Data Science Ethics

Peter.Flach@bristol.ac.uk

Peter A. Flach

www.cs.bris.ac.uk/~flach/

Department of Computer Science

Intelligent Systems Laboratory

University of Bristol

Can computers be racist?

http://www.fordfoundation.org/ideas/equals-change-blog/posts/can-computers-be-racist-big-data-inequality-and-discrimination/

What is ethics?

- What actions are right or wrong in particular circumstances?
- Philosophers have proposed different ways to formalise this; here we will take a practical, utilitarian view:

Ethical behaviour puts benefits to group or society above benefits to individual

  - E.g. queuing at supermarket, letting car in side road pull out before you
  - Based on shared values within the group or society
- "What one SHOULD do" rather than "what one CAN do" (legally)
  - Laws often follow ethics but not all ethics is regulated by law
- Slides in large part inspired by the edX course DS101x Data Science Ethics
  - https://courses.edx.org/courses/course-v1:MichiganX+DS101x+1T2017/course/

Data science and ethics: some key topics

- Human subjects research and informed consent
- Data ownership
- Privacy and anonymity
- Data validity, managing change
- Algorithmic fairness

Informed consent

- For consent to be valid it must be informed consent. For this to be the case it must be:
  - Given voluntarily (with no coercion or deceit)
  - Given by an individual who has capacity
  - Given by an individual who has been fully informed about the issue.
- http://ministryofethics.co.uk/?p=6#
- At UK universities this is overseen by Research Ethics Committees
  - http://www.bristol.ac.uk/red/research-governance/ethics/uni-ethics/

Informed consent: exceptions

- Not legally required in case of ordinary conduct of business
  - e.g. A/B testing by web companies
    - happens all the time!
  - e.g. 2012 Facebook/Cornell experiment
    - http://www.pnas.org/content/111/24/8788.full
    - https://www.theguardian.com/technology/2014/jun/30/facebook-emotion-study-breached-ethical-guidelines-researchers-say
  - e.g. OKCupid “We experiment on human beings”
    - https://theblog.okcupid.com/we-experiment-on-human-beings-5dd9fe280cd5
    - https://www.theguardian.com/technology/2014/jul/29/okcupid-experiment-human-beings-dating

Informed consent: limitations

- Consent may be given for one particular use of the data, but doesn't automatically extend to retrospective analysis or repurposing
  - hence there is a difference between recording and use
    - e.g. most people are used to CCTV in shops, but there is a shared expectation that recorded video will not be published.
    - e.g. mobile phone companies need to track you in order to provide their service, but there is a shared expectation that your whereabouts won't be used or shared.
  - Often these limitations are voluntary and non-contractual, but there is a considerable grey area and lack of societal consensus (e.g. government surveillance).

- Is consent informed if it is hidden in many pages of dense legalese?

Data ownership: one view

- The writer of a biography owns it, not the subject.
- Wikipedia owns the encyclopaedia, not the contributors.
- TripAdvisor owns the reviews, not the contributors.
- **The collector of personal data owns it, not the subject.**

Data ownership: the EU view

- The **general data protection regulation** will apply from 25 May 2018. It lists the rights of the **data subject**, that is the individual whose personal data is being processed. These strengthened rights give individuals more control over their personal data, including through:
  - the need for the individual's clear **consent** to the processing of personal data
  - easier **access** by the subject to his or her personal data
  - the rights to **rectification**, to erasure and ‘to be forgotten’
  - the right to **object**, including to the use of personal data for the purposes of ‘profiling’
  - the right to data **portability** from one service provider to another
- It also lays down the obligation for **controllers** (those who are responsible for the processing of data) to provide transparent and easily accessible information to data subjects on the processing of their data.
- http://www.consilium.europa.eu/en/policies/data-protection-reform/data-protection-regulation/

Privacy

- Privacy is a basic human need, unrelated to whether you have anything to hide or not. Nevertheless, it is hard to define ("the right to be left alone").
- In general, privacy has individual and societal benefits (e.g. voting).
- Privacy is sometimes related to non-disclosure, but more often about being able to control disclosure.
- A distinction is often made between data (the actual phone call) and metadata (the number you called, where you and they were at the time, duration of the call, etc.).
  - But metadata can reveal a lot!
    - e.g. car insurers tracking driver behaviour
  - Similarly, disaggregated data can reveal a lot!
    - e.g. 'smart' water, electricity and gas meters can reveal who is at home and what they are doing

From trust to design

- Traditional social norms dealt with privacy by trust
- Modern data systems must deal with privacy by design
  - data sharing must be contractual
  - many stakeholders with different interests, but no societal consensus yet

