

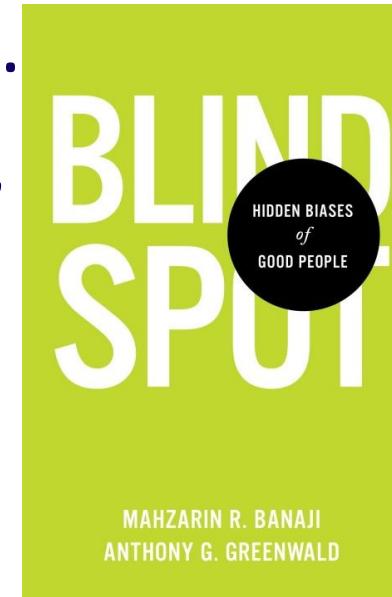


DSCI 531
SOCIAL DATA BIAS

Kristina Lerman
(slides courtesy of Fred Morstatter)
Spring 2025

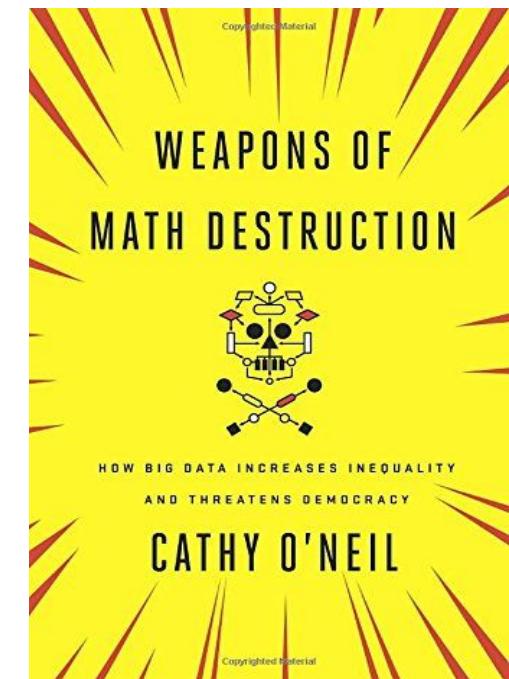
“Hidden Biases of Good People”

- Bias is subjective
 - It's not easy to point out one's own bias
- Hidden bias can influence our behavior without our noticing its influence
- Bias results from our limited knowledge, experience, and information overloading.
- If it is inevitable for humans to have bias, can machines help us to be balanced and fair?



Can Machines or Algorithms Help?

- Information overloading + our laziness → our reliance on machines or algorithms
- With algorithms, everyone is judged according to the same rules
 - In theory, bias should be eliminated
 - In reality, algorithms reinforce discrimination
- Ample examples are given in WMD and in “The Filter Bubble”
- Reasons include: opacity, lack of regulations, and ***data bias***



Data Bias – An Example

Biases innately exist in data

A beauty contest was judged by AI and the robots didn't like dark skin

The first international beauty contest decided by an algorithm has sparked controversy after the results revealed one glaring factor linking the winners



“...the main problem was that the data the project used to establish standards of attractiveness did not include enough minorities...”

the guardian

If We Can Detect Bias, We Can Mitigate It

- Visual bias example + 
 - We should look more, around and beyond
- “Orchestrating Impartiality”
 - “<5% all players in top 5 symphony orchestras are female musicians in 1970, ...
 - but 25% today”
 - The impact of “blind” auditions on female musicians
- In this lecture, we look at one kind of bias - social data bias

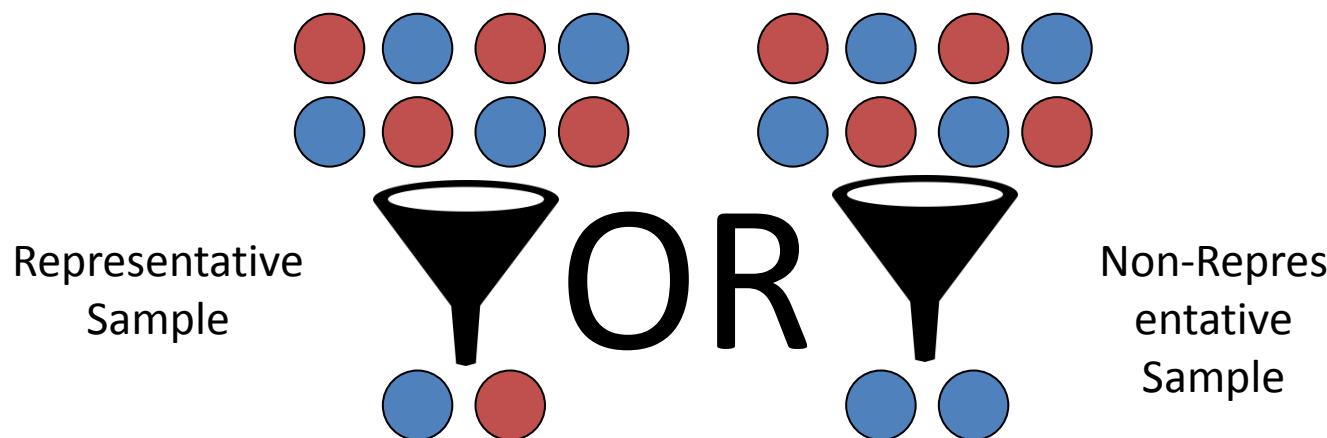
Bias Definition(s)

- Bias Definition:

Prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair.

- Operational Definition:

Non-representativity



Social Data Bias

Population Biases

- When the users under study do not match the population.

Behavioral Biases

- The features of the site dictate how it is used.

Content Biases

- Contextual behavioral bias. Example: Parents talk differently to their children than they do to their friends.

Linking Biases

- Network can influence behavior. Example: users behave differently as they follow more people on the platform.

Temporal Variations

- User behavior changes with time.

Redundancy

- Same content being posted multiple times.

Olteanu, Alexandra, et al. "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries." (2019).

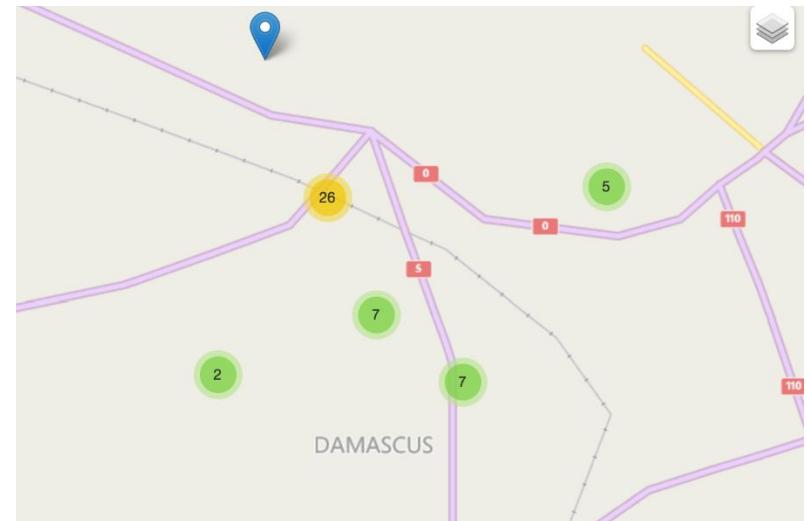
Before Social Data

- Smaller Samples
- WEIRD
 - Western, educated, industrialized, rich and democratic
- “Laboratory” setting
- Non-scalable content analysis
- ...

Social Media

- Damascus, Syria
- Who are the most important people?
- How does information flow through the city?

Geolocated tweets:

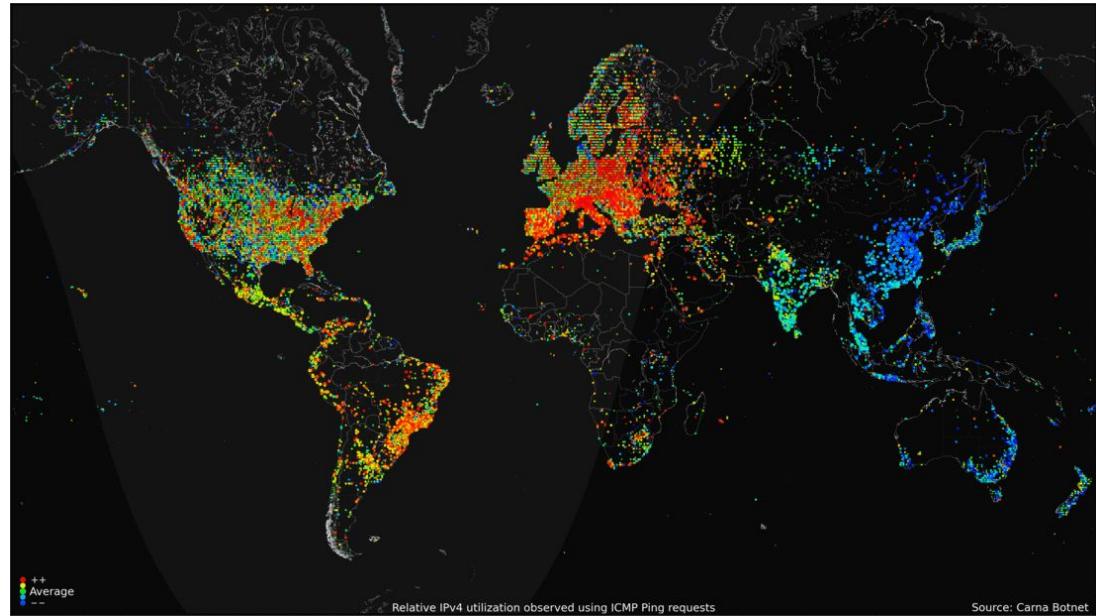
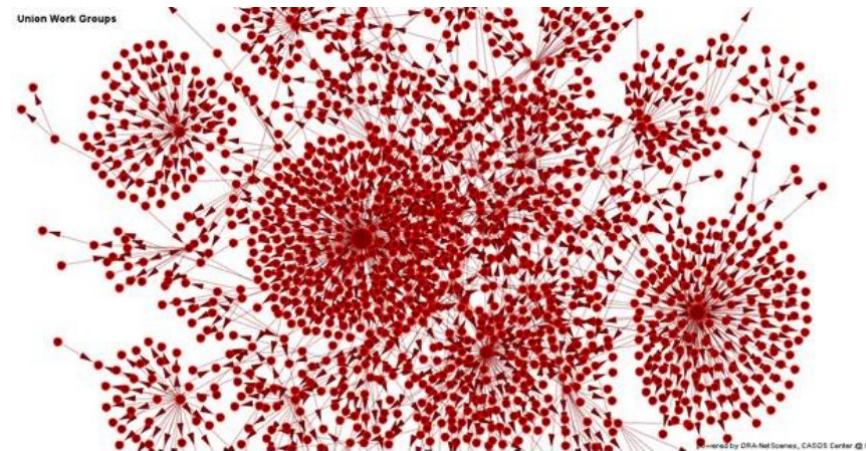


Top Hashtags:



Social Media Mining

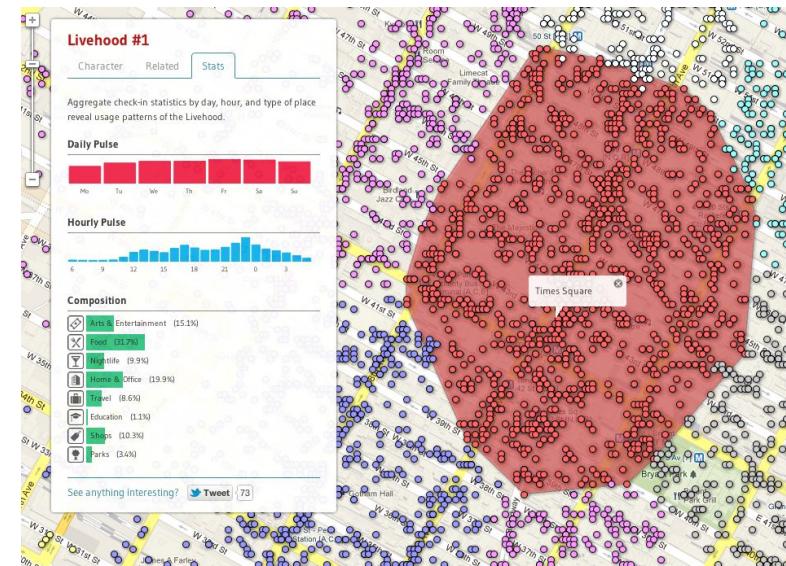
- 2.34 billion social media users in 2016.
- Enables us to answer societal questions at a global scale.



Source: <https://blog.kaspersky.com/amazing-internet-maps/10441/>

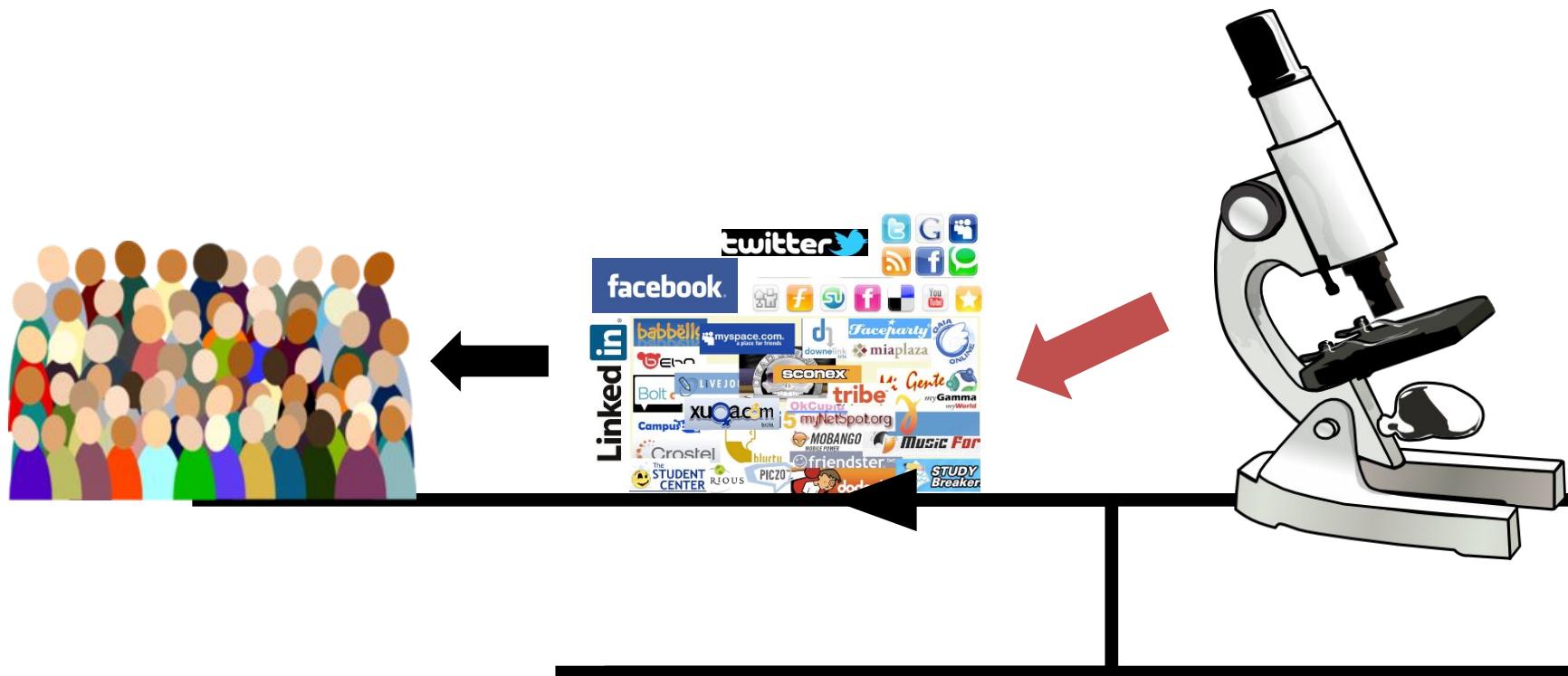
What are those questions?

- Information Diffusion – Who will buy a product?
 - Who will be infected by a disease?
- What are the dynamics of an urban area?
- What biases exist in society?
- Learn from **social data** to learn about **society**!



