Biases in Research Using Social Data

Based on:
Alexandra Olteanu, Carlos Castillo, Fernando Diaz, Emre Kıcıman (July 2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers on Big Data*
https://doi.org/10.3389/fdata.2019.00013
http://www.aolteanu.com/SocialDataLimitsTutorial/

Presented by:
Carlos Castillo
ICREA and Universitat Pompeu Fabra

Updated on May 2022, some materials/references added

frontiers in Big Data

REVIEW
published: 11 July 2019
doi: 10.3389/fdata.2019.00013

Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries

Alexandra Olteanu[1,2], Carlos Castillo[3], Fernando Diaz[2] and Emre Kıcıman[4]*

[1] Microsoft Research, New York, NY, United States, [2] Microsoft Research, Montreal, QC, Canada, [3] Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain, [4] Microsoft Research, Redmond, WA, United States

Social data in digital form—including user-generated content, expressed or implicit relations between people, and behavioral traces—are at the core of popular applications and platforms, driving the research agenda of many researchers. The promises of social data are many, including understanding "what the world thinks" about a social issue, brand, celebrity, or other entity, as well as enabling better decision-making in a variety of fields including public policy, healthcare, and economics. Many academics and practitioners have warned against the naïve usage of social data. There are biases and inaccuracies occurring at the source of the data, but also introduced during processing. There are methodological limitations and pitfalls, as well as ethical boundaries and unexpected consequences that are often overlooked. This paper recognizes the rigor with which these issues are addressed by different researchers varies across a wide range. We identify a variety of menaces in the practices around social data use, and organize them in a framework that helps to identify them.

"For your own sanity, you have to remember that not all problems can be solved. Not all problems can be solved, but all problems can be illuminated."—Ursula Franklin[1]

Keywords: social media, user data, biases, evaluation, ethics

OPEN ACCESS

1. INTRODUCTION

We use *social data* as an umbrella concept for all kind of digital traces produced by or about users, with an emphasis on content explicitly written with the intent of communicating or interacting with others. Social data typically comes from *social software*, which provides an intermediary or a focus for a social relationship (Schuler, 1994). It includes a variety of *platforms*—like for social media and networking (e.g., Facebook), question and answering (e.g., Quora), or collaboration (e.g., Wikipedia)—and *purposes* from finding information (White, 2013) to keeping in touch with friends (Lampe et al., 2008). Social software enables the *social web*, a class of websites "in which user participation is the primary driver of value" (Gruber, 2008).

The social web enables access to social traces at a scale and level of detail, both in breadth and depth, impractical with conventional data collection techniques, like surveys or user studies (Richardson, 2008; Lazer et al., 2009). On the social web users search, interact, and share information on a mix of topics including work (Ehrlich and Shami, 2010), food (Abbar et al., 2015), or health (De Choudhury et al., 2014); leaving, as a result, rich traces that form what Harford (2014)

[1] Quoted by M. Meredith in http://hb9.berlinbiennale.de/all-problems-can-be-illuminated-not-all-problems-can-be-solved/

Frontiers in Big Data | www.frontiersin.org 1 July 2019 | Volume 2 | Article 13

Universitat Pompeu Fabra Barcelona

The promise of online social data is huge

Larger samples

Samples that are not WEIRD:
white, **e**ducated, **i**ndustrialized, **r**ich, **d**emocratic

