Where Does Data Bias Come from?

Emre Kıçımán  emrek@microsoft.com
Microsoft Research  http://kiciman.org
@emrek
Collaborators

Based on survey paper and tutorial series written with Alexandra Olteanu, Carlos Castillo, Fernando Diaz
“All models are wrong, but some are useful.”
George Box, 1987

“Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.”

Should We Let Data Speak for Itself?
Many Disagree!

“Regardless of the size of a data set, it is subject to limitation and bias. Without those biases and limitations being understood and outlined, misinterpretation is the result.”

[boyd and Crawford, 2012]

“There are a lot of small data problems that occur in big data. They don’t disappear because you’ve got lots of the stuff. They get worse.”

[David Spiegelhalter, in Harford, T. (March 28, 2014) Big data: are we making a big mistake. The Financial Times]
“Both the data set and the algorithms reflect choices about data, connections, inferences, interpretation, thresholds for inclusion, etc., that advance a specific purpose.” [Dwork and Mulligan, 2013]

“We must ask difficult questions of Big Data’s models of intelligibility before they crystallize into new orthodoxies.” [boyd and Crawford, 2012]

“Social media big data is a powerful addition to the scientific toolkit [that] needs to be placed on firmer methodological and conceptual footing.” [Tufekci, 2014]

“There is (…) the need for increased awareness of what is actually analyzed” [Ruths and Pfeffer, 2014]
Validity

Threats to Social Data Analysis

**Construct validity:** Are you measuring what you think you are measuring?

**Internal validity:** Does your analysis correctly lead from the measurements to your study conclusions?

**External validity:** To what extent your findings are generalizable to other situations?

**Ecological validity:** Does your experimental setup properly reflect the real world phenomenon your are studying?

**Temporal validity:** Do changes over time in the measured constructs invalidate the conclusions?
**Example:** Tracking Self-esteem on Social Media

**Hypothesis:** Self-esteem increases with age.

**Construct validity:** Are textual proxies used to measure self-esteem actually measuring self-esteem?

**Internal validity:** Are conclusions valid if data cleaning may remove textual cues reflecting confidence, or if the training data comes from teenagers?

**External validity:** Are observations from a given social platform generalizable to a broader situation if the users sample or their behavior is not representative?
Tutorial goals

To present a taxonomy of challenges that can occur at different stages of social data analysis.

To recognize, understand, or quantify some major classes of limitations around data.

To give us food for thought, by looking critically at our work.
Data Quality Issues

**Sparsity**: e.g., many measures follow a power law distribution.

**Noise**: e.g., content that is not reliable, content that is incomplete or corrupted, typos, infrequent terms, stop words.

**Representativeness**: e.g., a sample that is not representative of the larger population.

This talk focuses on *data bias*
Data bias: a *systematic distortion* in data that compromises its use for a task.
Note: Bias must be considered relative to task

Gender in loan application

Gender discrimination is illegal

Gender in medical diagnosis

Gender-specific medical diagnosis is desirable
What does data bias look like?

Measure systematic distortions along 5 data properties

1. Population Biases
2. Behavioral Biases
3. Content Production Biases
4. Linking Biases
5. Temporal Biases
What does data bias look like?

Measure distortions along 5 data properties

1. **Population Biases**
   Differences in demographics or other user characteristics between a user population represented in a dataset or platform and a target population

2. Behavioral Biases

3. Content Production Biases

4. Linking Biases

5. Temporal Biases
Example:
Different user demographics on different social platforms

See [Hargittai’07] for statistics about social media use among young adults according to gender, race and ethnicity, and parental educational background.
Different demographics use social platforms differently

Figure from [Altenburger et al. ICWSM’17]
Systematic distortions must be evaluated in a task dependent way.

E.g., for many tasks, populations should **match target population**, to improve **external validity**.

But for other tasks, subpopulations require approximately **equal representation** to achieve task parity.

http://gendershades.org/
What does data bias look like?

Measure distortions along 5 data properties

1. Population Biases
2. **Behavioral Biases**
   - Differences in user behavior across platforms or contexts, or across users represented in different datasets
3. Content Production Biases
4. Linking Biases
5. Temporal Biases
Behavioral Biases from *Functional Issues*

Platform functionality and algorithms influence human behaviors and our observations of human behaviors.