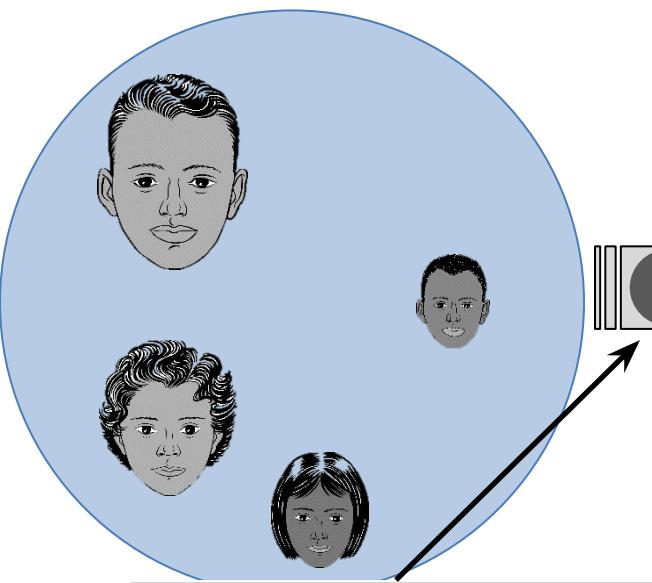
Social Media Mining

When we study social media data, we want to study *society*.
We only get *traces* of human behavior.



Driving Challenge: Representativity

The World

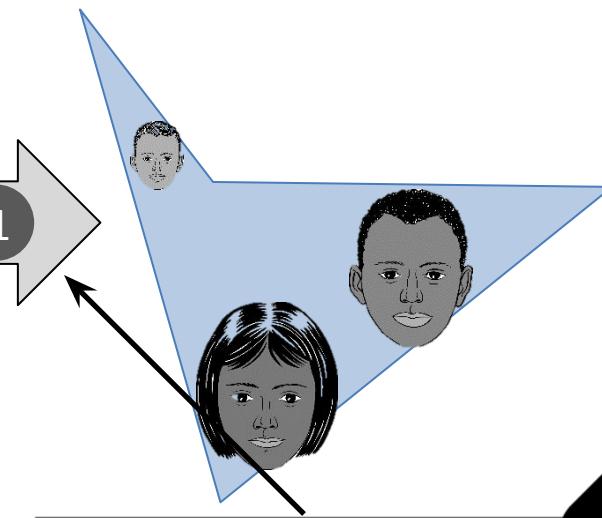


31% of online adults (26% of all Americans) use Pinterest

% of online adults who use Pinterest

All online adults	31%
Men	17
Women	45

Social Media

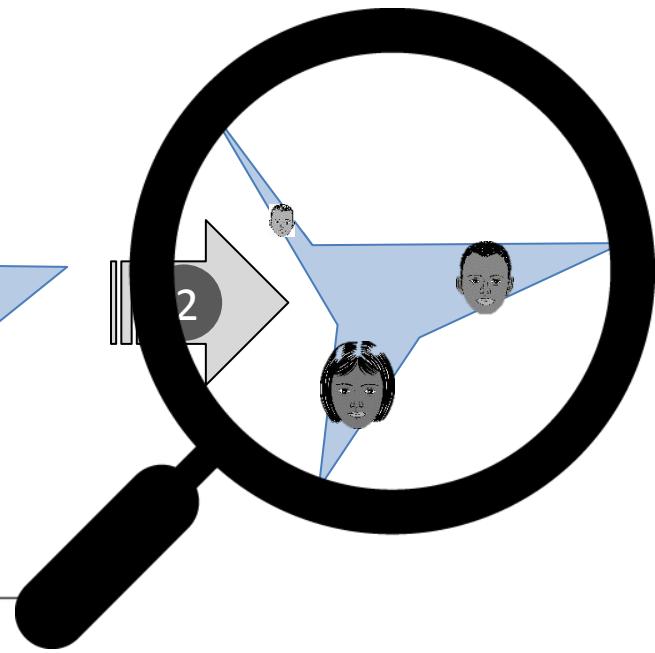


24% of online adults (21% of all Americans) use Twitter

% of online adults who use Twitter

All online adults	24%
Men	24
Women	25

Sampled Social Data



Goals of [SM] Analysis

- **Type I:** To understand phenomena specific to social software systems.
- **Type II:** To understand and influence phenomena beyond social platforms.

Threats to Validity of SM Analysis

- **Construct Validity:** Are you measuring what you think you are measuring?
- **Internal Validity:** Does your analysis lead from the measurements to your conclusions?
- **External Validity:** To what extent findings generalize?
 - *Ecological validity:* Does your experimental setup reflect the reality you are trying to study?
 - *Temporal validity:* Things change with time. Does this invalidate the conclusions?

Bias Taxonomy

What is affected by biases?

How do biases manifest?

Where do biases come from?

Type I research goals: understand/influence phenomena specific to social platforms

Construct validity

Type II research goals: understand/influence phenomena beyond social platforms

Internal validity

External validity



General biases and issues

Population biases

Behavioral biases

Content biases

Linking biases

Temporal biases

Redundancy



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)

How do biases manifest?

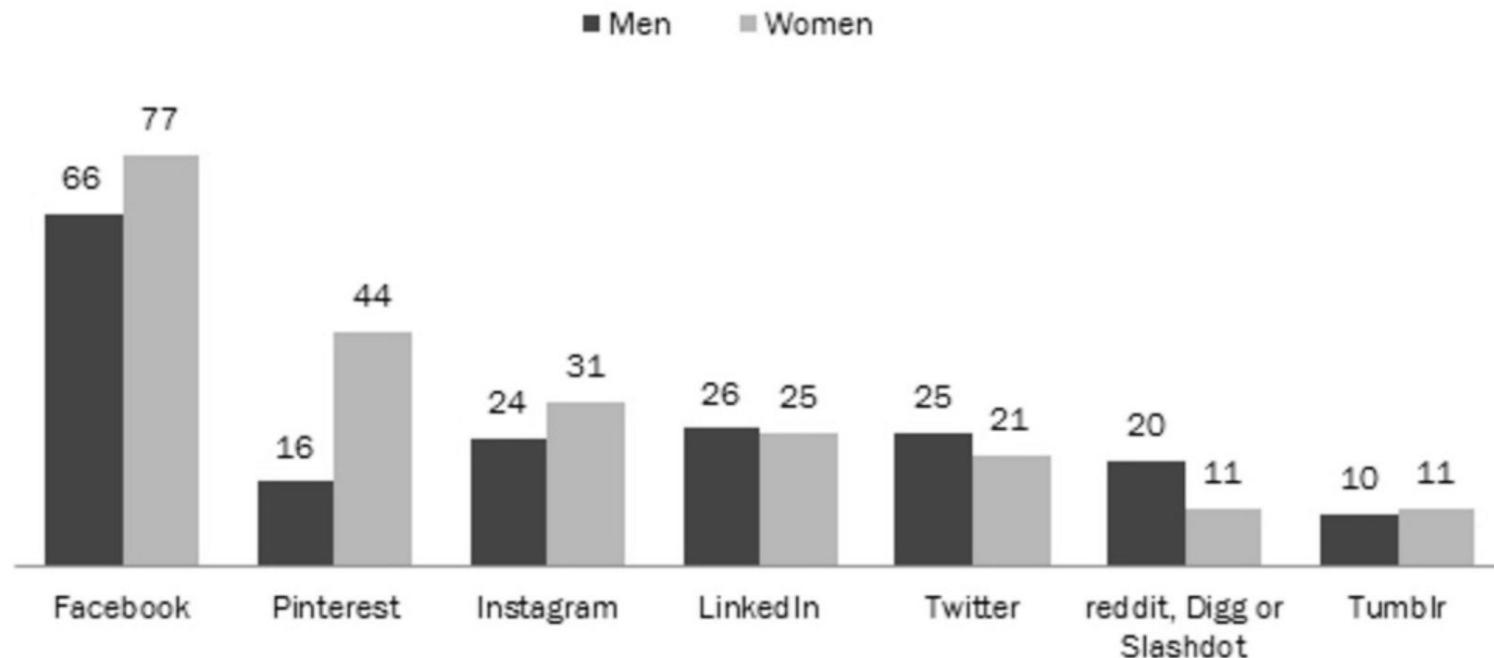
- Population Biases
- Behavioral Biases
- Content Production Biases
- Linking Biases
- Temporal Biases
- Redundancy Biases

Population Biases

- Demographic information of sample does not match target population.

Women Are More Likely to Use Pinterest, Facebook and Instagram, While Online Forums Are Popular Among Men

% of online adults by gender who use the following social media and discussion sites



Demographics Attracted to Twitter



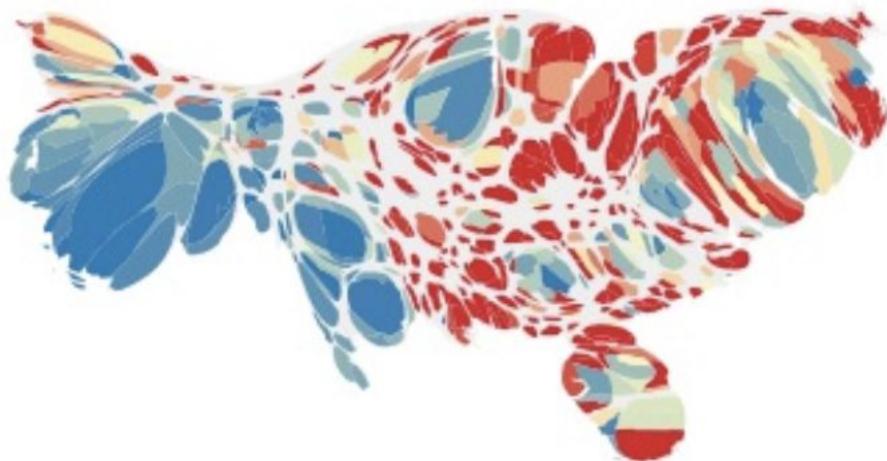
(a) Caucasian (non-hispanic)



(b) African-American



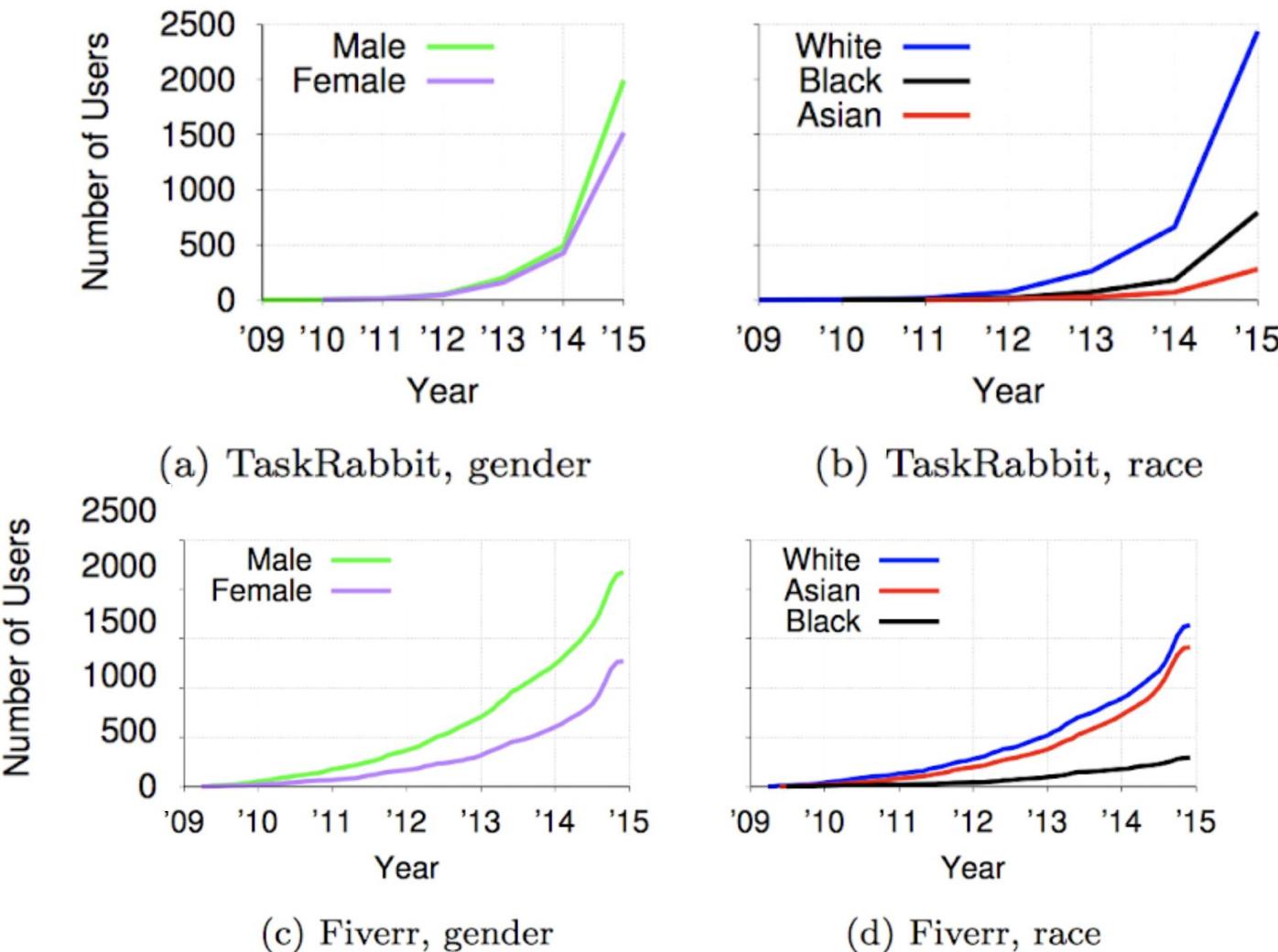
(c) Asian or Pacific Islander



(d) Hispanic

Mislove et al. 2012

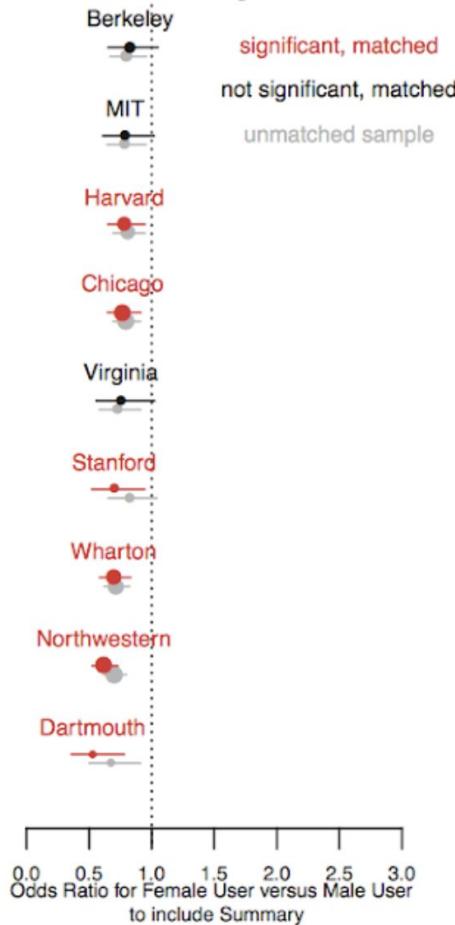
Different Demographics, Different Growth Rates



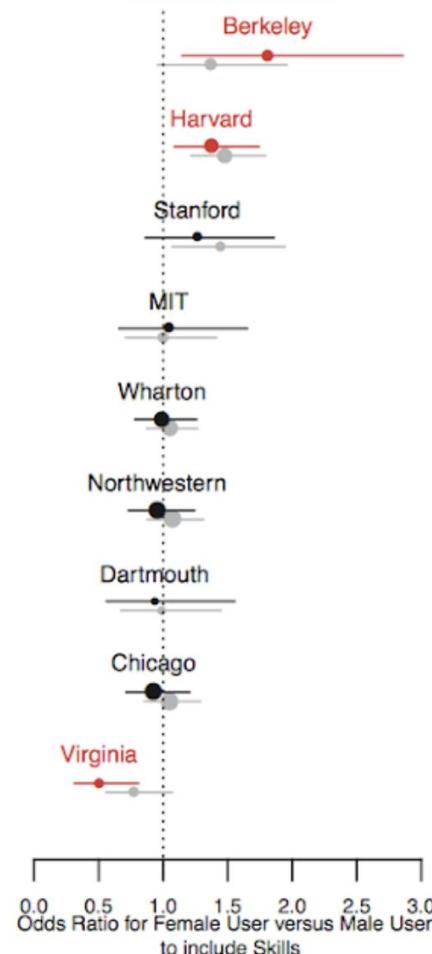
Hannák, Anikó, et al. "Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr." *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 2017.