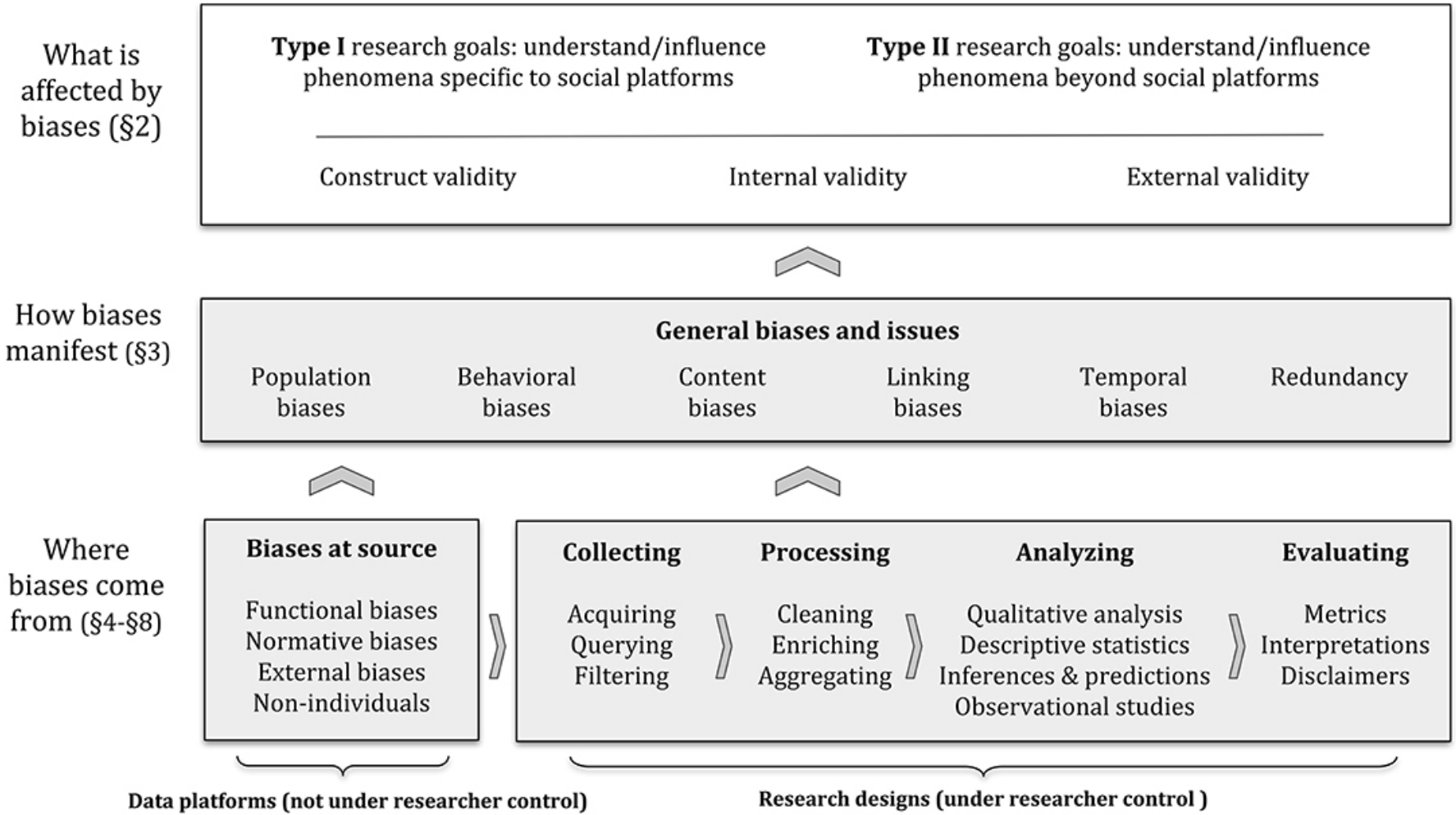
"Natural" settings

…

“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.**”

The end of theory, Chris Anderson, 2008, Wired Magazine.

Emphasis added.


What is affected by biases (§2)
Type I research goals: understand/influence phenomena specific to social platforms
Type II research goals: understand/influence phenomena beyond social platforms
Construct validity
Internal validity
External validity
How biases manifest (§3)
General biases and issues
Population biases
Behavioral biases
Content biases
Linking biases
Temporal biases
Redundancy
Where biases come from (§4-§8)
Biases at source
Functional biases
Normative biases
External biases
Non-individuals
Collecting
Acquiring
Querying
Filtering
Processing
Cleaning
Enriching
Aggregating
Analyzing
Qualitative analysis
Descriptive statistics
Inferences & predictions
Observational studies
Evaluating
Metrics
Interpretations
Disclaimers
Data platforms (not under researcher control)
Research designs (under researcher control)

