Responsible Data Management

Not in your textbook

Outline

- Privacy
- Equity

Data Value

- Data sets are bought and sold every day.
- Value is in the organization of data.
- Increase value by doing work:
  - Data cleaning
  - Data integration
  - More convenient access tools
  - …
- Very poor theory on how to price.

Surveillance Capitalism

- Do we really get stuff for free from companies?
- How do TV networks make money?
- How do Google, Facebook, etc. make money?
- How do these differ?

- Shoshana Zuboff

Privacy

- Ability to control sharing of information about self.
- Basic human need.
    - Even for people who have “nothing to hide”

Loss of Privacy

- Due to loss of control over personal data.
- I am OK with you having certain data about me that I have chosen to share with you or that is public, but I really do not want you to share my data in ways that I do not approve.

Anonymity

NETFLIX PRIZE

Closeted Lesbian Sues Netflix For Potential Outing

By Laura Northrup on December 19, 2009 3:00 PM

Here's the problem with anonymized data: if it were truly anonymized, it wouldn't be useful to anyone for anything. With enough data about a person—say, their age, gender, and zip code—it's not hard to narrow down who someone is. That's the idea behind a class-action lawsuit against Netflix regarding the customer data they released to the public as part of the Netflix Prize project, a contest to help create better movie recommendations. A closeted lesbian alleges that the data available about her could reveal her identity.

Consumerist.com

Anonymity is Impossible

- Anonymity is virtually impossible, with enough other data.
  - Diversity of entity sets can be eliminated through joining external data
  - Random perturbation works only if we can guarantee a one-time perturbation
  - Aggregation works only if there is no known structure among entities aggregated
- Faces can be recognized in image data.
  - Progressively, even under challenging conditions, such as partial occlusion

Anonymity Techniques

- K-Anonymity
  - Require at least k entries in a group about which information is revealed.
  - Hope that is enough to hide details about any one individual.
  - But not provably safe.
- Differential Privacy
  - Only respond to aggregate queries about the data.
  - Add carefully calibrated noise to the aggregate value being reported.
  - Can guarantee (with high probability) not revealing detail data about presence of any individual in the data set.

Differential Privacy

- Widely used today
  - E.g. US Census Bureau
- Only method with provable guarantees
- But, repeated queries are a worry
- Concept of fixed "epsilon" budget

used up some epsilon budget

- Also, complaints about added noise from some researchers

Facebook/Cambridge Analytica

- Your preferences can be predicted by the app, better than by your roommate, based on 70 ”like”s on Facebook. (Better than your spouse with 300 “like”s).
- Once someone has such a powerful app, they really know you, and can “push your buttons”.
- We need to limit such use if we are to feel free to share in the datafied world.

Choice May not be Yours to Make

- “The Golden State killer,” Joseph DeAngelo, was identified on account of partial matches with DNA his cousins had entered at a genealogy website.

No Option to Exit

- In the past, one could get a fresh start by:
  - Moving to a new place
  - Waiting till the past fades
    - Reputations can be rebuilt over time.
- Big Data is universal and never forgets anything!!
  - Way back machine for the web
- Can we develop techniques to forget?

Outline

- Privacy
- Equity

Algorithmic Fairness

- Do the data “speak for themselves”?
- Can algorithms be biased?
- Can we make algorithms unbiased?
  - Is training data set representative of the population?
  - Is past population representative of future population?
  - Are observed correlations due to confounding processes?

Validity

- Bad data leads to bad decisions.

buggy data

- But most data are dirty.
- If decision-making is opaque, results can be bad in the aggregate, and catastrophic for an individual.
- What if someone has a loan denied because of an error in the data analyzed?

though may be good aggregately

Third Party Data

- Material decisions can often be made on the basis of public data or data provided by third parties.
- There often are errors in these data.
- Does the affected subject have a mechanism to correct errors?
    - Credit rating data on steroids.
- Does the affected subject even know what data were used?
- “Right of recourse”