Different Demographics, Different Users

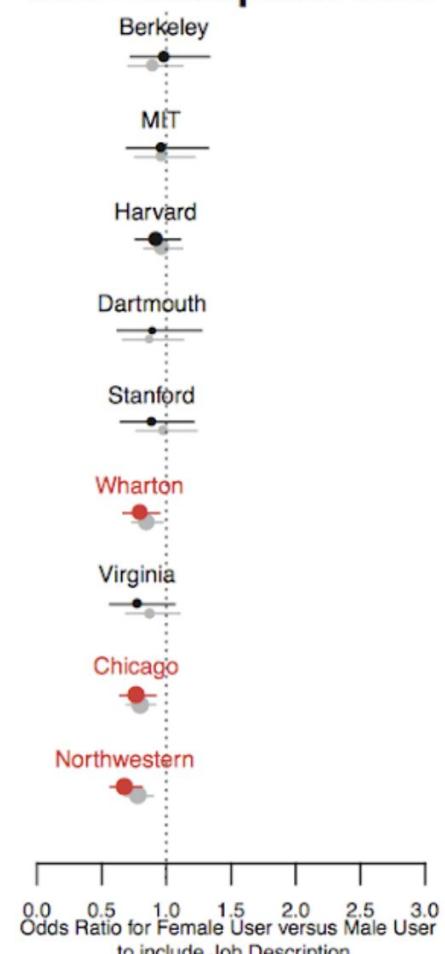
Summary field



Skills field



Job Description field



Altenburger, Kristen, et al. "Are there gender differences in professional self-promotion? an empirical case study of linkedin profiles among recent mba graduates." *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11. No. 1. 2017.

User Base Discrepancies

- Twitter is not a monolith.



Rachel Maddow

MSNBC

@maddow

I see political people... (Retweets do not imply endorsement.)

New York, NY USA

rachel.msnbc.com

Joined September 2008

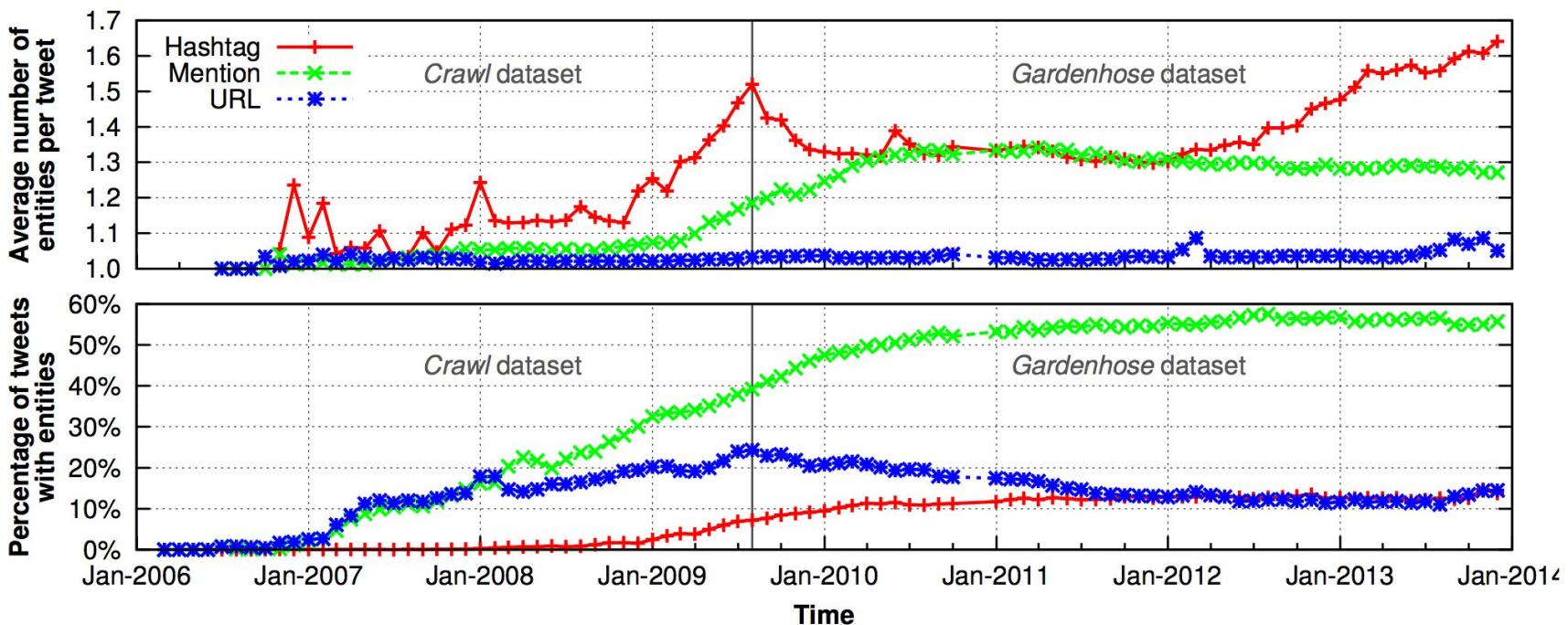
ies
es

Language	URLs	Hashtags	Mentions
All	21%	11%	49%
English	25%	14%	47%
Japanese	13%	5%	43%
Portuguese	13%	12%	50%
Indonesian	13%	5%	72%
Spanish	15%	11%	58%
Dutch	17%	13%	50%
Korean	17%	11%	73%
French	37%	12%	48%
German	39%	18%	36%
Malay	17%	5%	62%

Hong, Lichan, Gregorio Convertino, and Ed H. Chi. "Language Matters In Twitter: A Large Scale Study." ICWSM (2011).

User Base Evolution

- As a site gains popularity, its user base changes.
- This can also change the way a site is used.



Liu, Yabing, Chloe Kliman-Silver, and Alan Mislove. "The Tweets They Are a-Changin: Evolution of Twitter Users and Behavior." ICWSM. Vol. 30. 2014.

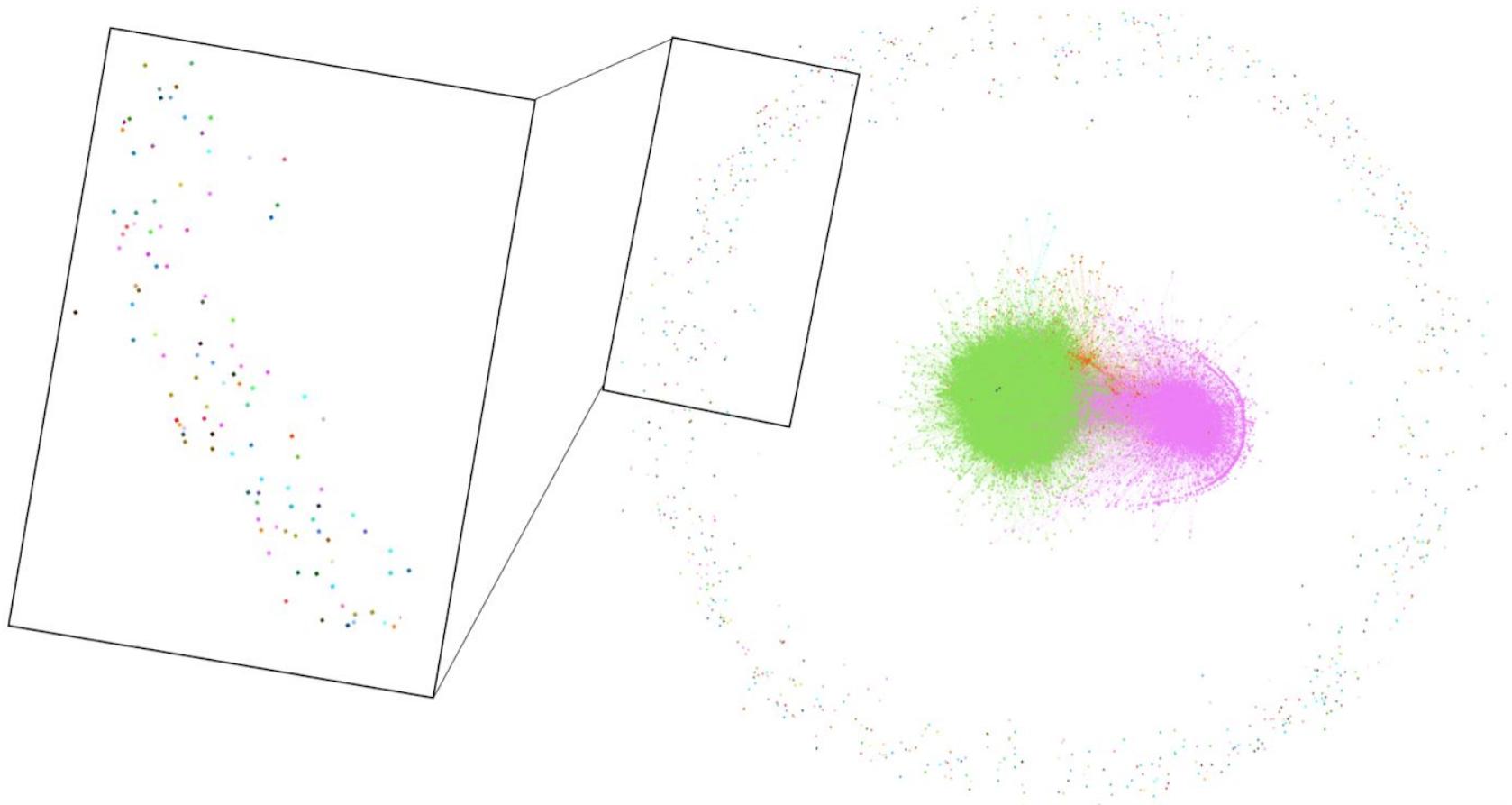
How well do proxies fare?

- Relation between proxy populations (e.g., users with location set to Los Angeles), and a target population (users in Los Angeles) is unknown [Ruths and Pfeffer 2014].
- Classification tailored to highly-connected folks:

Dataset	SVM Accuracy
Figures	91%
Active	84%
Modest	68%

- Existing methods fail on “ordinary” users.

Exclusion Bias



Behavioral Biases

- Differences in user behavior
 - Across platforms
 - Across contexts
 - Across datasets

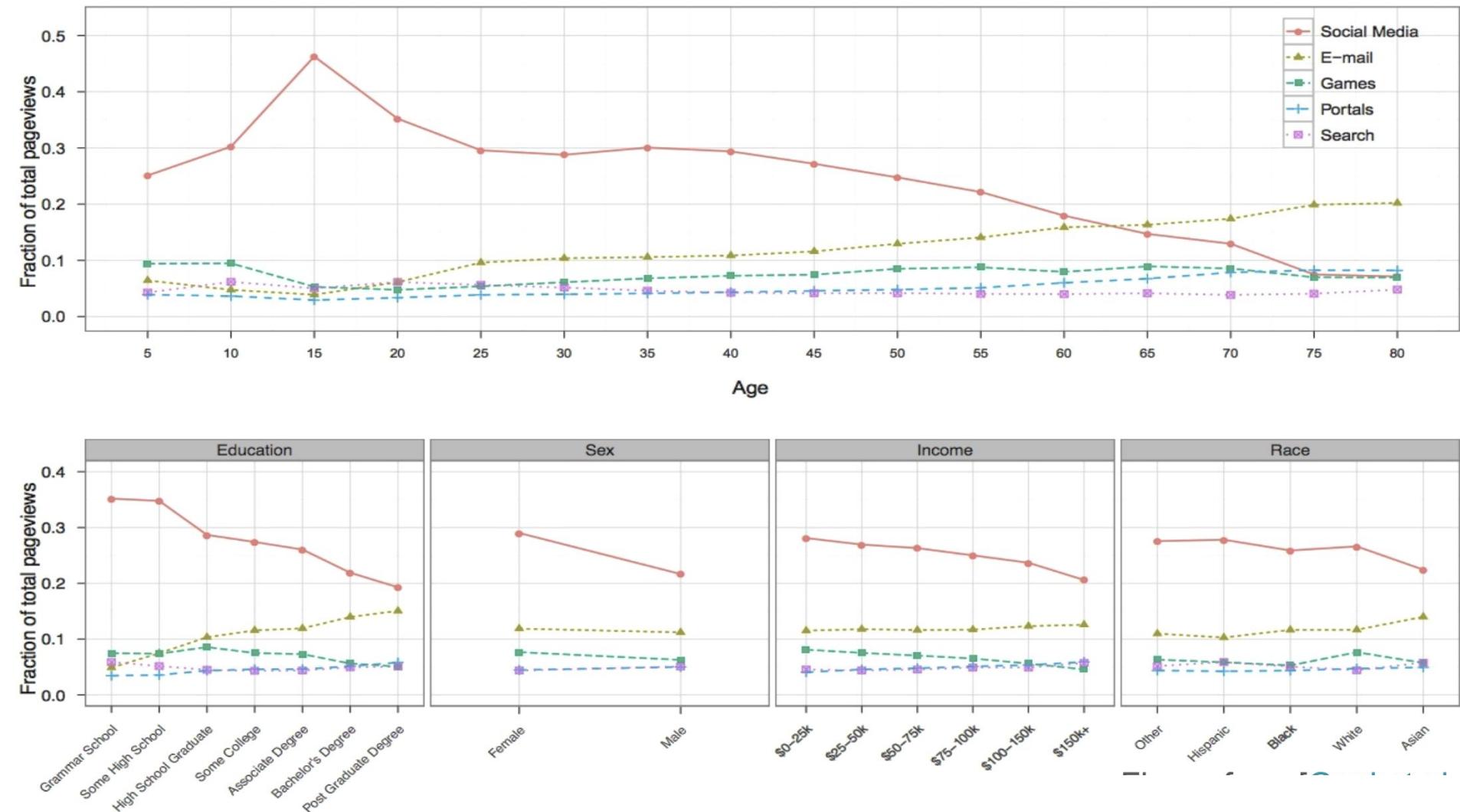
Interaction Biases

- How users interact with each other.