Prototypical Goals & Applications

A Growing Concern

Validity of Social Data Research

Issues Along a Prototypical Analysis Pipeline

Prototypical Goals & Applications

A Growing Concern

Validity of Social Data Research

Issues Along a Prototypical Analysis Pipeline

Social Data

All kinds of digital traces produced by or about users,
with an emphasis on content explicitly written with the intent of communicating or interacting with others

Social data comes from **social software**

Social Software

Software that provides an intermediary or a focus for a **social relationship**

Includes a variety of

- **Purposes** (e.g., communication, friendship maintenance, self-presentation)
- **Platforms** (e.g., social media and networking, recommendation and Q&A sites)
- **Data points meanings and semantics** (e.g., clicks, likes, shares, friendship links, visits, messages)

Prototypical Goals of Social Data Analysis

Social Computing

Computational Social Science

Type I: To understand phenomena specific to social software platforms, sometimes with the objective of improving them *(~ Social Computing)*

Type II: To understand and influence phenomena beyond social platforms, seeking to answer questions from sociology, psychology, or other disciplines *(~ Computational Social Science)*

Prototypical Goals & Applications

➡ A Growing Concern

Validity of Social Data Research

Issues Along a Prototypical Analysis Pipeline

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]

Data:
Source of Hypotheses

-or-

Tool to Test Hypotheses

Avoid HARKing: The practice of **h**ypothesizing **a**fter the **r**esults are **k**nown [Kerr'98]

Side question:
Do you "**work on x**" or you are "**trying to solve problem x**"?
Without a problem statement, "**working on x**" can mean anything.

Prototypical Goals & Applications

A Growing Concern

➡ Validity of Social Data Research

Issues Along a Prototypical Analysis Pipeline

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?

Known 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.

Data bias: a *systematic distortion* in the data that compromises its *representativeness*.

Prototypical Goals & Applications

A Growing Concern

Validity of Social Data Research

➡ Issues Along a Prototypical Analysis Pipeline

Issues When Working With Social Data

General Challenges

- **Population Biases:** differences in demographics
- **Behavioral Biases:** differences in user behavior
- **Content Biases:** lexical, syntactic, semantic biases in user content
- **Linking Biases:** differences in network connections, interactions
- **Temporal Biases:** changing biases over time
- **Redundancy:** duplicates, near duplicates

Challenges During Data Analysis (next slide)

- **Data Source:** biases at the source of social data
- **Data Collection:** biases due to data collection
- **Data Processing:** biases due to data preprocessing
- **Data Analysis:** validity threats due to methods selection and usage
- **Evaluation:** metrics selection, interpretation pitfalls

Challenges during data analysis

Data Source

- **Functional:** biases due to platform affordances and algorithms
- **Normative:** biases due to community norms
- **External:** biases due to phenomena outside social platforms
- **Non-individuals:** e.g., organizations, automated agents

Data Collection

- **Acquisition:** biases due to, e.g., API limits
- **Querying:** biases due to, e.g., query formulation
- **Filtering:** biases due to removal of data “deemed” irrelevant

Data Processing

- **Cleaning:** biases due to, e.g., default values
- **Enrichment:** biases from manual or automated annotations
- **Aggregation:** e.g., grouping, organizing, or structuring data

Data Analysis

- **Qualitative Analyses:** lack generalizability, interpret. biases
- **Descriptive Statistics:** confounding bias, obfuscated measurements
- **Prediction & Inferences:** data representation, perform. variations
- **Observational studies:** peer effects, select. bias, ignorability

Evaluation

- **Metrics:** e.g., reliability, lack of domain insights
- **Interpretation:** e.g., contextual validity, generalizability
- **Disclaimers:** e.g., lack of negative results and reproducibility

Population Biases

Differences in demographics or other user characteristics between a user population represented in a dataset or platform and a target population

