Techniques for Large-scale Data

2019-2020, GK-7

Big Data Ethics

Big Data Ethics

One of the course's learning outcomes:

- discuss large-scale data processing from an ethical point of view.

“Justice: What's The Right Thing To Do?”

This slide contains supplementary material that is outside the scope of this course.

If you want a moral philosophy perspective on ethical decision making:

- Professor Michael Sandel’s lectures from a course at Harvard introducing moral and political philosophy
- https://www.youtube.com/watch?v=kBdfcR-8hEY&list=PL30C13C91CFFEFEA6
- First lecture includes some “trolley car” scenarios

Different ethical frameworks exist

This slide contains supplementary material that is outside the scope of this course.

- Different ethical frameworks can lead to different recommendations
- e.g. Kantian ethics vs. Utilitarianism
- https://www.youtube.com/watch?v=7FR-FuhN2HM

Ten simple rules for responsible big data research

- Matthew Zook, Solon Barocas, danah boyd, Kate Crawford, Emily Keller, Seeta Peña Gangadharan, Alyssa Goodman, Rachelle Hollander, Barbara A. Koenig, Jacob Metcalf, Arvind Narayanan, Alondra Nelson, Frank Pasquale
- PLOS Computational Biology 13(3): e1005399 (March 2017). https://doi.org/10.1371/journal.pcbi.1005399
- ”we structure the first five rules around how to reduce the chance of harm resulting from big data research practices; the second five rules focus on ways researchers can contribute to building best practices that fit their disciplinary and methodological approaches.”

Ten simple rules for responsible big data research

1. Acknowledge that data are people and can do harm

- “most data represent or impact people”
- “even seemingly benign data can contain sensitive and private information”
- ”Even data that seemingly have nothing to do with people might impact individuals’ lives in unexpected ways”

Ten simple rules for responsible big data research

2. Recognize that privacy is more than a binary value

- ”creepy”: when social values and technical capabilities are not aligned
- ”Understand that your attitude towards acceptable use and privacy may not correspond with those whose data you are using, as privacy preferences differ across and within societies.”

Ten simple rules for responsible big data research

3. Guard against the reidentification of your data

- Well-intended attempts to anonymise data have often turned out to be insufficient.

Ten simple rules for responsible big data research

4. Practice ethical data sharing

- In some projects ”questions of the larger public good … provide compelling arguments for sharing data”
- ”even when broad consent was obtained upfront, researchers should consider the best interests of the human participant, proactively considering the likelihood of privacy breaches and reidentification issues”

Ten simple rules for responsible big data research

5. Consider the strengths and limitations of your data; big does not automatically mean better

- Context is important in all steps of the big data value chain
- “understand both the source of the data and the rules and regulations with which they were gathered”

Ten simple rules for responsible big data research

6. Debate the tough, ethical choices

- ”Discussion and debate of ethical issues is an essential part of professional development—both within and between disciplines—as it can establish a mature community of responsible practitioners.”
- ” Why might one set of scholars see [some ethical case] as a relatively benign approach while other groups see significant ethical shortcomings? Where do researchers differ in drawing the line between responsible and irresponsible research and why?”

Ten simple rules for responsible big data research

7. Develop a code of conduct for your organization, research community, or industry

- “Public attention to the ethical use of these data should not be avoided; after all, these datasets are based on an infrastructure that billions of people are using to live their lives, and there is a compelling public interest that research is done responsibly.”
- ” Issues that might otherwise be ignored until they blow up — e.g., Are we abiding by the terms of service or users’ expectations? Does the general public consider our research “creepy”? — can be addressed thoughtfully rather than in a scramble for damage control. This is particularly relevant to public-facing private businesses interested in avoiding potentially unfavorable attention.”

Ten simple rules for responsible big data research

8. Design your data and systems for auditability

- “The goal of auditability is to clearly document when decisions are made and, if necessary, backtrack to an earlier dataset and address the issue at the root (e.g., if strategies for anonymizing data are compromised).”
- ”Designing for auditability also brings direct benefits to researchers by providing a mechanism for double-checking work and forcing oneself to be explicit about decisions, increasing understandability and replicability.”

Ten simple rules for responsible big data research

9. Engage with the broader consequences of data and analysis practices

- e.g. consider environmental impact energy requirements of big data analytics
- ”While all big data research need not take up social and cultural questions, a fundamental aim of research goes beyond understanding the world to considering ways to improve it.”

UN sustainable development goals

Ten simple rules for responsible big data research

10. Know when to break these rules

- “For example, in times of natural disaster or a public health emergency, it may be important to temporarily put aside questions of individual privacy in order to serve a larger public good. Likewise, the use of genetic or other biological data collected without informed consent might be vital in managing an emerging disease epidemic.”
- ”Ethics is often about finding a good or better, but not perfect, answer, and it is important to ask (and try to answer) the challenging questions.”

Discussing ethical issues

- **Identify stakeholders**
- **Identify benefits and possible harm for each stakeholder**
- **Weigh benefits against possible harm**

- Consider actions that can mitigate harm
- Get input from others

Openness regarding ethical assessment

- It is assumed that organisations will normally want to operate in an ethical manner and some might even consider it beneficial to be seen as having high ethical standards

- Would you want an internal ethical assessment to be seen outside the organisation (e.g. by customers or the media)?