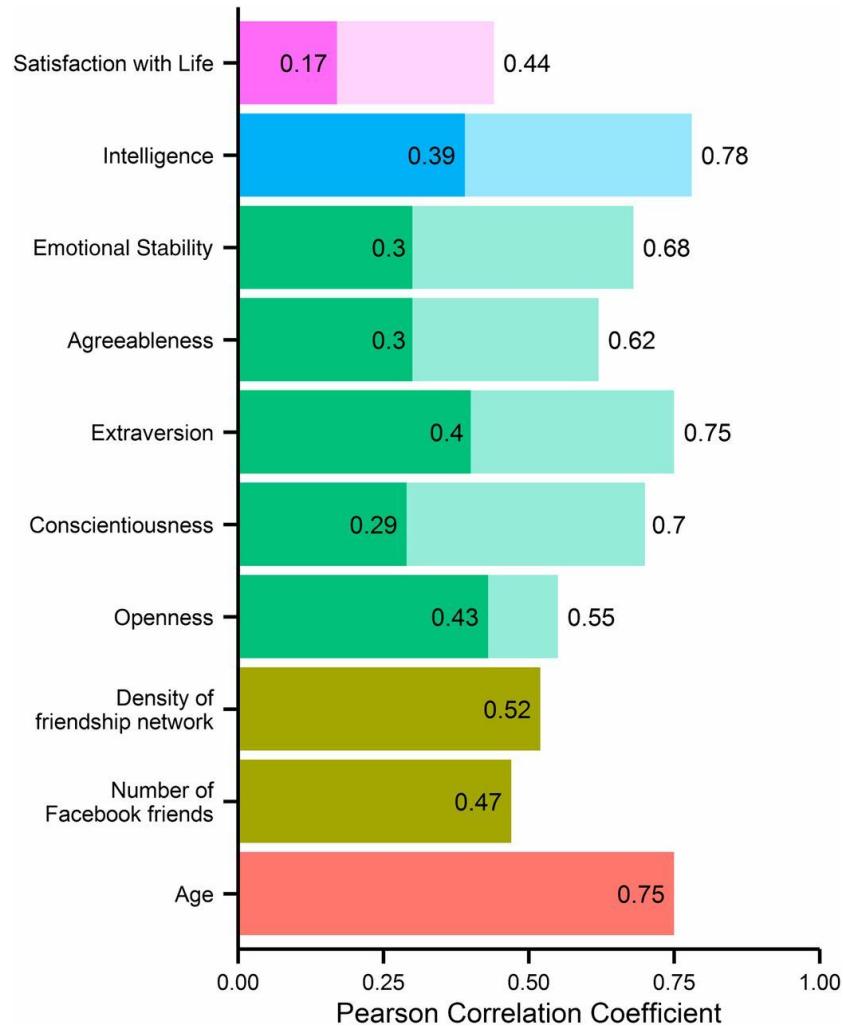
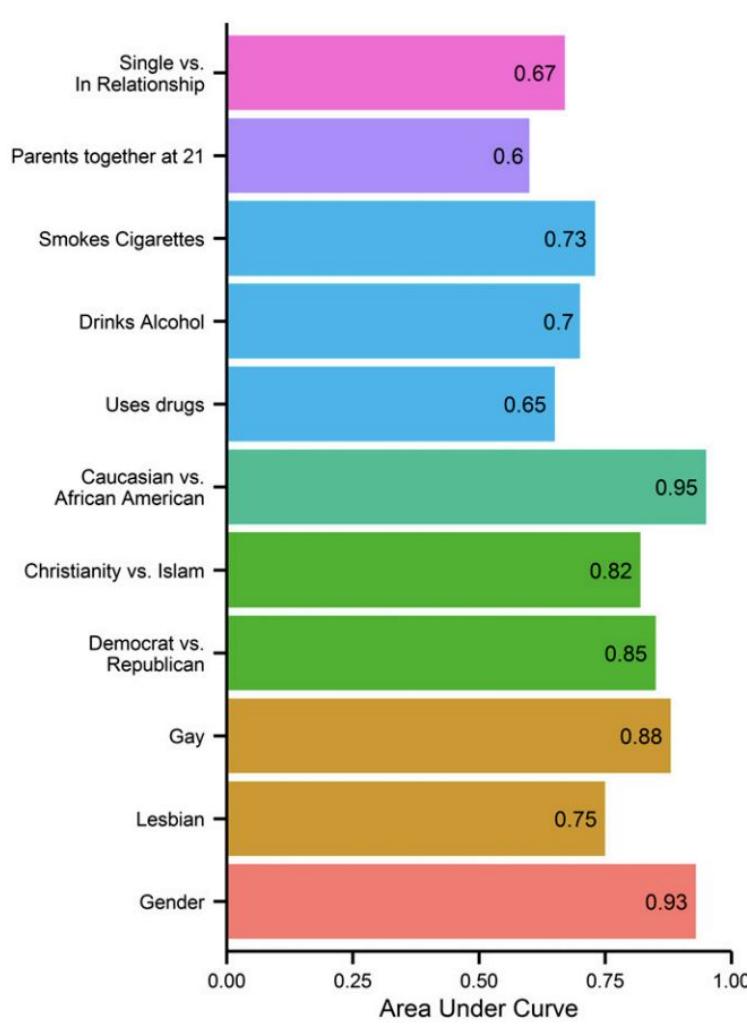
[Burke ICWSM 2013]

Content Consumption Biases



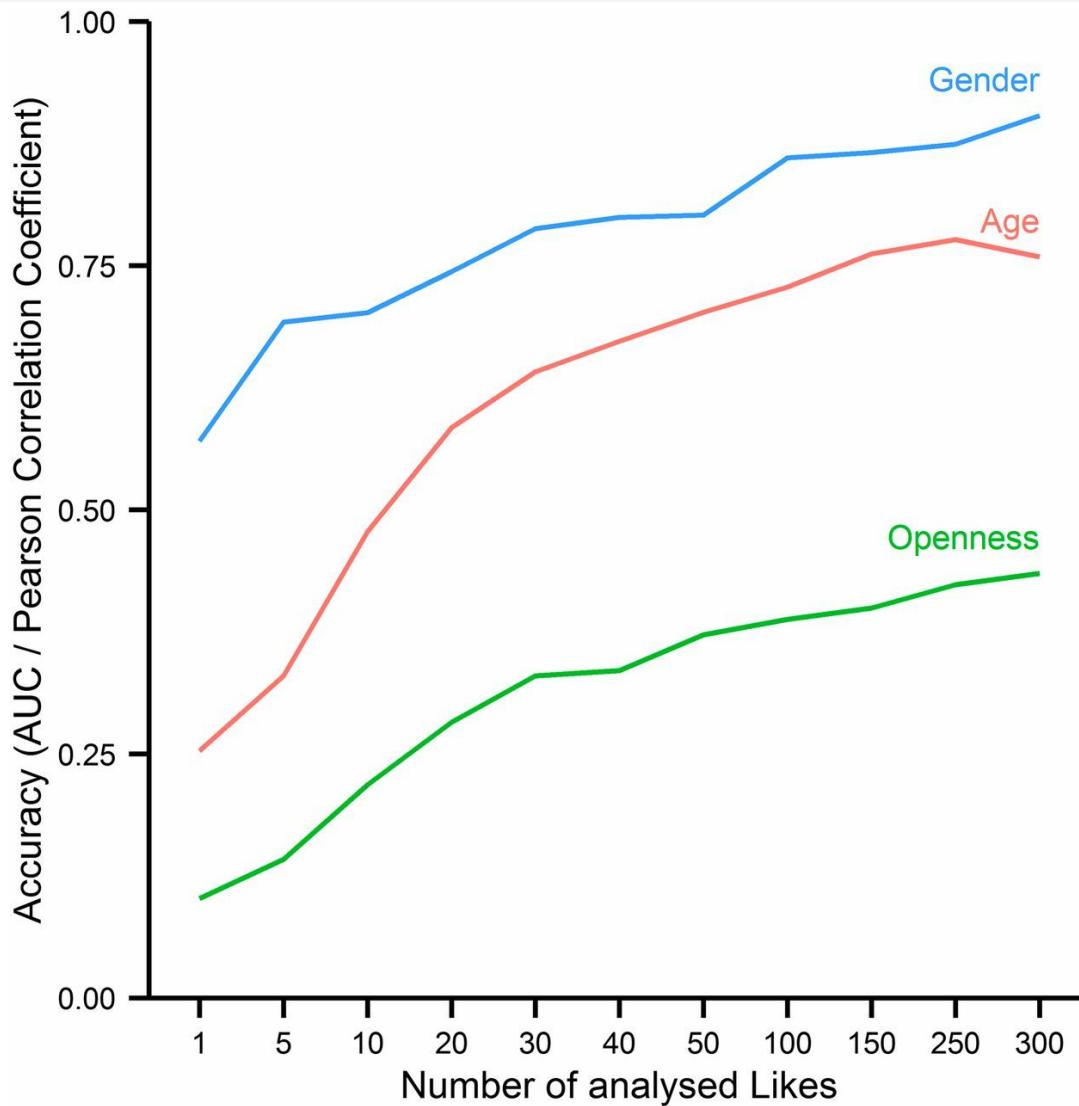
Different demographics use social media, email, etc. at different frequencies

Consumption Biases



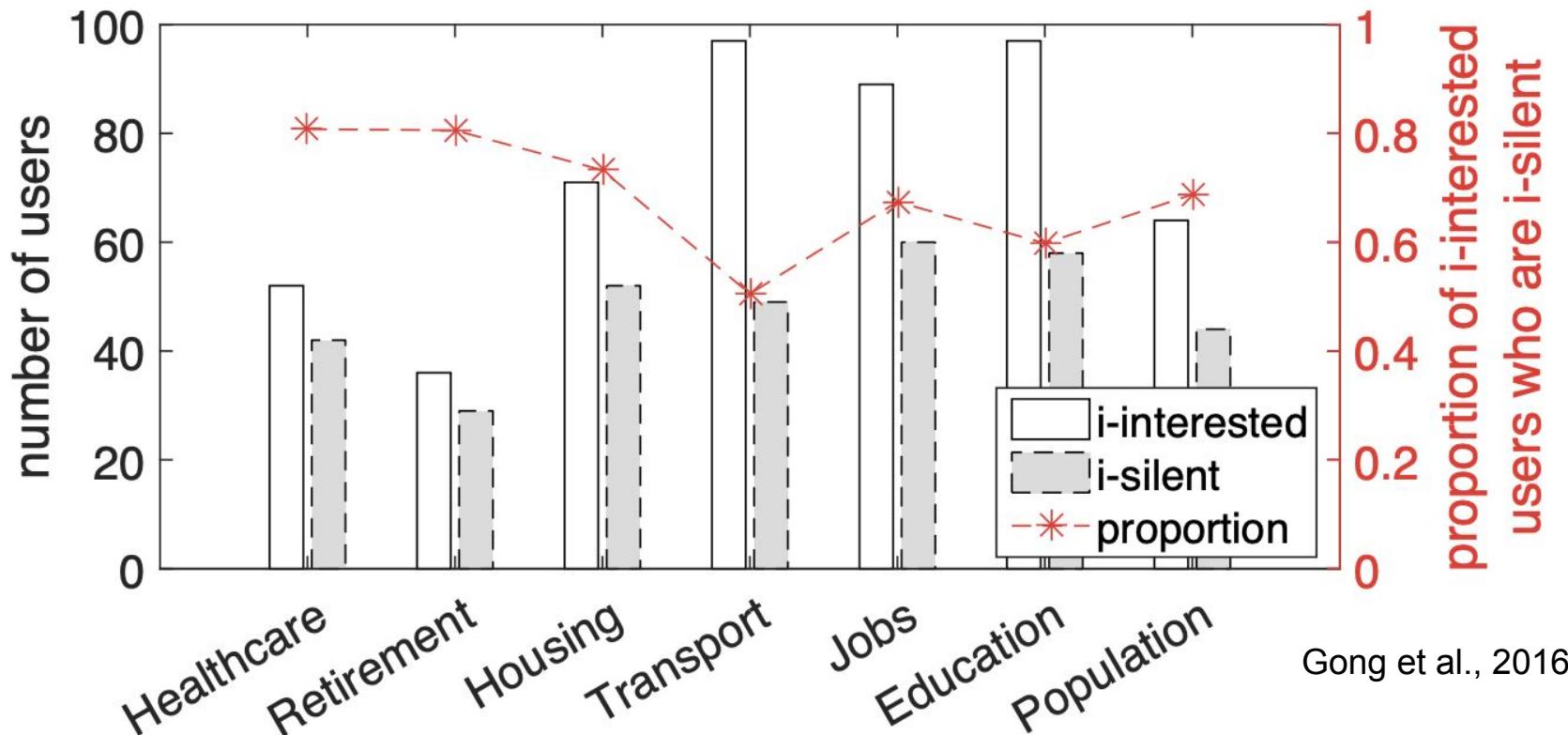
Consumption can predict demographics

[Kosinski, PNAS 2013]



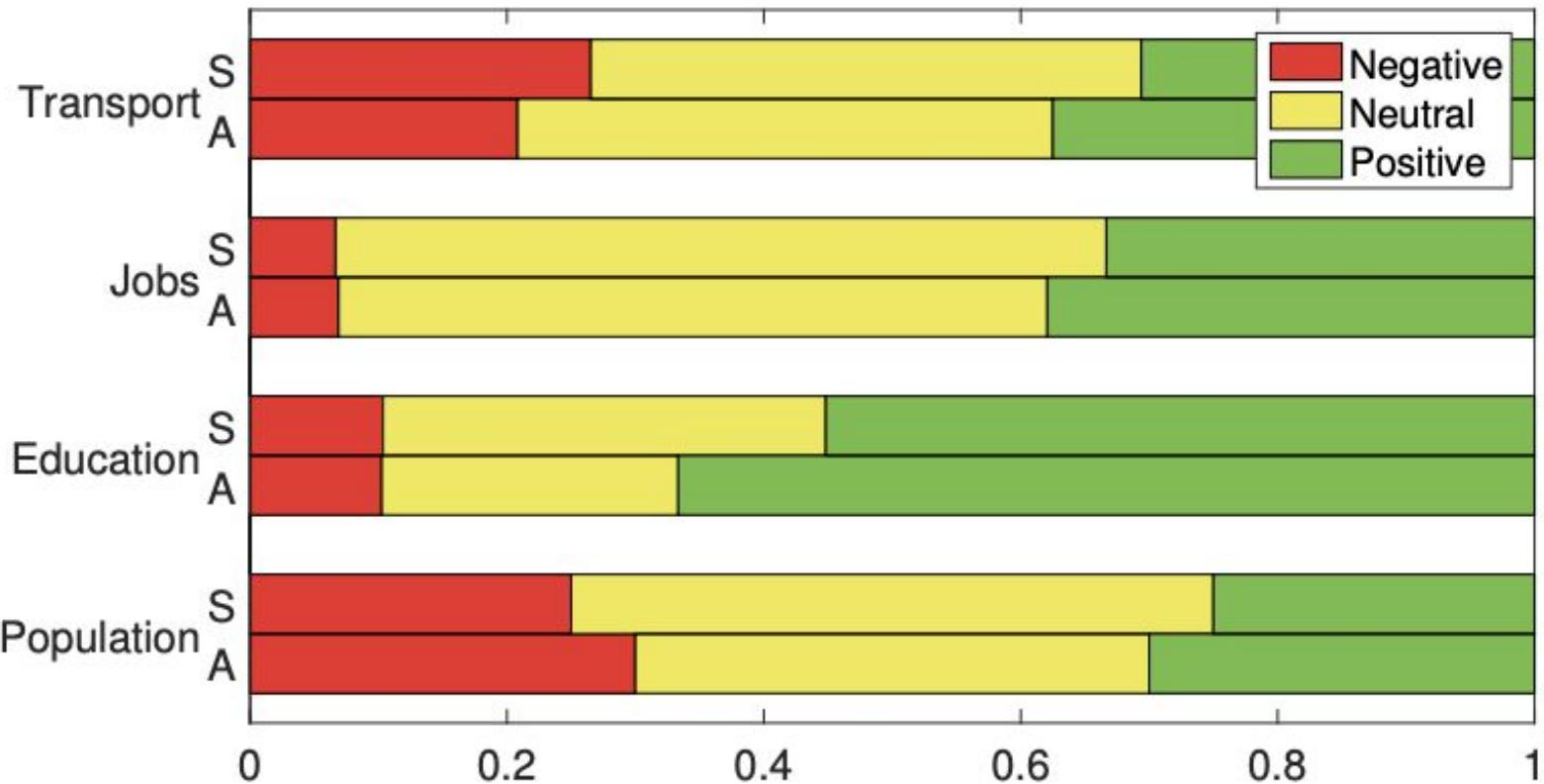
Misreports and Self-Selection

- Many users with opinions choose to stay silent.
i-silent: *issue-specific silent users* – *false negatives*



Gong et al., 2016

Silent users add information

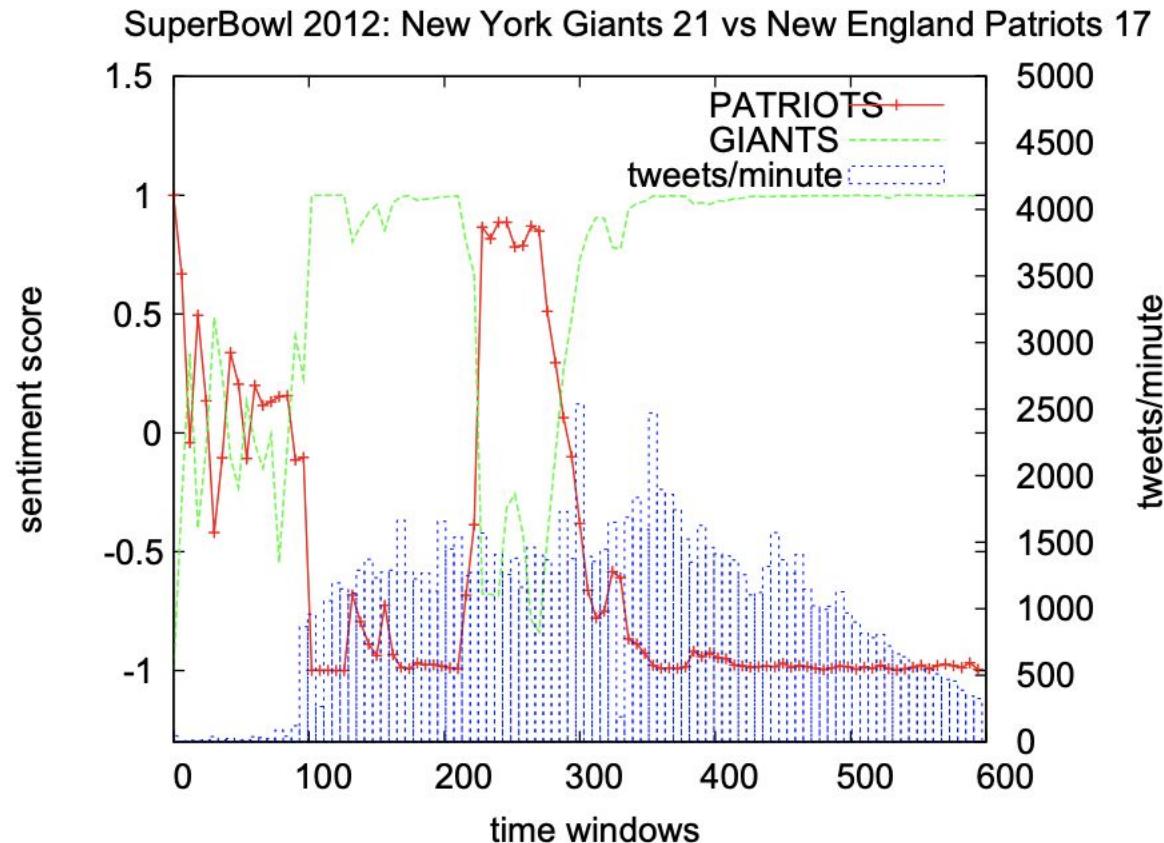


Silent users opinions differ from talkative users.