Different user demographics are drawn to different social platforms

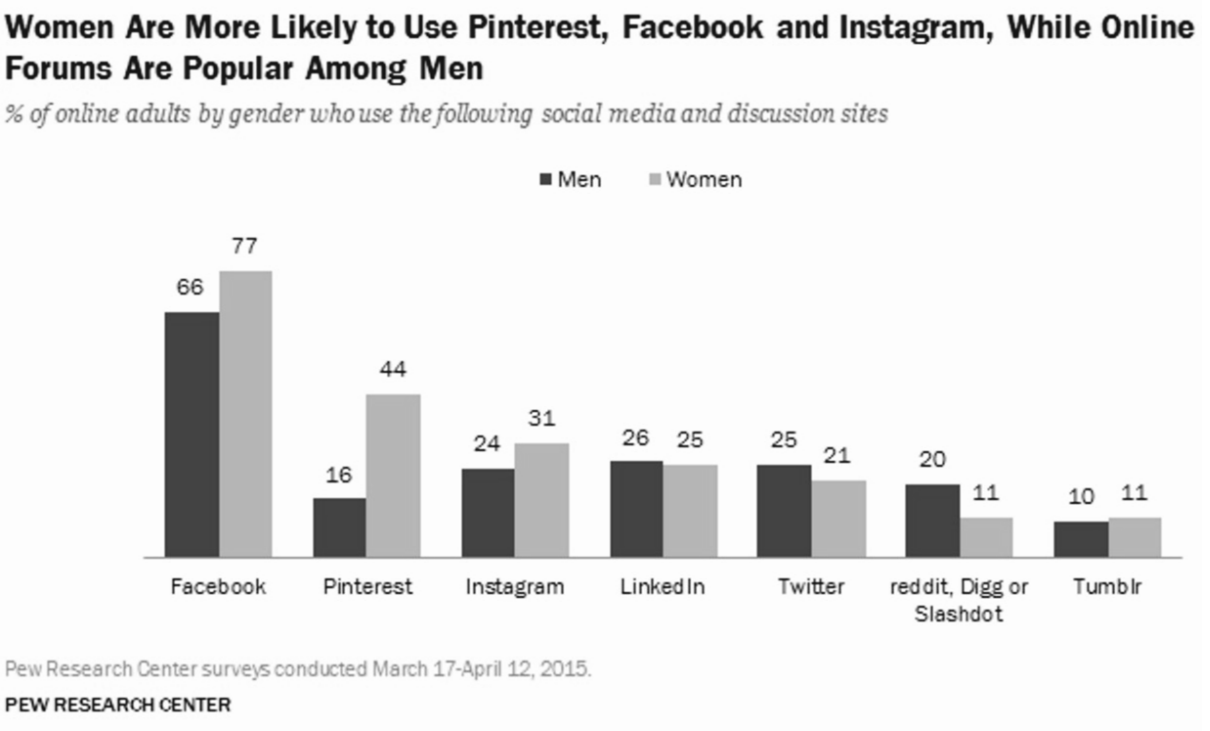


Figure from http://www.pewinternet.org/2016/11/11/social-media-update-2016/

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

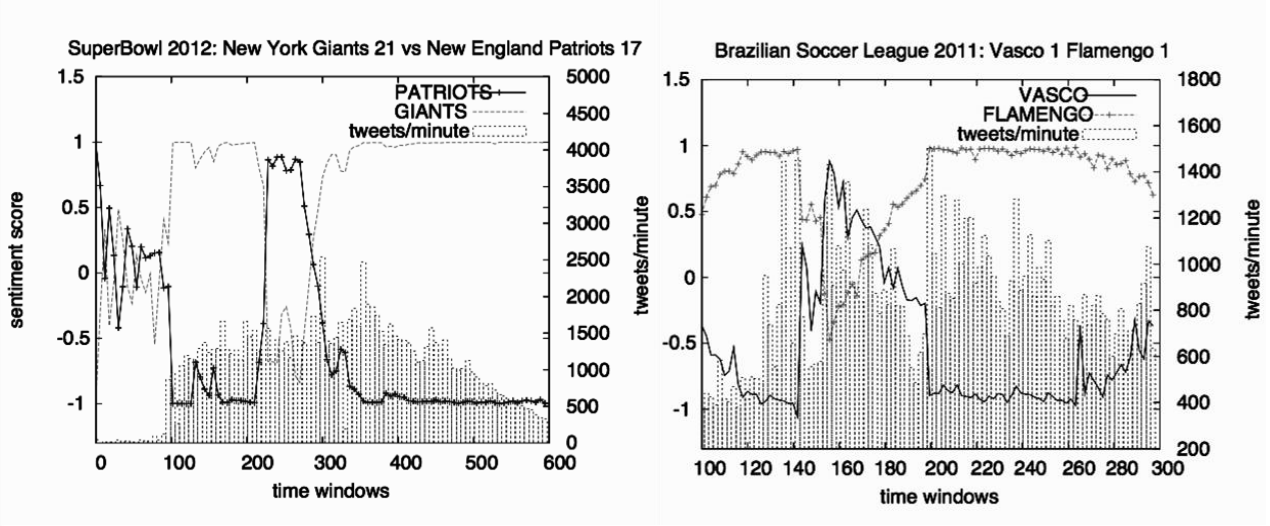
Behavioral Biases

Differences in user behavior across platforms or contexts, or across users represented in different datasets

Misreports and self-selection may occur due to behavioral biases

