storing-private-data
- Our everydayness quantified
- Incursions into legally and socially undefended territory
- Google has the largest unpaid number of employees
  - "You're not the customer, you're the product .."

Process 2: Data commodification: secondary markets and hidden value chains

- Sell personal data until it turns into waste
- Big data as a new industry (secondary markets)



(Martin 2015)

Examples

- http://intelligence.towerdata.com/
- http://www.iriworldwide.com/en-us/
- https://www.intelius.com/

Process 3: Decision Making

- Big Data Quality (Veracity)
  - Data accuracy for aggregated data
    - What's the quality criteria for a social media post?
  - Completeness of our digital identity
    - Mosaic effect
    - "When is a boat not a boat?"
    - Game: http://celebrityguesswho.com/#2
  - Meaning dependent on the context
    - Is how I act on social media a true representation of who I am?

Process 3: Decision Making

- Data Analysis
  - Predictions based on the past
  - How can I redefine myself?
  - In what context is is legitimate to make a prediction about someone?
  - Predictions often based on correlations (not causations)
  - What about outliers? (what if I don't fit into a predefined category??)

- Data Visualization
  - Decision making and presentation biases

Example 1

- https://www.theguardian.com/technology/2016/nov/02/admiral-to-price-car-insurance-based-on-facebook-posts

Admiral to price car insurance based on Facebook posts

Insurer's algorithm analyses social media usage to identify safe drivers in unprecedented use of customer data

The insurer will examine posts and likes by the Facebook user, although not photos, looking for habits that research shows are linked to these traits. These include writing in short concrete sentences, using lists, and arranging to meet friends at a set time and place, rather than just "tonight".

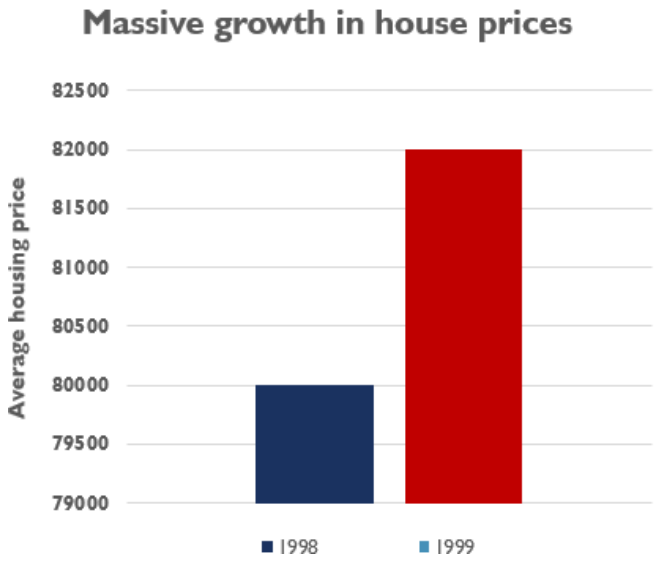
In contrast, evidence that the Facebook user might be overconfident - such as the use of exclamation marks and the frequent use of "always" or "never" rather than "maybe" - will count against them.

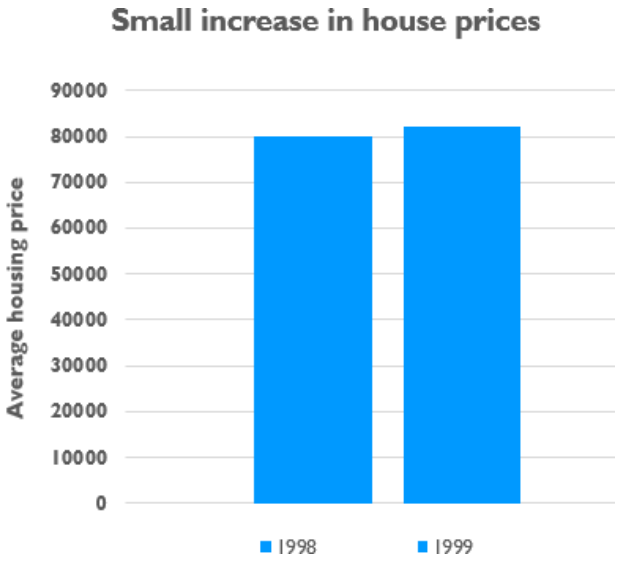
Process 3: Decision Making

- Data Analysis
  - Predictions based on the past
  - How can I redefine myself?
  - In what context is is legitimate to make a prediction about someone?
  - Predictions often based on correlations (not causations)
  - What about outliers? (what if I don't fit into a predefined category??)