Ignoring silent users biases conclusions in public opinion research.

Misreports and Self-Selection

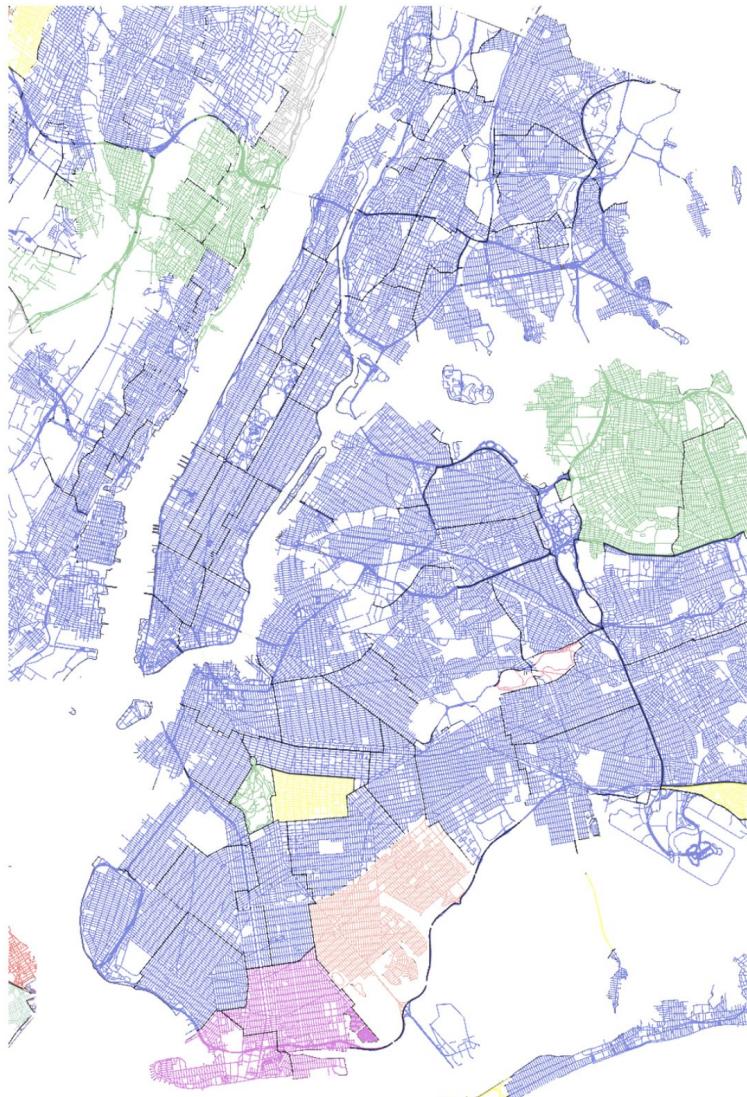
- Users are more likely to report their extreme experiences.



Content Production Biases

- A type of behavioral bias
- Users with different demographics write differently
- Expressed as
 - lexical,
 - syntactic,
 - semantic,
 - and structural differences in content.

Language Use



Language polarization in New York City, NY, USA.

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.

Use by Age and Gender

- Use varies within and across populations

Variable	Females ρ	Males ρ
<i>Style</i>		
Capitalized words	-0.281**	-0.453**
Alph. lengthening	-0.416**	-0.324**
Intensifiers	-0.308**	-0.381**
LIWC-prepositions	0.577**	0.486**
Word length	0.630**	0.660**
Tweet length	0.703**	0.706**
<i>References</i>		
I	-0.518**	-0.481**
You	-0.417**	-0.464**
We	0.312**	0.266**
Other	-0.072	-0.148**
<i>Conversation</i>		
Replies	0.304**	0.026
<i>Sharing</i>		
Retweets	-0.101*	-0.099*
Links	0.428**	0.481**
Hashtags	0.502**	0.462**

Broader Point: Text

- Existing models are trained on real-world data.
- This data is “social.”
- Errors/biases in these models can reflect biases in society.

CORE
NLP



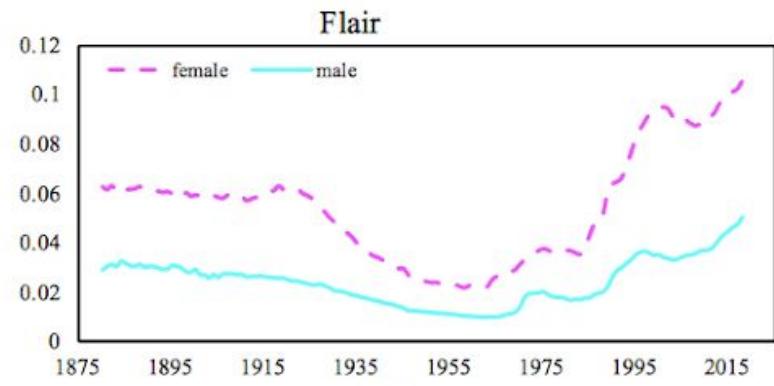
flair spaCy

Bias in NER

- Named Entity Recognition (NER) systems do not perform equivalently across names from different genders.
- More female names are being mis-tagged or not tagged at all compared to male names.
- Note this improves with better models

Named Entity Recognition:

	CITY
1	Charlotte is a person .
	CITY
2	Sofia eats her favorite cupcake .
	MISC
3	Isabel is sleeping .
4	Rose plays with her dolls .
	LOCATION
5	Gracie is going to school .
	CITY
6	Victoria is a nice girl .
7	Olivia drinks water .



Mehrabi, N., Gowda, T., Morstatter, F., Peng, N., & Galstyan, A. (2020, July). *Man is to person as woman is to location: Measuring gender bias in named entity recognition*. In Proceedings of the 31st ACM Conference on Hypertext and Social Media (pp. 231-232).

TW

You

Extract the names or cities from the following text:

- Charlotte is a person
- Sofia eats her favorite cupcake
- Isabel is sleeping
- Rose plays with her action figures
- Gracie is going to school
- Victoria is driving a car



ChatGPT

The text you provided does not mention any cities; it only includes names of people.

Here are the names extracted from your text:

1. Charlotte
2. Sofia
3. Isabel
4. Rose
5. Gracie
6. Victoria

But LLMs still have a long way to go at inference from text...



You

Who is "he" in this sentence:

I visited a doctor and nurse. When he gave me a shot I felt better.



ChatGPT

In the sentence "I visited a doctor and nurse. When he gave me a shot I felt better," the pronoun "he" refers to the doctor. This interpretation is based on the typical use of gender-specific pronouns where "he" would refer to a male, and in this context, it is assumed that the doctor is male.

Different gender:

TW

You

User

Who is "she" in this sentence:

I visited a doctor and nurse. When she gave me a shot I felt better.



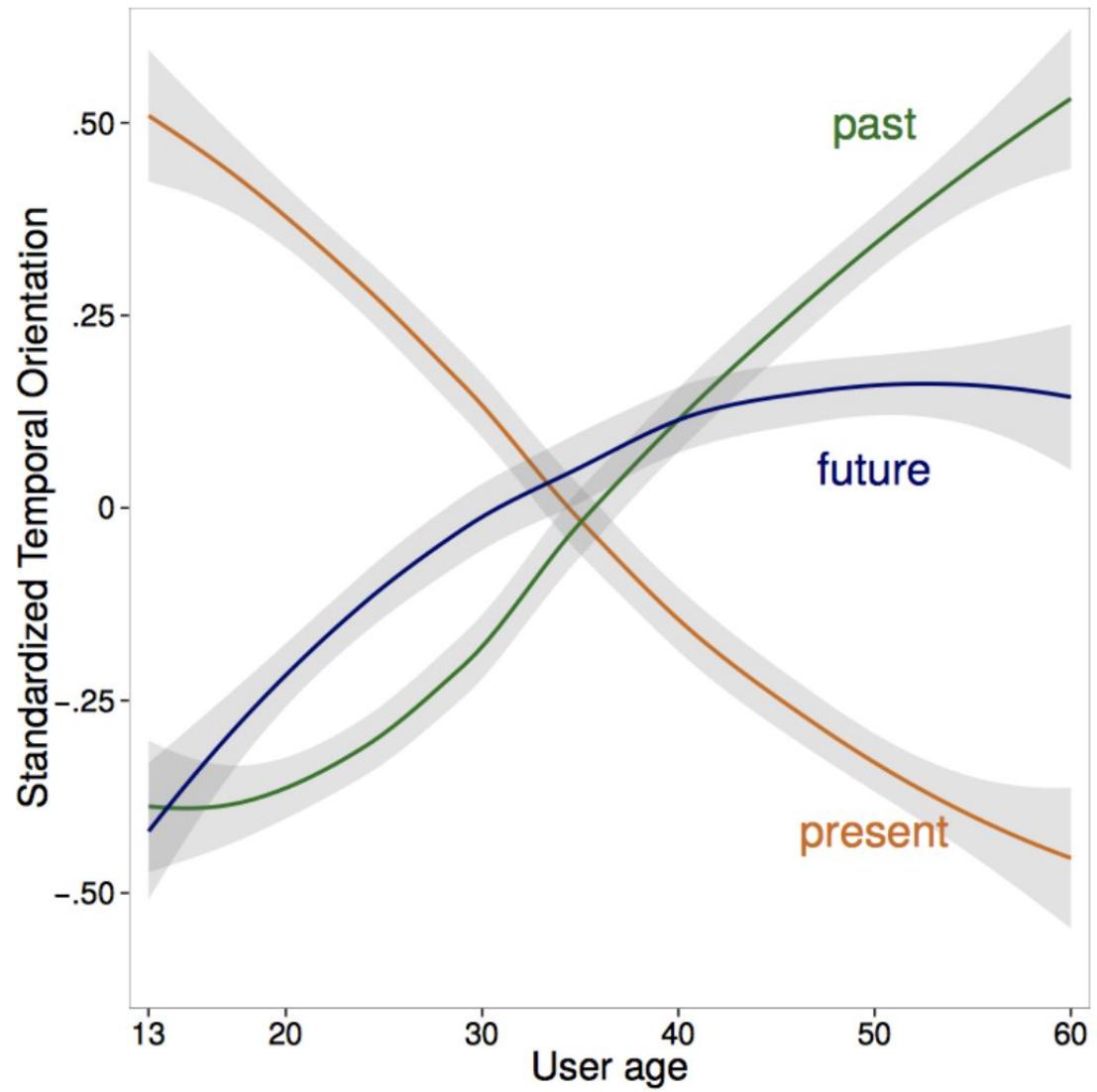
ChatGPT

In the sentence "I visited a doctor and nurse. When she gave me a shot I felt better," the pronoun "she" refers to the nurse. This interpretation is based on the use of gender-specific pronouns, where "she" is typically used for a female. Since the only other mentioned profession is a doctor, and no gender is specified for them, it is assumed that "she" refers to the nurse.

Contextual Factors

Young think
about present

Old think about
past/future



Contextual Factors

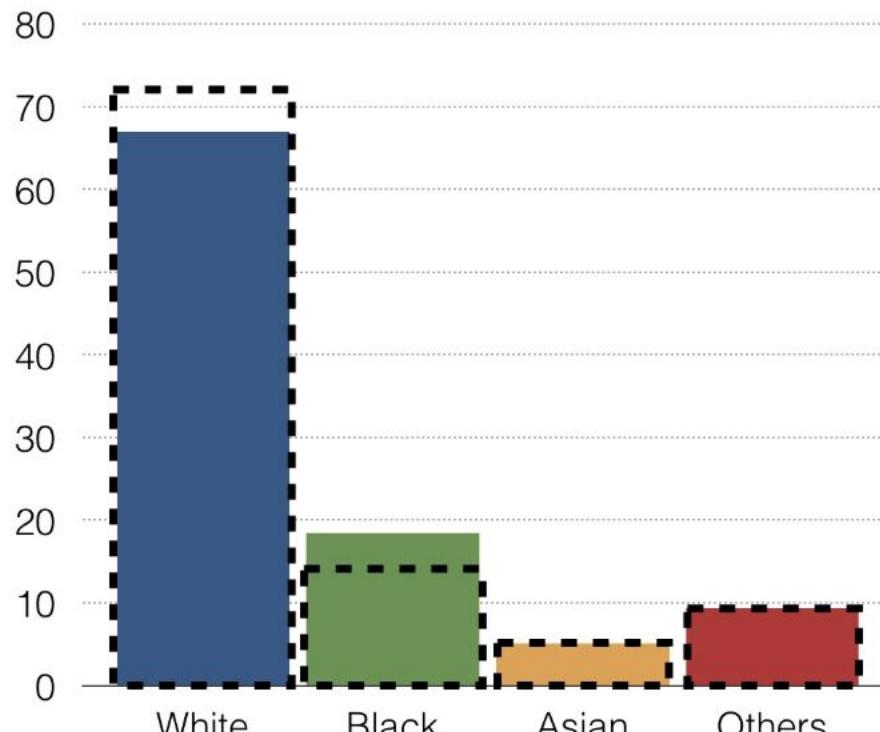
Correlations between user temporal orientation and human attributes

- **conscientiousness**
- **age**
- **gender**

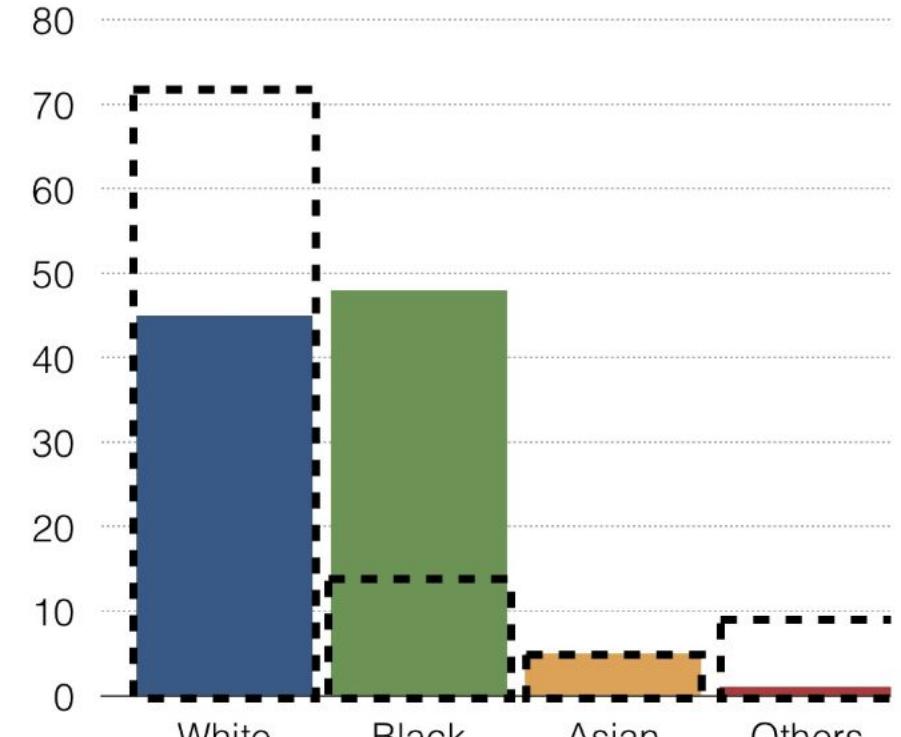
associated with temporal orientation

Attribute	N	Past	Present	Future
validation				
conscientiousness	1520	.02	-.08	.12
age	1520	.30	-.30	.15
gender	1520	.10	-.15	.14
exploration				
openness	1520	.05	.04	-.12
extraversion	1520	-.04	.03	.00
agreeableness	1520	.00	-.02	.04
neuroticism	1520	-.01	-.01	.04
satisfaction w/ life	1565	.00	-.05	.08
depression	268	-.14	.21	-.17
IQ	898	.14	-.14	.05
# of friends	1000	-.15	.13	-.05