Figure from: [http://grouplens.org/blog/investigating-the-potential-for-miscommunication-using-emoji/](http://grouplens.org/blog/investigating-the-potential-for-miscommunication-using-emoji/)
Abby using a Google Nexus, texting Bill:

Bill using an iPhone, texting Abby:

Figure from: http://grouplens.org/blog/investigating-the-potential-for-miscommunication-using-emoji/
Cultural elements and social contexts are reflected in social datasets.

The way users are perceived affects their interaction patterns (e.g., more or less content sharing/followers).

Women’s code changes are more likely to be accepted in Github, unless they are identified as women.

Figure from [Terrel et al., pre-print]
Societal biases embedded in behavior can be amplified by algorithms

System presents options, influencing user choice

Users pick biased options

Biased actions are used as feedback

System learns to mimic biased options
Autocomplete for Search Interfaces

See also: Seth Stephens-Davidowitz. Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are (2017)
What does data bias look like?

Measure distortions along 5 data properties

1. Population Biases
2. Behavioral Biases
3. **Content Production Biases**
   - Lexical, syntactic, semantic, and structural differences in the contents generated by users
4. Linking Biases
5. Temporal Biases
The use of language(s) varies across and within countries and populations.

The second language by district or municipality (in the case of New Jersey state) is shown. Blue - Spanish, Light Green - Korean, Fuchsia - Russian, Red - Portuguese, Yellow - Japanese, Pink - Dutch, Grey - Danish, Coral - Indonesian.

Figure source [Mocanu et al. PlosOne 2013]
The use of language(s) varies across and within countries and populations.

<table>
<thead>
<tr>
<th>Feature</th>
<th>#female/#male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emoticons</td>
<td>3.5</td>
</tr>
<tr>
<td>Ellipses</td>
<td>1.5</td>
</tr>
<tr>
<td>Character repetition</td>
<td>1.4</td>
</tr>
<tr>
<td>Repeated exclamation</td>
<td>2.0</td>
</tr>
<tr>
<td>Puzzled punctuation</td>
<td>1.8</td>
</tr>
<tr>
<td>OMG</td>
<td>4.0</td>
</tr>
</tbody>
</table>

References

[Rao et al., Workshop on Search and Mining User Generated Contents 2010]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Females $\rho$</th>
<th>Males $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Style</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Capitalized words</td>
<td>-0.281**</td>
<td>-0.453**</td>
</tr>
<tr>
<td>Alph. lengthening</td>
<td>-0.416**</td>
<td>-0.324**</td>
</tr>
<tr>
<td>Intensifiers</td>
<td>-0.308**</td>
<td>-0.381**</td>
</tr>
<tr>
<td>LIWC-prepositions</td>
<td>0.577**</td>
<td>0.486**</td>
</tr>
<tr>
<td>Word length</td>
<td>0.630**</td>
<td>0.660**</td>
</tr>
<tr>
<td>Tweet length</td>
<td>0.703**</td>
<td>0.706**</td>
</tr>
<tr>
<td><strong>References</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>-0.518**</td>
<td>-0.481**</td>
</tr>
<tr>
<td>You</td>
<td>-0.417**</td>
<td>-0.464**</td>
</tr>
<tr>
<td>We</td>
<td>0.312**</td>
<td>0.266**</td>
</tr>
<tr>
<td>Other</td>
<td>-0.072</td>
<td>-0.148**</td>
</tr>
</tbody>
</table>

Conversation

| Replies                | 0.304**        | 0.026        |

Sharing

| Retweets               | -0.101*        | -0.099*      |
| Links                  | 0.428**        | 0.481**      |
| Hashtags               | 0.502**        | 0.462**      |

Pearson correlation with the age of the tweet author. Table from [Nguyen et al. ICWSM 2013]
Content Biases from Normative Issues

Community norms and societal biases influence observed behavior and vary across online and offline communities and contexts.

What kind of pictures would you share on Facebook, but not on LinkedIn?

Are individuals comfortable contradicting popular opinions?