- Data Visualization
  - Decision making and presentation biases

Example 2

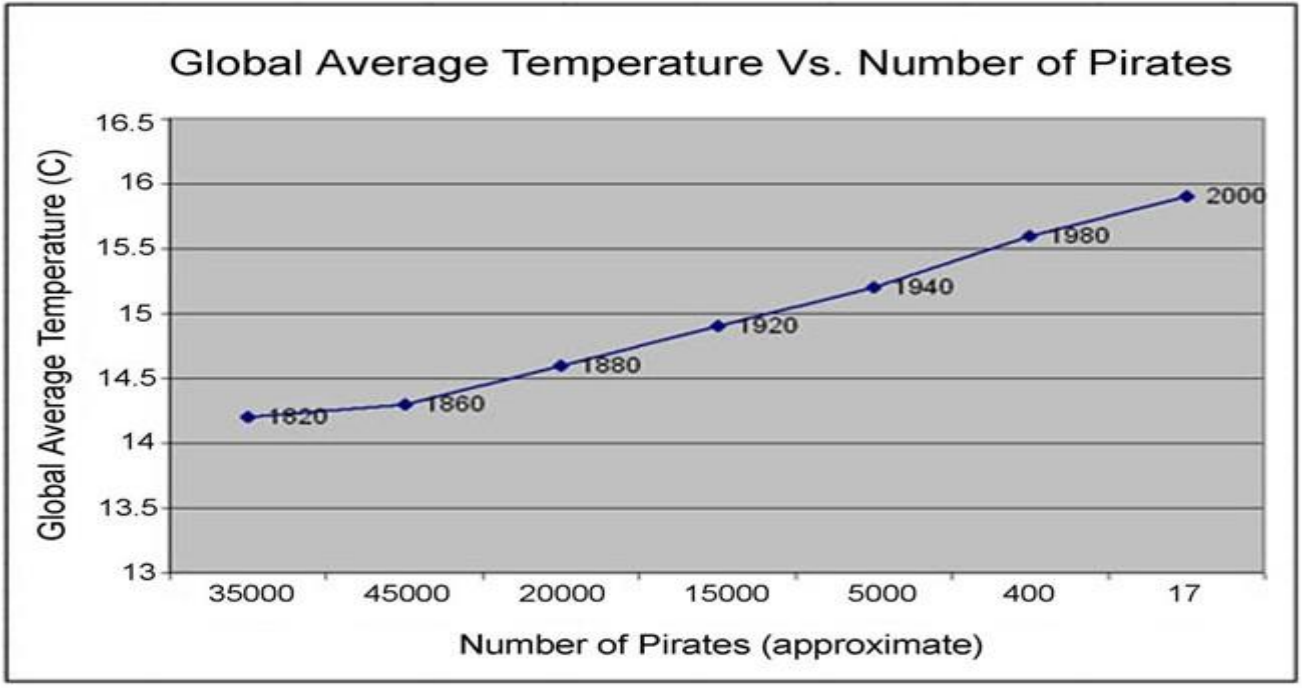

Massive growth in house prices
Average housing price
82500
82000
81500
81000
80500
80000
79500
79000
1998
1999


Small increase in house prices
Average housing price
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
1998
1999

Example 3

- Do you want to stop global warming? **BECOME A PIRATE**!

correlation ≠ causation



See also: https://www.tylervigen.com/spurious-correlations

Process 4: Control and monitoring

- Pervasive monitoring now possible using sensors, Internet of Things technology
  - Everyone is observed, organisations make money of observing others, collect data, sell data, make offers, induce dependence
- What happens to social trust?
- Surveillance is the precise opposite of the trust-based relationships
- Free market economy versus Surveillance Economy
  - Who I am, what I am, what I like to do, where I like to go, who I know, …
  - http://theconversation.com/someones-looking-at-you-welcome-to-the-surveillance-economy-16357