Studies relying on self-reports can be biased due to **what** users report, **when** they report, and **how** they report.

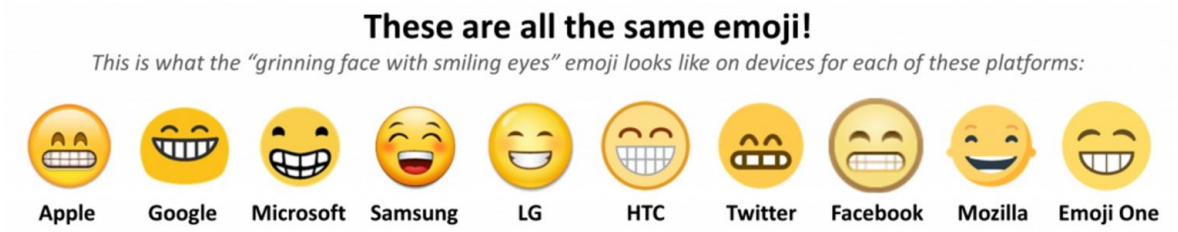
Users are more likely to mention their extreme or positive experiences, rather than their negative or common experiences [Kiciman ICWSM'12, Guerra et al. WSDM'14 (also the source of the figures)]

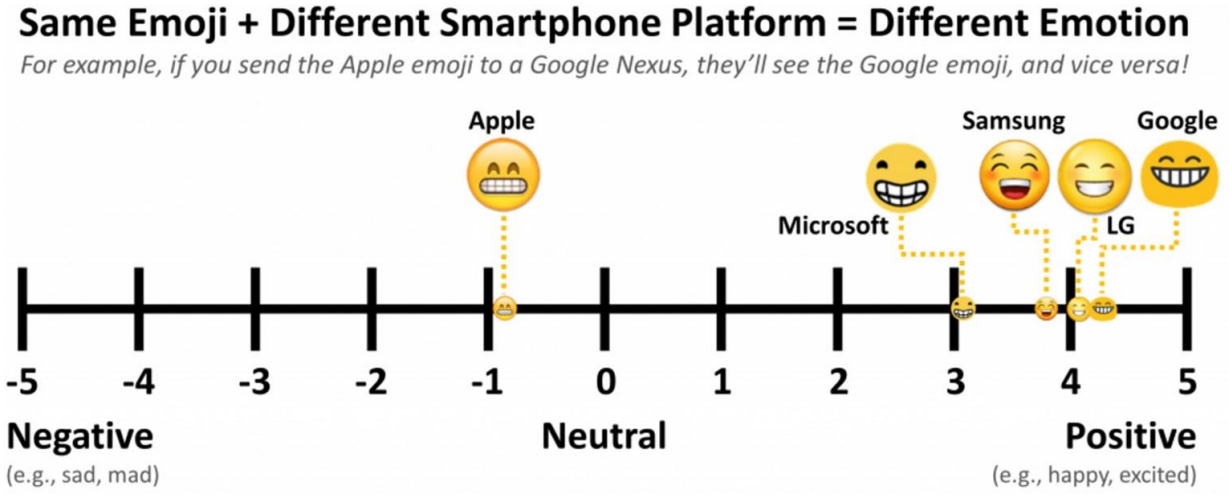


Functional Biases

Biases that are a result of platform-specific mechanisms or affordances, that is, the possible actions within each system or environment