Biased Data

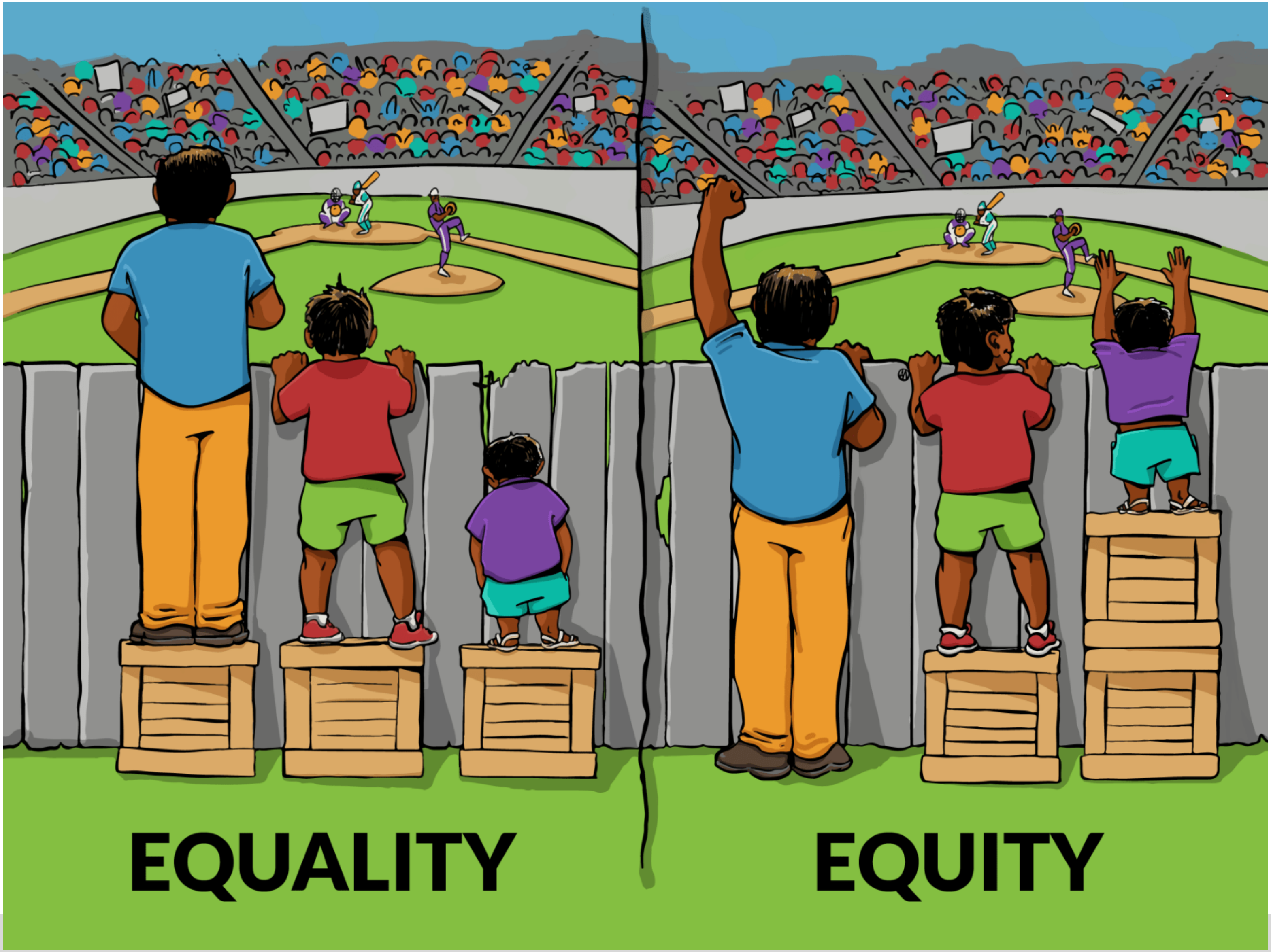
- Data collection mechanisms often result in biases.
  - Whether these matter requires thought.
- Social media posts are not representative of the general population
  - Skew younger, better educated, more tech-savvy.
  - Over-represent people with strong opinions
- Medical tests often at one (or a few) local site(s)
  - But results are claimed to apply throughout the world.
  - Most humans are indeed alike.
  - But what about racial/genetic differences?
  - Environmental differences between rich and poor nations.

Equity

Treat people differently based on their circumstances to achieve comparable outcomes.

Equity ≠ Fairness

Equity vs Equality



Interaction Institute
for Social Change
interactioninstitute.org

Artist: Angus Maguire
madewithangus.com

Example of Equity

- It is fair to spend an equal number of dollars per student in a public school.
    - Aggregate budget allocations often made this way.
- Equity requires addition spending on children with special needs.

Example of Equity

- It is fair to give each student in the class the same amount of time to take an exam.
- Equity requires allowing extra time for some students.

Example of Model Equity

- It is fair to measure every applicant's knowledge/potential through a standardized test, such as GRE.
- Equity requires taking into account studies showing the strong correlation between test performance and socio-economic status (and gender and race and …).

Example of Data Equity

- It is fair to create a training data set that is an unbiased sample of the population: each minority group is represented in proportion to its size in the population.
- Equity may require over-sampling of small minorities. If a small minority group "behaves" differently than others, the model may minimize aggregate error by ignoring the minority group.

Conclusion

- Data-driven automation can do a lot more, and do it a lot faster. But the "it" needs to be defined carefully.