Database design and implementation

CMPSCI 645

Lecture 22: Fairness, diversity, and transparency

slide credit: Data Responsibly, EDBT 2016 (Julia Stoyanovich, Serge Abiteboul, Gerome Miklau)
The promise of Big Data

**power**
- enormous data sets
- massively parallel processing

**goal:** progress

**opportunity**
- improve people’s lives
- accelerate scientific discovery
- boost innovation
- transform society
- optimize business
Illustration: Big Data and health

Analysis of a person’s medical data, genome, social data

personalized medicine
personalized care and predictive measures

personalized insurance
expensive, or unaffordable, for those at risk

the same technology makes both possible!
Is data analysis impartial?

- Big data is algorithmic, therefore it cannot be biased! And yet...

- All traditional evils of discrimination, and many new ones, exhibit themselves in the big data ecosystem

- We need novel technological solutions to identify and rectify irresponsible data analysis practices

Responsible data analysis

- Fairness
- Diversity
- Transparency
- Neutrality
It was the same Swingline stapler, on the same Staples.com website. But for Kim Wamble, the price was $15.79, while the price on Trude Frizzell’s screen, just a few miles away, was $14.29.

A key difference: where Staples seemed to think they were located.

Lower prices offered to customers who live in affluent neighborhoods.
Fairness is lack of bias

- Where does bias come from?
  - Data collection
  - Data analysis

- Analogy: Scientific data analysis
  - collect a representative sample
  - do sound reproducible analysis
  - explain data collection and analysis

When data is about people, bias can lead to discrimination
Assigning outcomes to populations

Population

Assignments

Positive outcomes
40% of the population

Individual with negative outcome

Individual with positive outcome
Sub-populations may be treated differently

<table>
<thead>
<tr>
<th>hair color</th>
<th>red</th>
<th>not red</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive outcomes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40% of the whole population</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20% of red haired</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60% of not red haired</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

disparate impact on red-haired people

statistical parity fails
Enforcing statistical parity

The demographics of the individuals receiving any outcome are the same as demographics of the underlying population.
Redundant encoding

Now consider the assignments under both hair color (protected) and hair length (innocuous).

The vendor has adversely impacted red-haired people, but claims that outcomes are assigned according to hair length.
Redundant encoding

Let’s replace hair color with **race** (protected), hair length with **zip code** (innocuous)
Discrimination may be unintended

Staples website estimated user’s location, offering discounts to those near rival stores, leading to discrimination w.r.t. to average income.

<table>
<thead>
<tr>
<th>income</th>
<th>rival store proximity</th>
<th>positive outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>close</td>
<td>20% of low income</td>
</tr>
<tr>
<td></td>
<td>far</td>
<td>60% of high income</td>
</tr>
<tr>
<td>high</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Result diversity

- **Dating query:** female, 40 or younger, at least some college, in order of decreasing income

- **Results** are homogeneous at top ranks. Both the seeker (asking the query) and the matches (results) are dissatisfied.

- Crowdsourcing, crowdfunding, ranking of Web search results, ...
  - all subject to this problem

the rich get richer, the poor get poorer
Data diversity

- Data must be representative

- Bias in data collection may be amplified in data analysis perpetuating original bias
Racially identifying names

[Latanya Sweeney; CACM 2013]

racially identifying names trigger ads suggestive of an arrest record
Transparency and accountability

- Users and regulators must be able to *understand* how raw data was selected, and what operations were performed during analysis.

- Users want to *control* what is recorded about them and how that information is used.

- Users must be able to *access* their own information and correct any errors (US Fair Credit Reporting Act).

- **Transparency** facilitates **accountability** - verifying that a service performs as it should, and that data is used according to contract.
The European Union accused Google on Wednesday of cheating competitors by distorting Internet search results in favour of its Google Shopping service and also launched an antitrust probe into its Android mobile operating system.
Neutrality

- Net neutrality
  - the network is transporting data with no bias based on source, destination, content ...

- Platform neutrality
  - big internet platforms should not discriminate in favor of their own services
Should the government regulate the big data industry?
- regulate
- define good practices
- evaluate responsibility

Issues:
- which government?
- lack of competence, agility

lots of gray areas, much work remains, enforcement is problematics since few auditing tools exist
Education

- Concepts
  - **understanding** data acquisition methods and data analysis processes
  - **verifying** the data and the process: provenance, credit attribution, trust
  - **interpreting** results

- Tools: computer science, probability and statistics, what people need to know about **data science**!
Statistics scares people, big data REALLY scares people!
Education: correlation vs causation

Global Average Temperature vs. Number of Pirates

https://commons.wikimedia.org/wiki/File:PiratesVsTemp(en).svg
Education: Data visualization

OBAMACARE ENROLLMENT

6,000,000

AS OF MARCH 27

7,066,000

MARCH 31 GOAL

SOURCE: HHS

Education: Data visualization