The same mechanism can embed different meanings in different contexts [Tufekci ICWSM’14]

[the meaning of retweets or likes] “could range from affirmation to denunciation to sarcasm to approval to disgust”

E.g., after singer Prince died, most SNs showed public mourning. But not anonymous site PostSecret
The awareness of being observed by other impacts user behavior: **Privacy and safety concerns**

**Privacy concerns** affect what content users share, and, thus, the type of patterns we observe.

![Bar charts showing frequency of visits to different places](image)

<table>
<thead>
<tr>
<th>Place</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restaurants</td>
<td>Less than once a month (20), Once a month (30), Once a week (50), More than once a week (60)</td>
</tr>
<tr>
<td>Bars</td>
<td>Less than once a month (20), Once a month (30), Once a week (50), More than once a week (60)</td>
</tr>
<tr>
<td>School</td>
<td>Less than once a month (20), Once a month (30), Once a week (50), More than once a week (60)</td>
</tr>
<tr>
<td>Doctor</td>
<td>Less than once a month (20), Once a month (30), Once a week (50), More than once a week (60)</td>
</tr>
<tr>
<td>Home</td>
<td>Less than once a month (20), Once a month (30), Once a week (50), More than once a week (60)</td>
</tr>
<tr>
<td>Work</td>
<td>Less than once a month (20), Once a month (30), Once a week (50), More than once a week (60)</td>
</tr>
</tbody>
</table>

Foursquare/Image from [Lindqvist et al. CHI’11]
Adversaries also manipulate content

**Misinformation** is false information, unintentionally spread

**Disinformation** is false information, deliberately spread

Hoaxes on Wikipedia: *(left)* impact as number of views per day for hoaxes surviving at least 7 days, and *(right)* time until a hoax gets detected and flagged.

Figures from [Kumar et al. 2016]
What does data bias look like?

Measure distortions along 5 data properties

1. Population Biases
2. Behavioral Biases
3. Content Production Biases
4. Linking Biases
   Differences in the attributes of networks obtained from user connections, interactions, or activity
5. Temporal Biases
Behavior-based and connection-based social links are different

Figure 14. Graph measurements for four interaction graphs compared to the entire Facebook social network.

Figure from [Wilson et al. EuroSys’09]
Online social networks formation also depends on factors external to the social platforms

- Geography & distance
- Co-visits
- Dynamics of offline relations
- [...]

Figure from [Gilbert and Karahalios CHI 2009]
What does data bias look like?

Measure distortions along 5 data properties

1. Population Biases
2. Behavioral Biases
3. Content Production Biases
4. Linking Biases
5. **Temporal Biases**
   
**Differences in populations and behaviors over time**
Different demographics can exhibit different growth rates across and within social platforms.

TaskRabbit and Fiverr are online freelance marketplaces. Figure from [Hannak et al. CSCW 2017]

Figure 1: Member growth over time on TaskRabbit and Fiverr, broken down by gender and race.
E.g., Change in Features over Time

Introducing a new feature or changing an existing feature impacts usage patterns on the platform.
5 measurements of biases

1. Population Biases
2. Behavioral Biases
3. Content Production Biases
4. Linking Biases
5. Temporal Biases
Biases can come in at any step along the data analysis pipeline.
Best Practices for Bias Avoidance/Mitigation

Consider **team composition** for diversity of thought, background and experiences
Best Practices for Bias Avoidance/Mitigation

Understand the **task**, **stakeholders**, and potential for **errors and harm**
Best Practices for Bias Avoidance/Mitigation

Check data sets
Consider data provenance
What is the data intended to represent?
Verify through qualitative, experimental, survey and other methods
**Best Practices for Bias Avoidance/Mitigation**

**Check models and validate results**

**Why** is the model making decision? What **mechanisms** would explain results? Is supporting evidence consistent?

**Twyman’s law**: The more unusual the result, more likely it’s an error.
Best Practices for Bias Avoidance/Mitigation

Post-Deployment

Ensure **optimization and guardrail metrics consistent w/responsible practices and avoid harms**

**Continual monitoring**, including customer feedback

Have a **plan to identify and respond to failures and harms** as they occur
Summary advice:

Be careful!
Key Takeaways

• Many, complex biases at all stages of data collection and analysis
  • Population, Behavioral, Content Production, Linking Temporal Biases

• Mitigate through deeper investigation, understanding

• Read more: Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries, Olteanu, Castillo, Diaz and Kıcıman