The way in which contents are presented influences user behavior



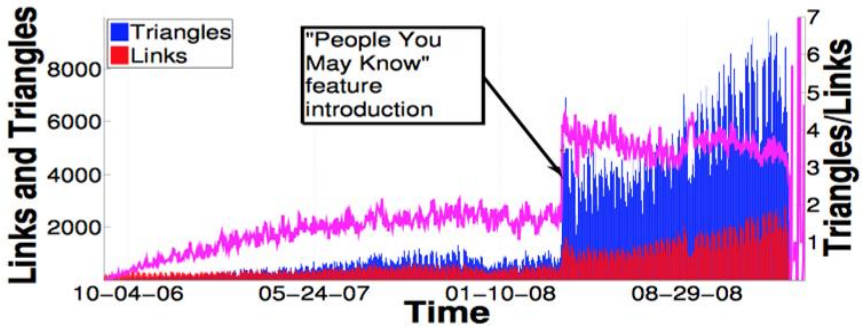


Paper [Miller et al. ICWSM'16]

Figure from: http://grouplens.org/blog/investigating-the-potential-for-miscommunication-using-emoji/

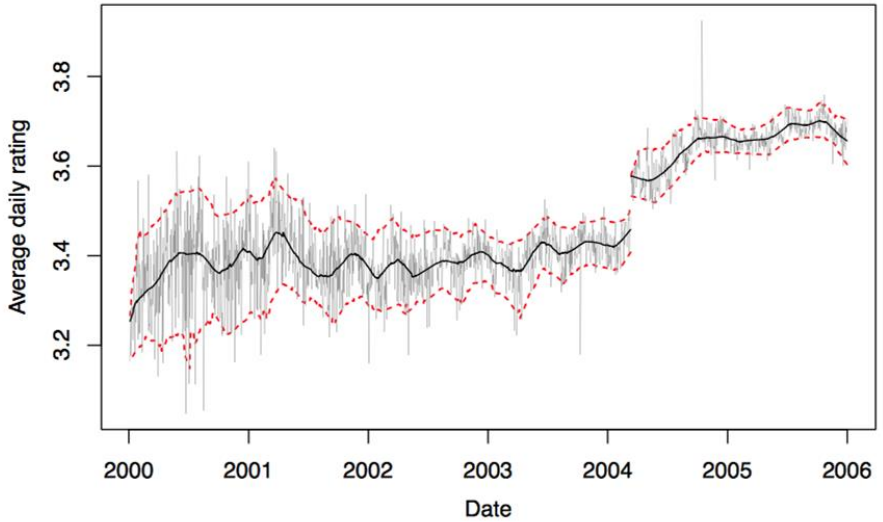
Platform design and features shape user behavior

Introducing a new feature or **changing an existing feature** impacts usage patterns on the platform.



(b) Facebook New Orleans

Figure from [Zignani et al. ICWSM 2014]



Netflix movie rating dataset.

Figure from [Malik and Pfeffer ICWSM 2016]

Normative Biases

Biases that are a result of written norms or expectations about unwritten norms describing acceptable patterns of behavior on a given platform

Social conformity, “anchoring” and “herding” happen in social platforms

The “up-treated” means that the treatment group was shown scores increased by 1.