(Zuboff 2015 and Martin 2015)

Process 5: Experiments

- Where did experiments traditionally take place?
  - We have been seeing big companies running a large number of experiments on its users investigating different aspects

- Rewards and punishments

while for big data,
take experiment online
oversight,

(Zuboff 2015)

Societal actors need to provide oversight and regulate

- EU General Data Protection Regulation (enforced since May 2018)

- Aims to regulate and protect data privacy for all EU citizens.
  - Penalty 4% of annual global turnover of the organizations.

- The consent
  - should be clear, concise, not too long and intelligibly written—should attach the reasons of data collection and analyses.
  - individuals have the right to withdraw the consent with the same easiness that they have previously agreed with.

- Accessing individual's data
  - Individuals have the right to ask for a copy of their personal data together with information regarding the processing and purpose of data collection and analyses from a controller
  - Individuals have the right of data portability, which means that they can transfer their data from one controller to another.

http://www.eugdpr.org/eugdpr.org.html

10 simple rules for responsible big data research

- From Zook et al (Plos Comp Bio 2017)
  - http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005399

Rule 1

- Acknowledge that data are people and can do harm
  - All data are people until proven otherwise
    - Social media
    - Heart rates from Youtube videos
    - Ocean measurements that change property risk profiles

Rule 2

- Recognize that privacy is more than a binary value
  - Privacy is contextual and situational
  - Single Instagram photo versus entire history of social media posts
  - Privacy preferences differ across individuals and societies

Rule 3

- Guard against the reidentification of your data
  - Metadata associated with photos
  - Reverse image search – connect dating and professional profiles
  - Difficult to recognize the vulnerable points a-priori!
    - Battery usage on a phone – can reveal a person's location
      - Unintended consequence of 3rd party access to phone sensors
  - When datasets thought to be anonymized are combined with other variables, it may result in unexpected reidentification

Rule 4

- Practice ethical data sharing
  - Seeking consent from participants to share data

Rule 5

- Consider the strengths and limitations of your data; big does not automatically mean better
  - Document the provenance and evolution of your data. Do not overstate clarity; acknowledge messiness and multiple meanings.
    - is a Facebook post or an Instagram photo best interpreted as an approval/disapproval of a phenomenon, a simple observation, or an effort to improve status within a friend network?

Rule 6

- Debate the tough, ethical choices/issues
  - importance of debating the issues within groups of peers
    - Examples mentioned earlier
      - Facebook emotional contagion
      - Exposing teen girl's pregnancy
    - More recently, Google Duplex
      - https://www.youtube.com/watch?v=D5VN56jQMWM

Rule 7

- Develop a code of conduct for your organization, research community, or industry
  - Are we abiding by the terms of service or users' expectations?
  - Does the general public consider our research "creepy"?

how to use information?

ensure privacy

Rule 8

- Design your data and systems for auditability
  - Plan for and welcome audits of your big data practices.
  - Systems of auditability clarify how different datasets (and the subsequent analysis) differ from each other, aiding understanding and creating better research.
    - "For example, many types of social media and other trace data are unstructured, and answers to even basic questions such as network links depend on the steps taken to collect and collate data."

you need to make sure that you are able to explain about the process and how to use

Rule 9

- Engage with the broader consequences of data and analysis practices
  - Recognize that doing big data research has societal-wide effects

Rule 10

- Know when to break these rules
  - Natural disaster
  - Public health emergency
  - Hostile enemy
  - …

  - It may be important to temporarily put aside questions of individual privacy in order to serve a larger public good.

Summary: What can we do?

- We need to empower individuals
  - Educate individuals, raise social awareness
  - Provide data access and control (e.g. Google activity)
    - http://www.abc.net.au/4corners/stories/2017/04/10/4649443.htm
- Define and develop a culture of acceptable data use
  - Organizations should internalize the costs
  - Genuine consent from individuals
  - Be transparent and clearly communicate intent of data collection and analytics/ai
  - Adopt rules for responsible big data research
  - Provide data control to individuals