- Would you publish an internal ethical assessment on the organisation's web site?
- With open source software there's a view that "given enough eyeballs, all bugs are shallow" (Raymond, 1999). Could an organization crowdsource a better ethical policy?

Raymond, E. (1999). *The cathedral and the bazaar*. Knowledge, Technology & Policy, 12(3), 23-49.

Determining user personality characteristics from social networking system communications and characteristics

- ”A social networking system obtains linguistic data from a user's text communications on the social networking system. For example, occurrences of words in various types of communications by the user in the social networking system are determined. The linguistic data and non-linguistic data associated with the user are used in a trained model to predict one or more personality characteristics for the user. The inferred personality characteristics are stored in connection with the user's profile, and may be used for targeting, ranking, selecting versions of products, and various other purposes.”

US patent US8825764B2

Emotion recognition in video conferencing

- ”… when faces or other parts of video conference participants are detected, the present technology determines an emotional status of video conference participants. This may include identification of facial expressions or changes in facial expressions over time. The emotional status can be also partly based on speech recognition or voice analysis. If it is determined that the emotional status is negative, an alert communication can be generated and transmitted to one of a video conference participant or a third party.”

US patent US9576190B2

Car insurance prices based on Facebook posts

- “Admiral Insurance will analyse the Facebook accounts of first-time car owners to look for personality traits that are linked to safe driving. For example, individuals who are identified as conscientious and well-organised will score well.”
- “These [traits] 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.”
- “The scheme is voluntary, and will only offer discounts rather than price increases”

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

Tracking staff behaviour

- A company is considering using a system that will allow management to track staff behaviour in real-time, gather data on when and who they email, who accesses and edits files, and who meets whom when. The HR (human resources, or personnel) department at the company is in favour of introducing this new system and believes it will benefit the staff at the company. The company's management is also in favour.
- Discuss whether introducing this system raises potential ethical issues.

Feet, not faces?

- ”market leading technology provides 95% footfall accuracy and is 80% accurate at identifying gender. The technology is unobtrusive and sensors can be easily installed throughout and outside a store. Cameras point to the ground, so intelligence can be gathered while preserving people's privacy. The proprietary technology can track outside traffic, in-store occupancy, dwell times, group size, demographic details and brand recognition to provide real-time data with the potential for rich analysis.”

[Hoxton Analytics web site, 2018]

Only a source of ethical problems?

- Techniques for large-scale data can offer tools for investigating possible ethical issues

- Can help to shine a light on possible unethical behaviour in other areas

Bank fraud detection

- Graph databases can be used to uncover fraud rinds in real time
- Example: finding a group of people who share an address (not suspicious), or a phone number (possible), or a social security number (a red flag!).
- This example is described in a “graph gist”:
  https://neo4j.com/graphgist/bank-fraud-detection
- and in a short video:
  https://www.youtube.com/watch?v=ujimD6MP87I
- and in a longer video [13:48-18:33]:
  https://www.youtube.com/watch?v=AeNufTq1W5I

Some other fraud detection examples

- The longer video mentioned on the previous slide describes three kinds of fraud.
- For each kind, the presenters explain how the fraud is carried out, and then they describe how the fraud situation can appear as a pattern in a graph database:

  - First party fraud [13:48-18:33]
  - Insurance fraud [26:45-27:48]
  - eCommerce fraud [27:50-29:15]

Global Tax Evasion

The Neo4j graph database was used to investigate the Panama Papers

https://neo4j.com/case-studies/the-international-consortium-of-investigative-journalists-icij/

https://neo4j.com/blog/analyzing-panama-papers-neo4j/

- https://www.youtube.com/watch?v=8VJ4njrA0TQ
- Talk starts at [2:42]
- Neo4j from about [14:00], including an overview of Neo4j features
- Panama papers feature from [34:00]
- Demo using Neo4j browser from [39:34]

2017 Datathon winner

- Revealing hidden links through open data
  - https://www.datasciencesociety.net/datathon-winner-revealing-hidden-links-through-open-data/

- Using RDF, GraphDB and SPARQL

- If someone managing a government entity is also a related party (such as, for example, owner) of a private contractor of the government entity, there might be a conflict of interest.
- Cliques of companies that share board members, and individuals who are on the boards of many dozens of companies: might indicate tax evasion.

General Data Protection Regulation (GDPR)

REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 27 April 2016

on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

- https://eur-lex.europa.eu/eli/reg/2016/679/oj
- 88 pages

Principles relating to processing of personal data (Article 5, GDPR)

Personal data shall be:

a) processed lawfully, fairly and in a transparent manner …
b) collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes …
c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed …
d) accurate and … up to date
e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed …
f) processed in a manner that ensures appropriate security of the personal data …

Example of exception

- “(112) Those derogations [exceptions] should in particular apply to data transfers required and necessary for important reasons of public interest, for example in cases of international data exchange between competition authorities, tax or customs administrations, between financial supervisory authorities, between services competent for social security matters, or for public health, for example in the case of contact tracing for contagious diseases …”

Illegal vs. Unethical

- *Illegal* and *unethical* are not the same thing.
- While we might want there to be good agreement between a society's ethical values and its laws, it should be anticipated that the fast pace at which technology is developing can lead to technical possibilities and the law becoming out of step.
- Contribute to discussions about technology and the law – it is important for those who understand technology to participate in this debate, and not to leave the platform to those who lack your insights into the technologies being legislated on.