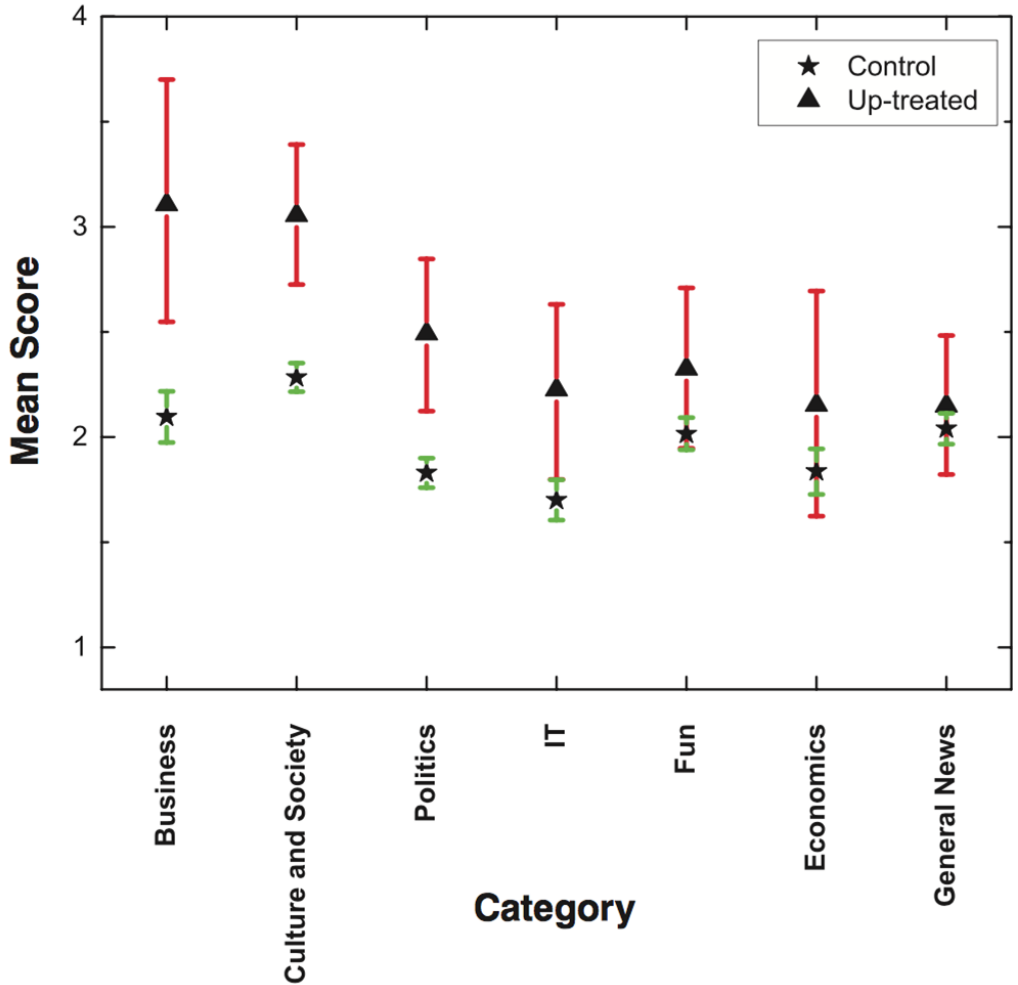


Figure from [Muchnik et al. Science 2013]

Issues Introduced During Data Collection

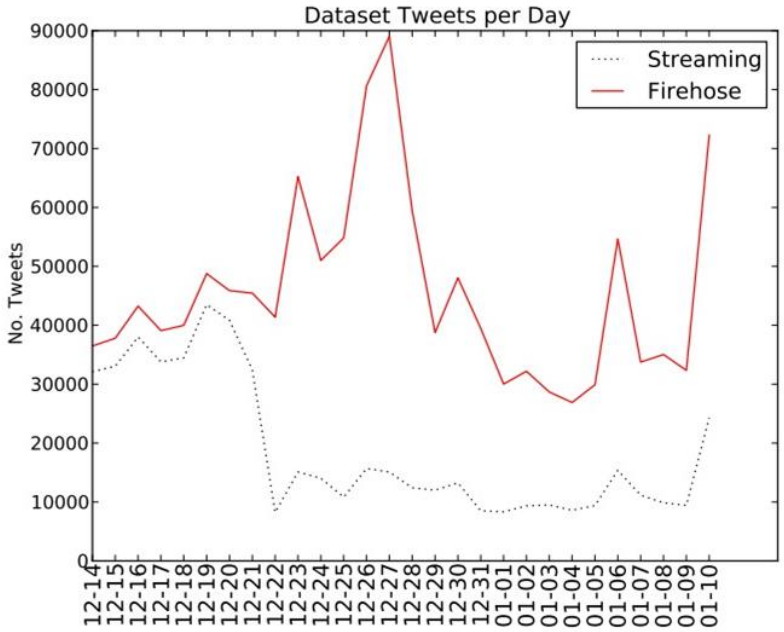
Source Selection Bias

Data Acquisition

Data Querying

Data Filtering

Platforms' sampling strategies are often opaque



Data sampled with a list of hashtags related to Syria.
Figures from [Morstatter et al. ICWSM'13]