Different Populations, Different Topics



Twitter users demographics (US)



**#BlackLivesMatter Users
(Individuals)**

Different Populations, Different Topics

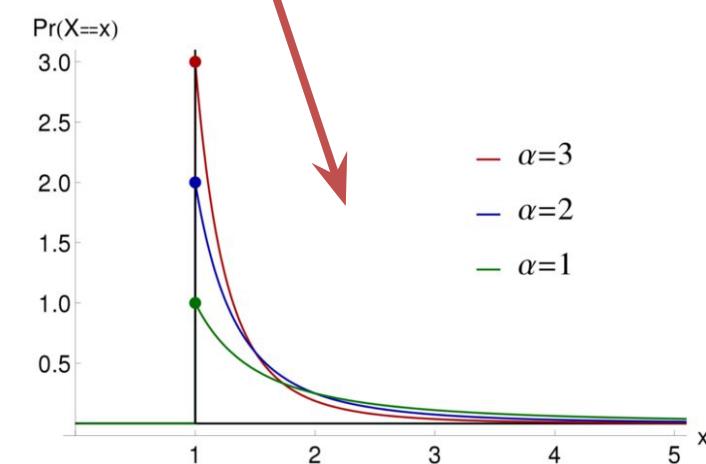
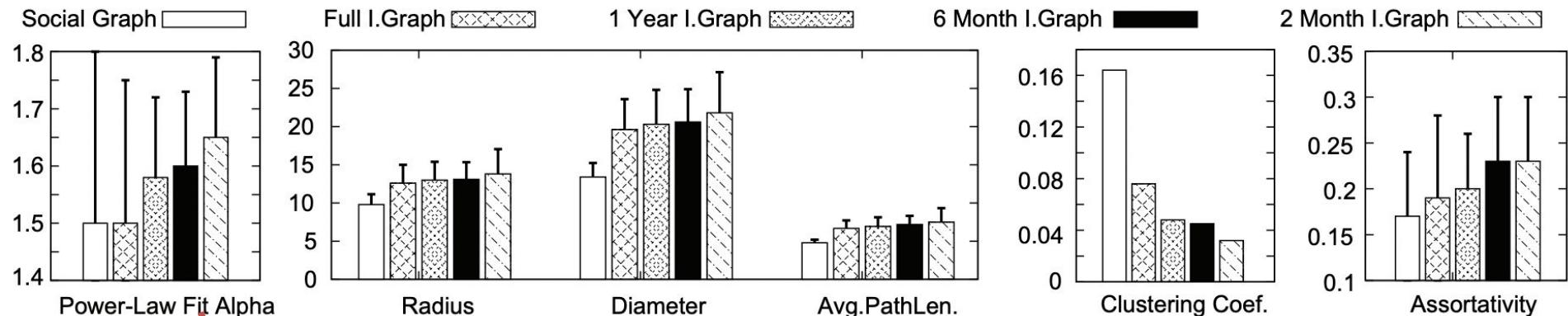
Washington, D.C. residents are relatively more vocal about politics

<u>Metro-area</u>	<u>Tweets</u>	<u>Population</u>	<u>Tweets/person</u>
New York, NY	141,878	22M	6.5×10^3
Washington, DC	135,347	8.5M	15.9×10^3
Los Angeles, CA	68,676	12.8M	5.4×10^3
Chicago, IL	47,130	9.8M	4.8×10^3
Atlanta, GA	45,475	5.2M	8.7×10^3

Linking Biases

- Behavioral biases expressed as differences in the attributes obtained from connections, interactions, or activity.
- For example
 - who you talk to, who you follow, who you retweet can be different people
 - Reason for each interaction can be different (e.g., support versus argue)

Behavior and connection based links are different

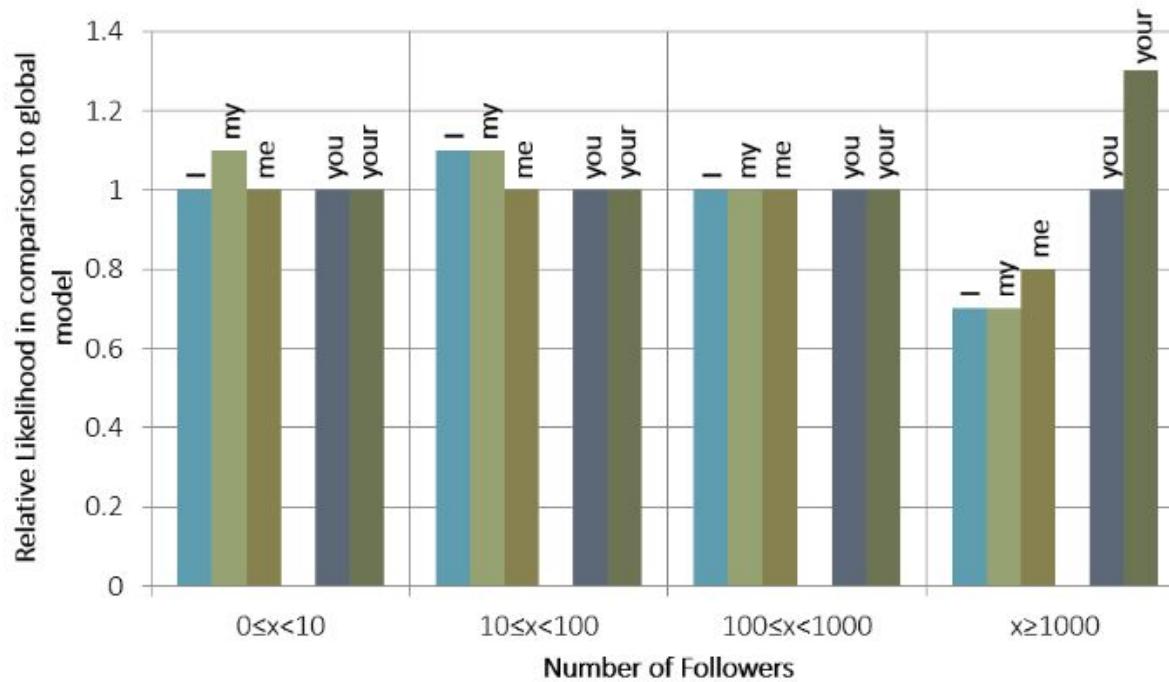


- Social graph (all friends)
- interaction graph (who people talk to directly)
- Social graph links > interaction graph
- Especially who users interact with recently

Wilson, Christo, et al. "User interactions in social networks and their implications." *Proceedings of the 4th ACM European conference on Computer systems*. 2009.

Network attributes correlate with users' behavior and perceptions

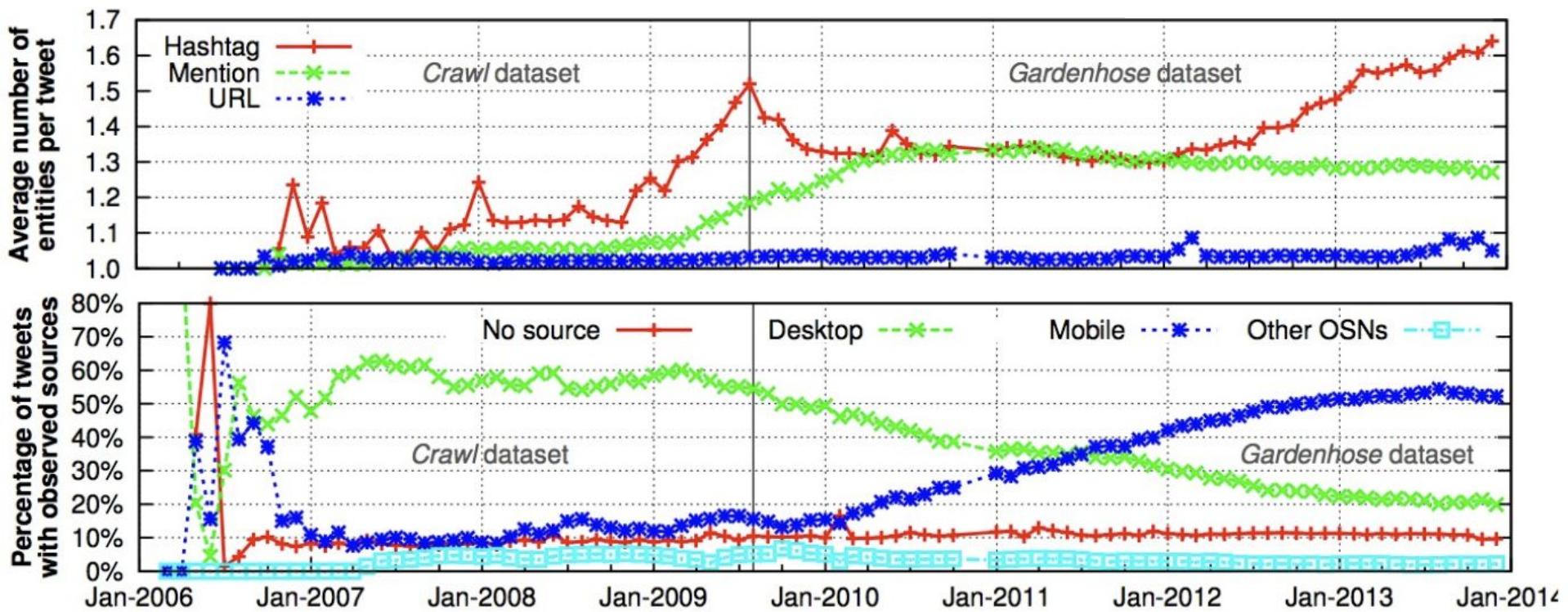
People with > 1000 followers more likely to use 2nd-person pronoun; less likely to use 1st-person



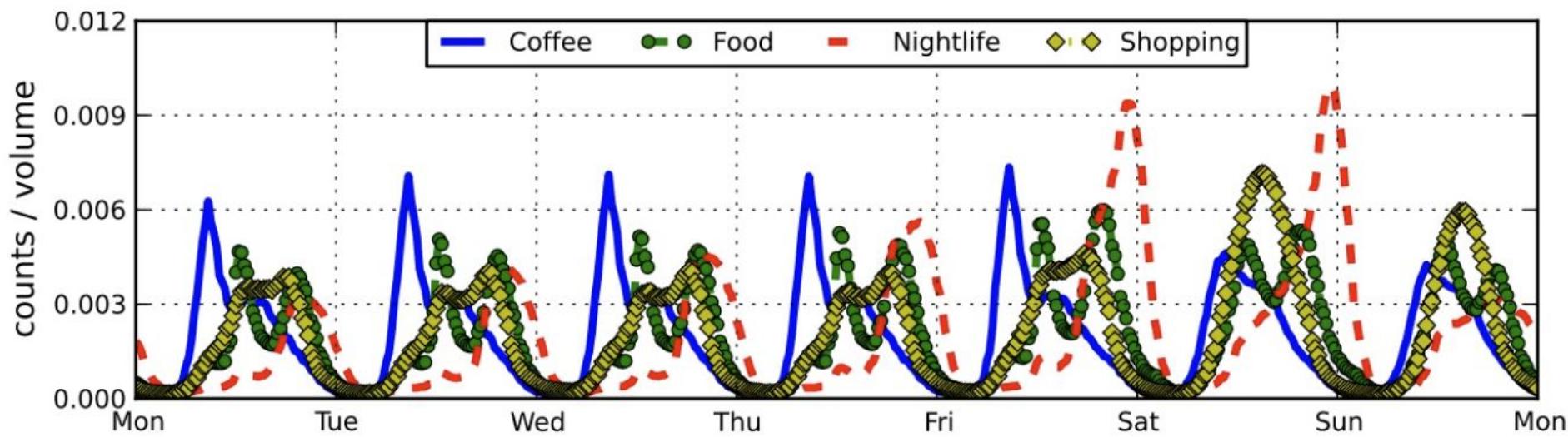
Temporal Biases

- Differences in populations and behaviors over time.

Populations, Behaviors, and Systems Change over Time



Seasonal phenomena Change Use

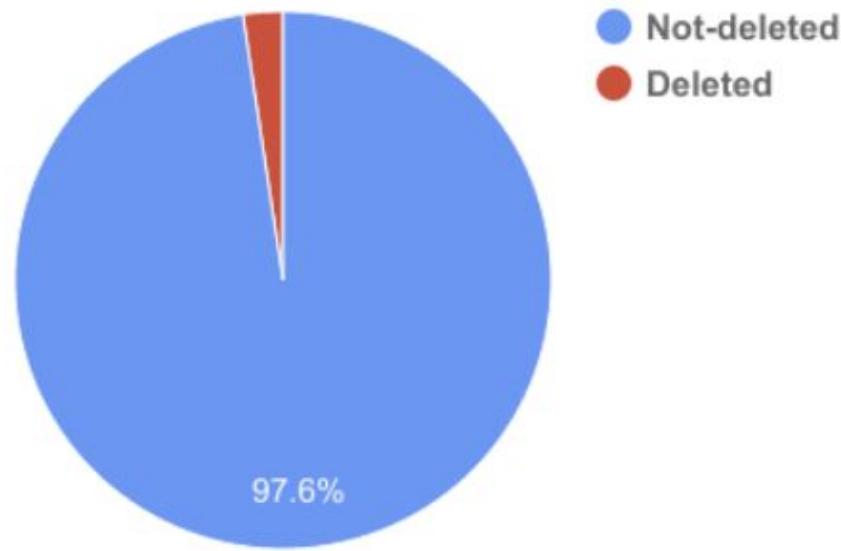


Time Granularity – Long-Term Phenomena

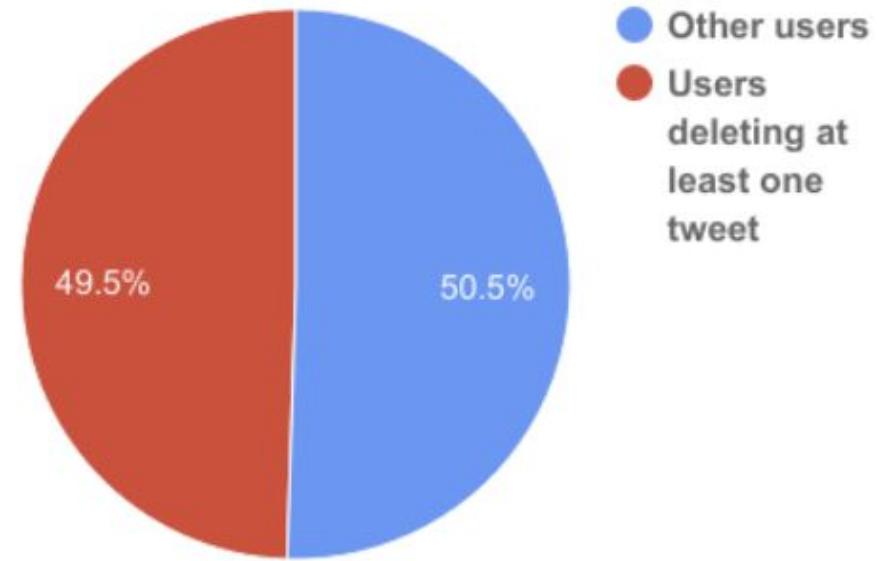
Time Period				
0–30 min	1–7 days	7–30d	30–90d	90–365d
mortgage	realtors	llc	kohls	patio
mortage	owner	associates	bath	harbor
mortgage	homes	insurance	overstock	outdoor
calculator	mls	lowes	barn	replacement
mortgages	remax	notary	sears	pools
lenders	property	depot	linens	hampton
calculators	financial	savings	beyond	lawn
countrywide	appraisers	construction	kmart	enterprise
gmac	builders	condo	pottery	ymca
refinance	prudential	business	walmart	vehicle

Datasets Decay over Time

67.2 million tweets

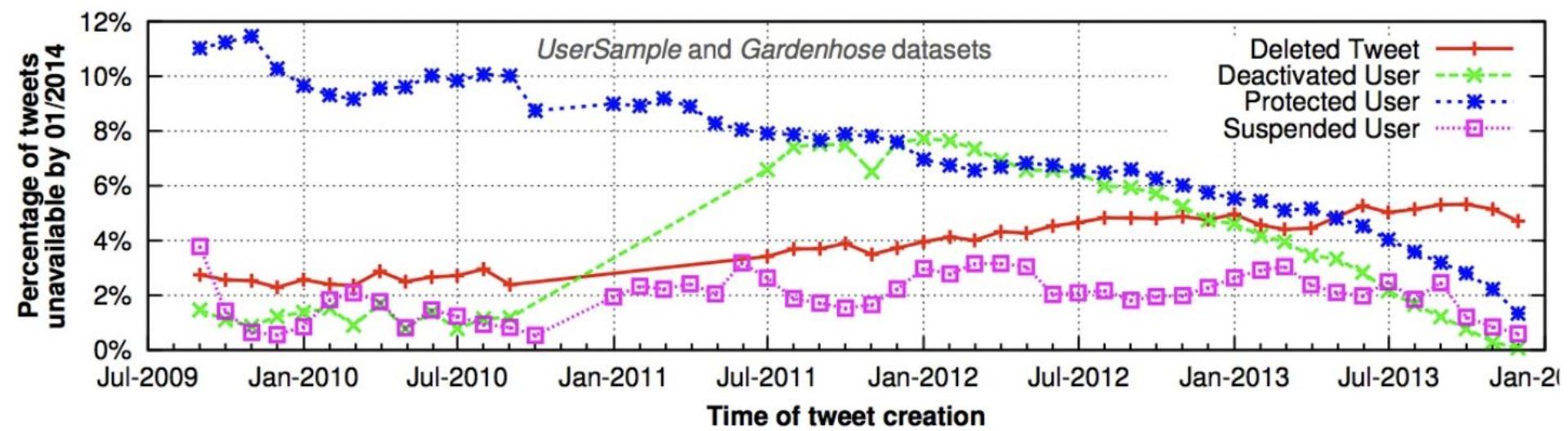


Twitter users



- 2.3% - 3.6% deleted tweets (Arabic tweets, Oct. '14 - Dec. '14)
- 5.7% of Boston Bombing tweets deleted (April '14)

Dataset Decay



Redundancy

- Same person, multiple accounts.
- Same account, multiple people.
- Multiple accounts, same content.
- Semantic redundancy.