- Absolute anonymity is probably impossible, but at least we can avoid casual identification by properly de-identifying the data.
  - "You have zero privacy anyway. Get over it." (Scott McNealy, Sun CEO, 1999)

Data validity

- Bad data and bad models can lead to bad (possibly harmful) decisions
- Many possible sources of error
  - choice of representative sample
    - e.g. are Twitter users representative of the population? Are tweets representative of Twitter users?
    - may need to rebalance important attributes (e.g. gender, race)
    - drift means that data that once was representative may no longer be
  - errors in the data
    - e.g. 26% of consumers had at least one error in their credit report; 29% of consumers had credit scores that differ by at least fifty points between credit bureaus.
    - "In the Heisenberg-meets-Kafka world of credit scoring, merely trying to figure out possible effects on one's score can reduce it." (Frank Pasquale, The Black Box Society, 2015)
      https://books.google.co.uk/books?id=TumaBQAAQBAJ&pg=PA24&lpg=PA24&dq=kafka+meets+heisenberg

Gender bias at UC Berkeley?

	Men applied	Men admitted	%	Women applied	Women admitted	%
	2691	1198	45%	1835	557	30%

http://vudlab.com/simpsons/

Gender bias at UC Berkeley?

	Men applied	Men admitted	%	Women applied	Women admitted	%
	2691	**1198**	**45%**	**1835**	**557**	**30%**
A	825	512	62%	108	89	**82%**
B	560	353	63%	25	17	**68%**
C	325	120	**37%**	593	202	34%
D	417	138	33%	375	131	**35%**
E	191	53	**28%**	393	94	24%
F	373	22	6%	341	24	**7%**

http://vudlab.com/simpsons/

Managing change

- Campbell's Law: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."
- Goodhart's Law: "When a measure becomes a target, it ceases to be a good measure."
  - This also happens in data science → be critical of bake-offs!
- E.g. Google Flu Trends
  - launched in 2008 to detect flu outbreaks from Google search queries
  - started performing poorly in 2013, to a large extent caused by people changing their search behaviour (and by overfitting to seasonal search terms that stopped being correlated with flu occurrences)
  - https://youtu.be/e60sEYNikPk

Algorithmic fairness

- Big data can be used to facilitate proxy discrimination by means of non-protected attributes (e.g. postcode) that correlate strongly with protected attributes (e.g. race)
  - but also to detect and address this!
- The following examples are taken from the KDD 2016 tutorial on Algorithmic bias: from discrimination discovery to fairness-aware data mining
  - http://francescobonchi.com/algorithmic_bias_tutorial.html

UCI datasets can be biased!

- German credit (https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data))
  - N = 1,000 records of bank account holders
  - Class label: good/bad creditor (grant or deny a loan)
  - Attributes: numeric/interval-scaled: duration of loan, amount requested, number of installments, age of requester, existing credits, number of dependents; nominal: result of past credits, purpose of credit, personal status, other parties, residence since, property magnitude, housing, job, other payment plans, own telephone, foreign worker; ordinal: checking status, saving status, employment

Examples of discrimination

- PD rules A & B → C
  - A is potentially discriminated (PD) group, B is context, C denies benefit
  - e.g. gender=female & saving_status=no_known_savings → credit=no
- Favouritist PD rules A & B → C
  - A is favoured group, B is context, C grants benefit
  - e.g. gender=male & saving_status=no_known_savings → credit=yes
- Indirect discrimination
  - suppose neighbourhood=10451 & city=NYC → benefit=deny
  - and neighbourhood=10451 & city=NYC → ethnicity=african_american
  - then neighbourhood=10451 & city=NYC & ethnicity=african_american → benefit=deny

Measuring discrimination

- lift(A→C) = conf(A→C)/conf(true→C)
  - conf(X→Y) = support(X&Y)/support(X)
- elift_B(A&B→C) = conf(A&B→C)/conf(B→C) = conf(B&C→A)/conf(B→A)
- We want the elift of PD-rules to be less than some threshold α
  - e.g. conf(city=NYC → benefit=deny) = 0.25
  - conf(city=NYC & ethnicity=african_american → benefit=deny) = 0.75
  - hence elift = 3.0

German credit examples

- conf(saving_status=no_known_savings → credit=no) = 0.18
- conf(personal_status=female_div/sep/mar & saving_status=no_known_savings → credit=no) = 0.27
- elift = 1.5
- conf(purpose=used_car → credit=no) = 0.17
- conf(age≥52.6 & personal_status=female_div/sep/mar & purpose=used_car → credit=no) = 1
- elift = 6.0

Outlook

- Data science is pervasive
- Only small part of ethical issues will be regulated
- We yet have to reach societal consensus about many of them
- The Data Scientist has a large responsibility here