Streaming API

Firehose

Challenges When Processing Data

Improve construct validity

Detect and correct errors and inconsistencies in the data

Data Cleaning

Data Enrichment

Data Aggregation

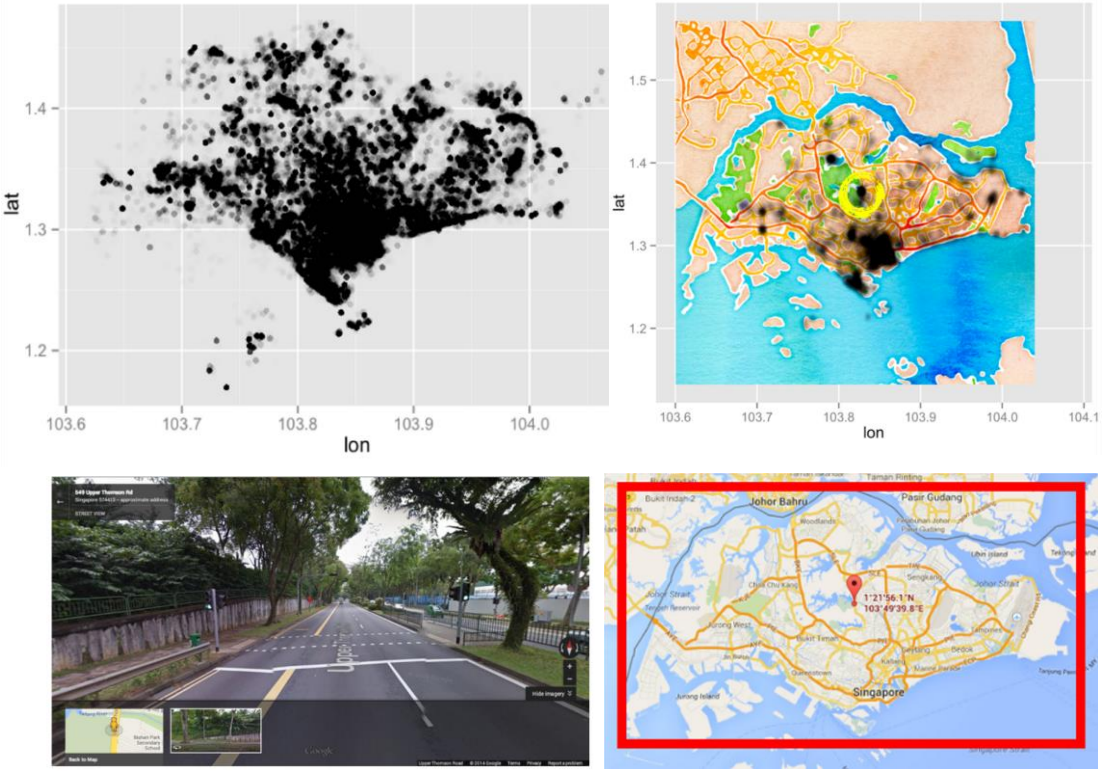
The normalization of geographical references may introduce biases

Story: Failed geo-lookups on a IP-to-Geo service default to the center of the geo-bounding box of a larger region e.g. country.

Result: ~600 million IP addresses map to a farm in Kansas

Details:http://fusion.kinja.com/how-an-internet-mapping-glitch-turned-a-random-kansas-f-1793856052

Where are Flickr pictures geo-located in Singapore?



Figures from Medium post by David Shamma: https://medium.com/@ayman/the-social-concerns-of-geo-located-rectangles-9b361f34811d

Can all biases be avoided through careful sampling?

No

Ideas on how to mitigate biases

- Engage with **domain experts**
- Perform **qualitative analysis**
  - In-depth, open-ended, questioning *how* and *why* something happens
- Use **multiple** datasets and metrics
- Use appropriate **statistical** methods
- Emphasize extensively the **limitations**
- Allow others to **reproduce** and to meaningfully *challenge* your results

KEEP CALM AND QUESTION EVERYTHING