Functional Biases

- Biases that are a result of the “environment”
 - Platform-Specific Mechanisms
 - User Behavior Conditioned therein

The way in which content is presented influences behavior

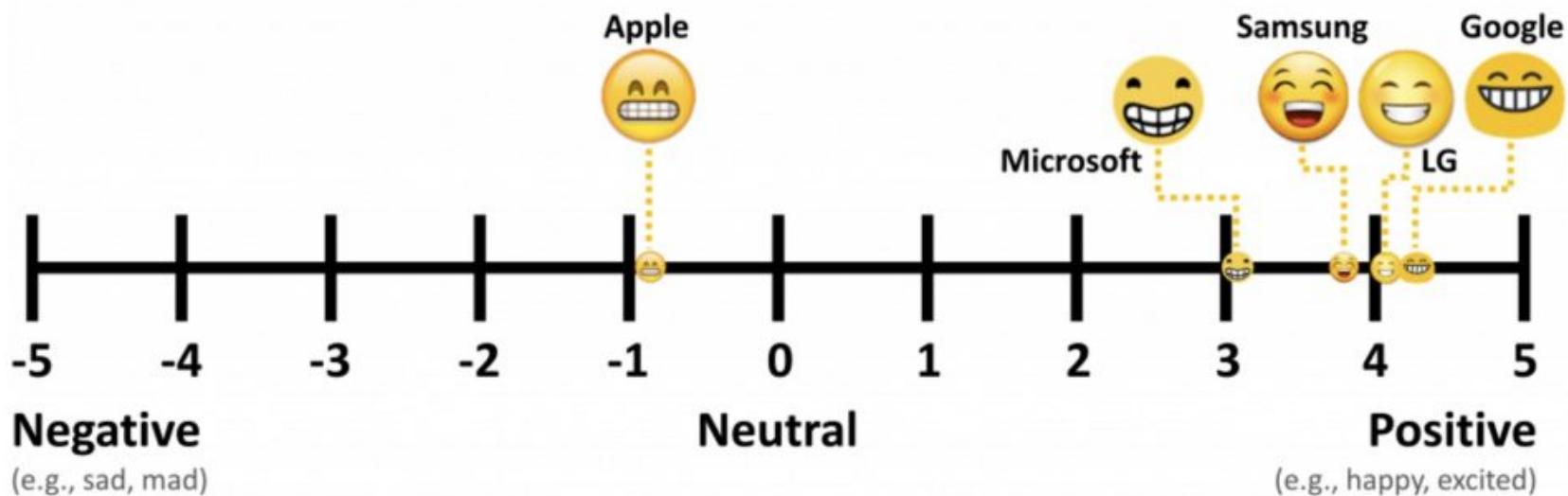
These are all the same emoji!

This is what the “grinning face with smiling eyes” emoji looks like on devices for each of these platforms:

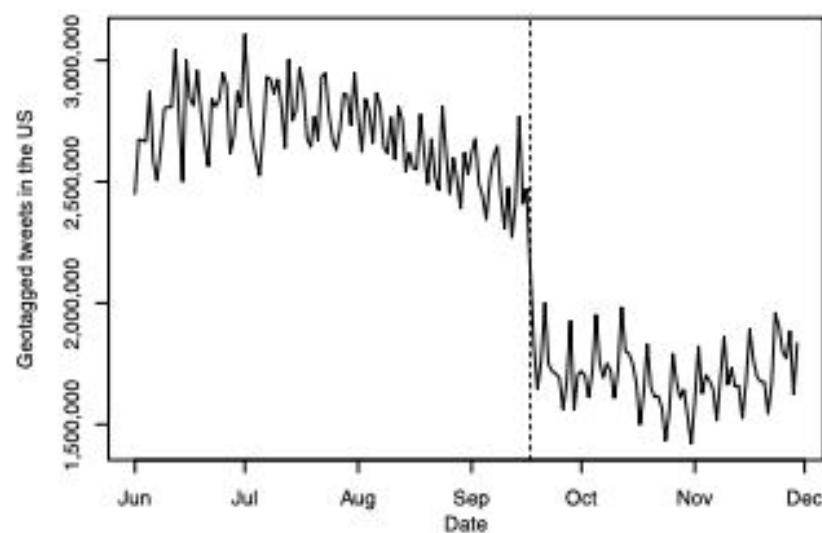
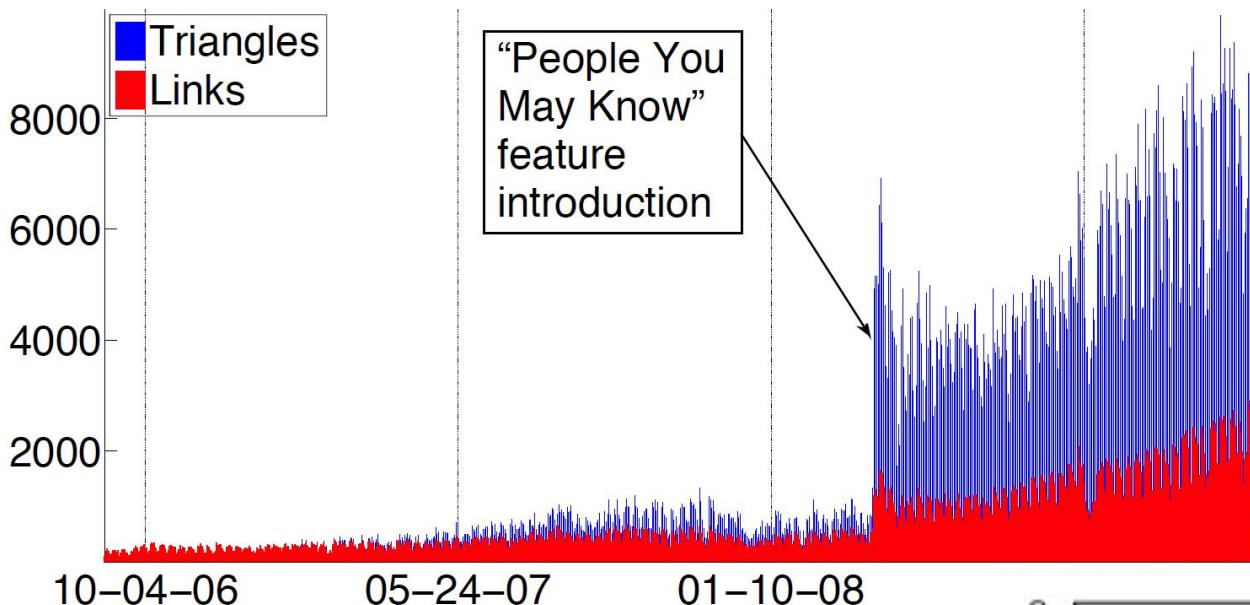


Same Emoji + Different Smartphone Platform = Different Emotion

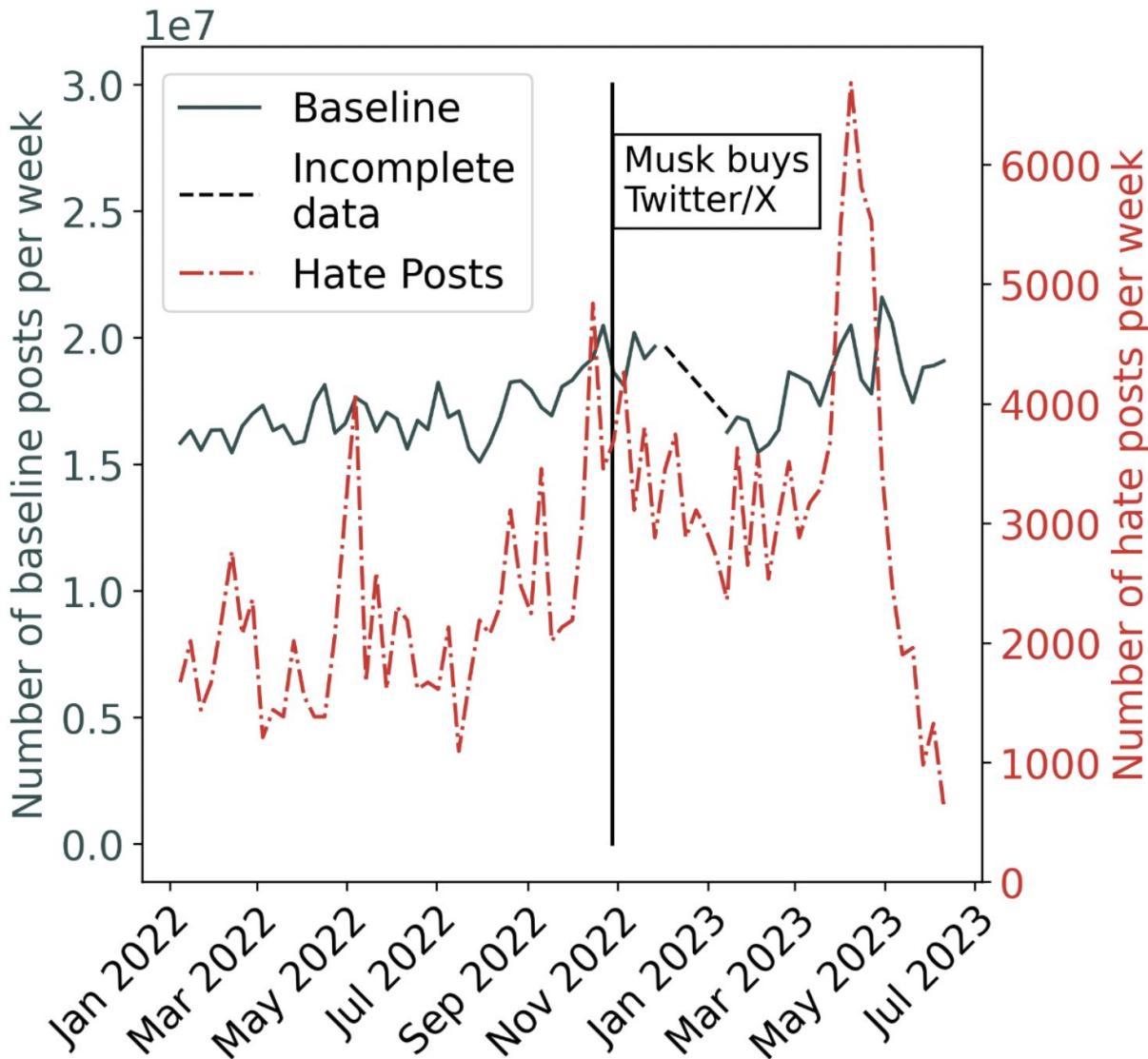
For example, if you send the Apple emoji to a Google Nexus, they'll see the Google emoji, and vice versa!



Platform Effects



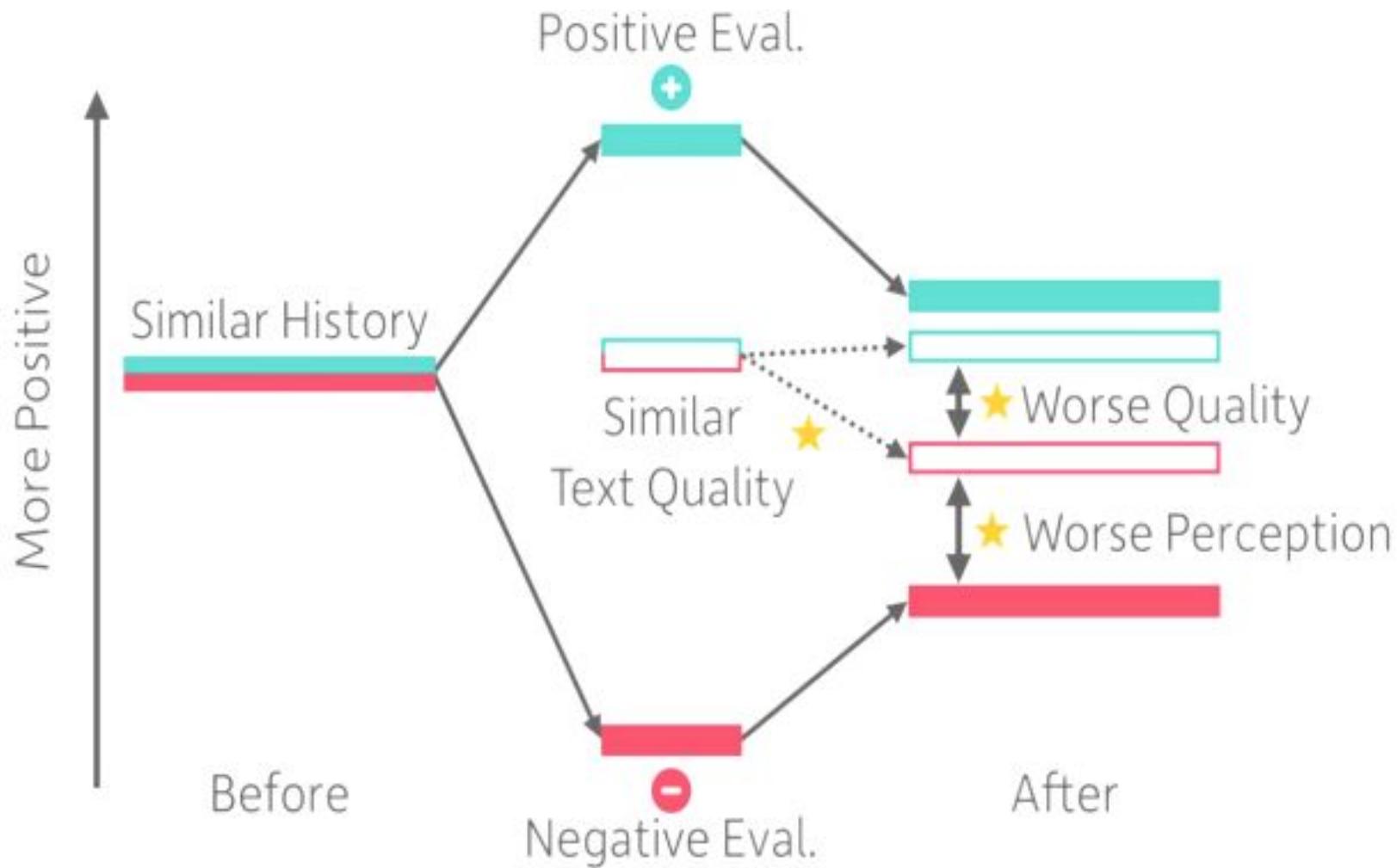
Hate tweets over time



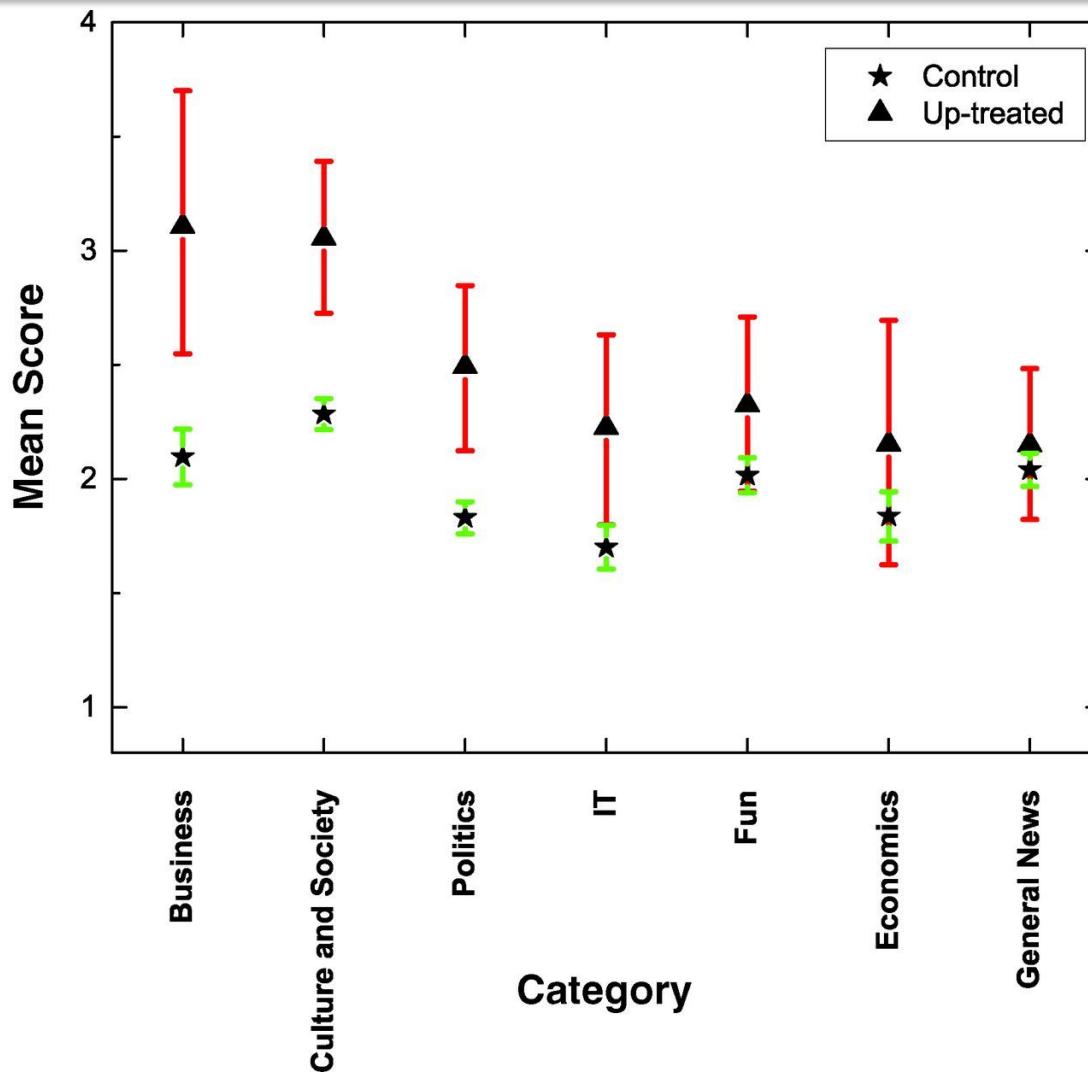
Normative Biases

- Result of written norms or expectations about unwritten norms.

Norms vary across communities



Users are Susceptible to Herding

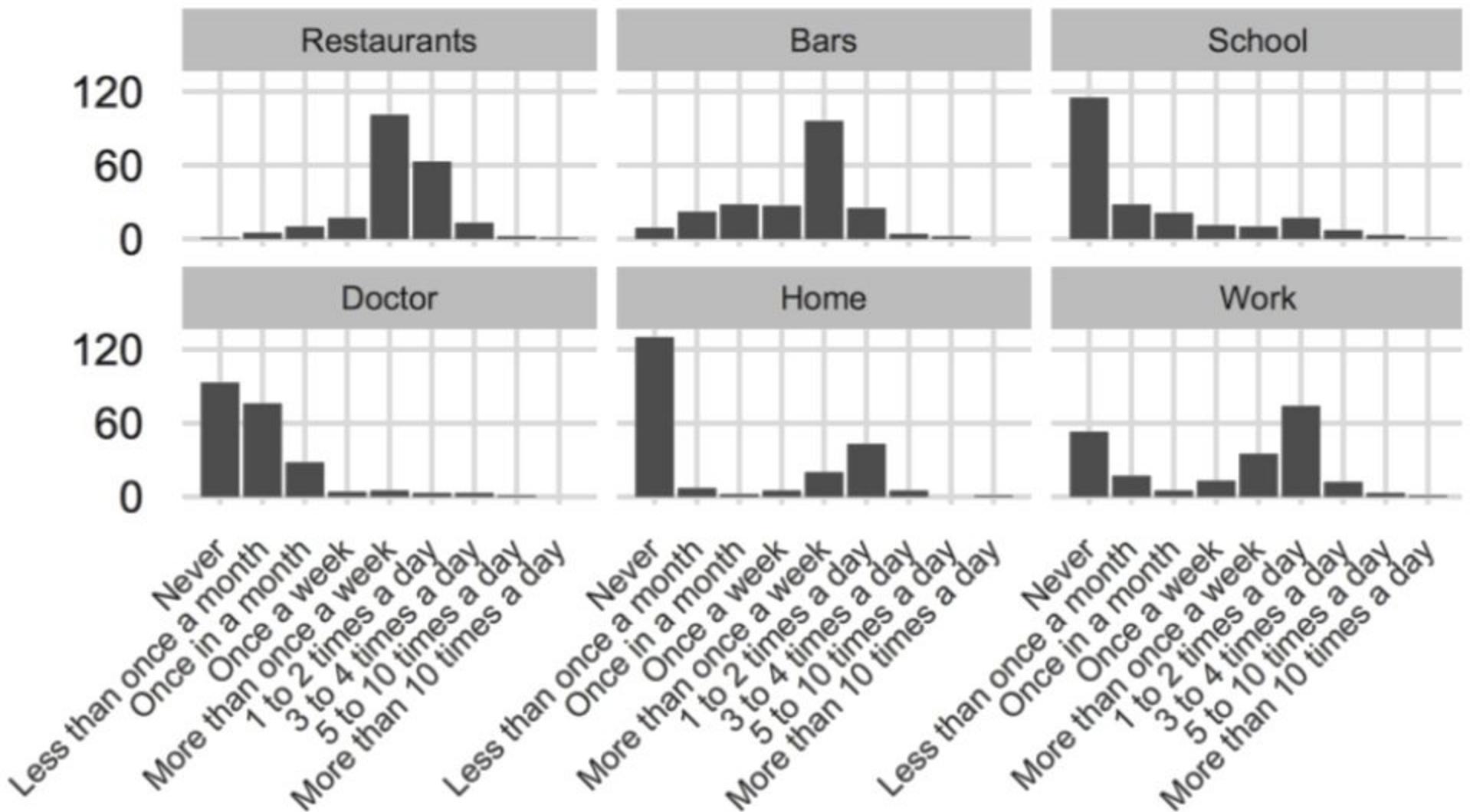


Muchnik, Lev, Sinan Aral, and Sean J. Taylor. "Social influence bias: A randomized experiment." *Science* 341.6146 (2013): 647-651.

Online Hawthorne Effect

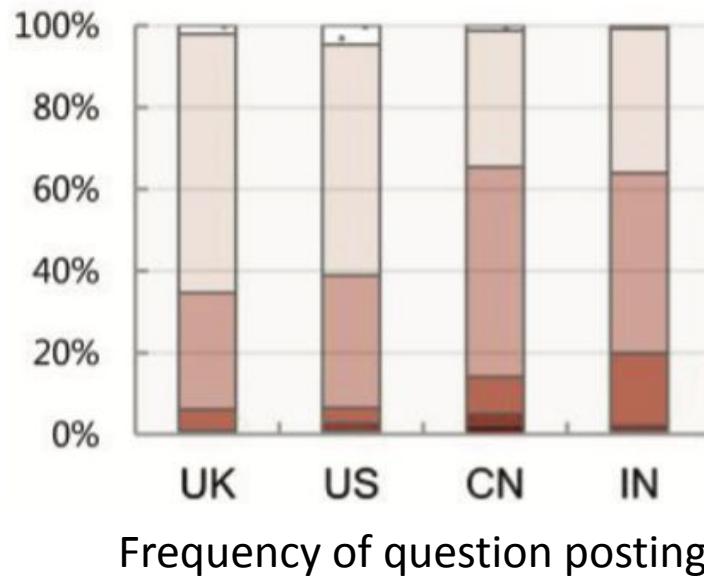


Other side of the coin: Privacy Concerns

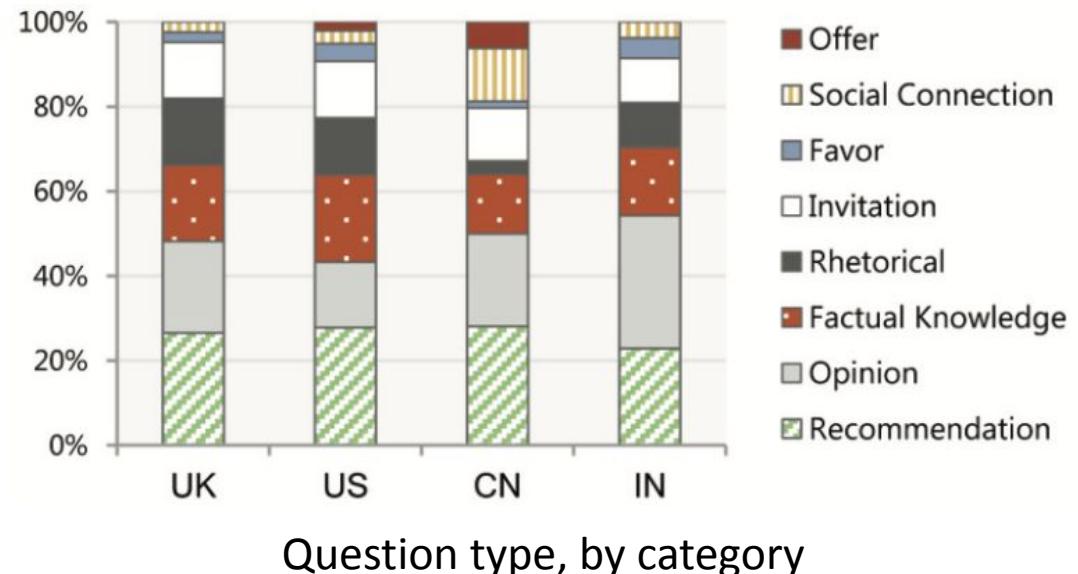


Learn about people from, e.g., their Foursquare check-ins

Cultural Elements are Reflected in Social Datasets

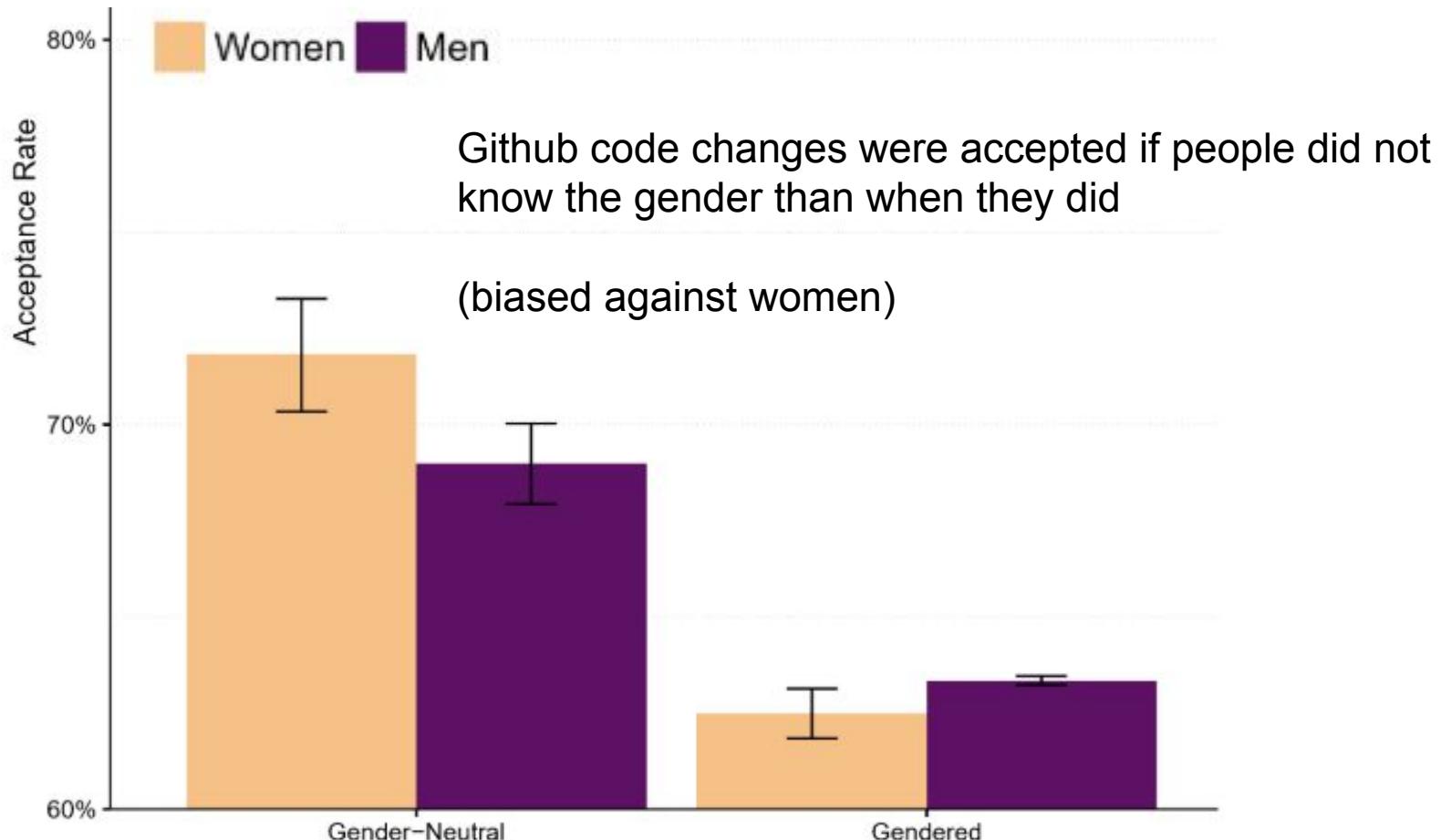


□ Never
■ a few times a year
■ a few times a month
■ a few times a week
■ daily
■ several times a day



Cultural Elements are Reflected in Social Datasets

Code Change Acceptance -- Github



Issues from Data Collection

- **Source Selection Bias**
- Data Acquisition
- Data Querying
- Data Filtering

Data Acquisition Bias

- Some platforms offer no programmatic access (APIs).
- Creates gaps between data and what users see.
- LinkedIn TOS:

LinkedIn sues anonymous data scrapers

Kate Conger @kateconger / 1:09 pm PDT • August 15, 2016

 Comment



Data Acquisition Bias – Rate Limits

GET endpoints

The standard API rate limits described in this table refer to GET (read) endpoints. Note that endpoints not listed in the chart default to 15 requests per user. All request windows are 15 minutes in length. These rate limits apply to the standard API endpoints only, does not apply to premium endpoints.

Endpoint	Resource family	Requests / window (user auth)	Requests / window (app auth)
GET account/verify_credentials	application	75	0
GET application/rate_limit_status	application	180	180
GET favorites/list	favorites	75	75
GET followers/ids	followers	15	15
GET followers/list	followers	15	15
GET friends/ids	friends	15	15
GET friends/list	friends	15	15

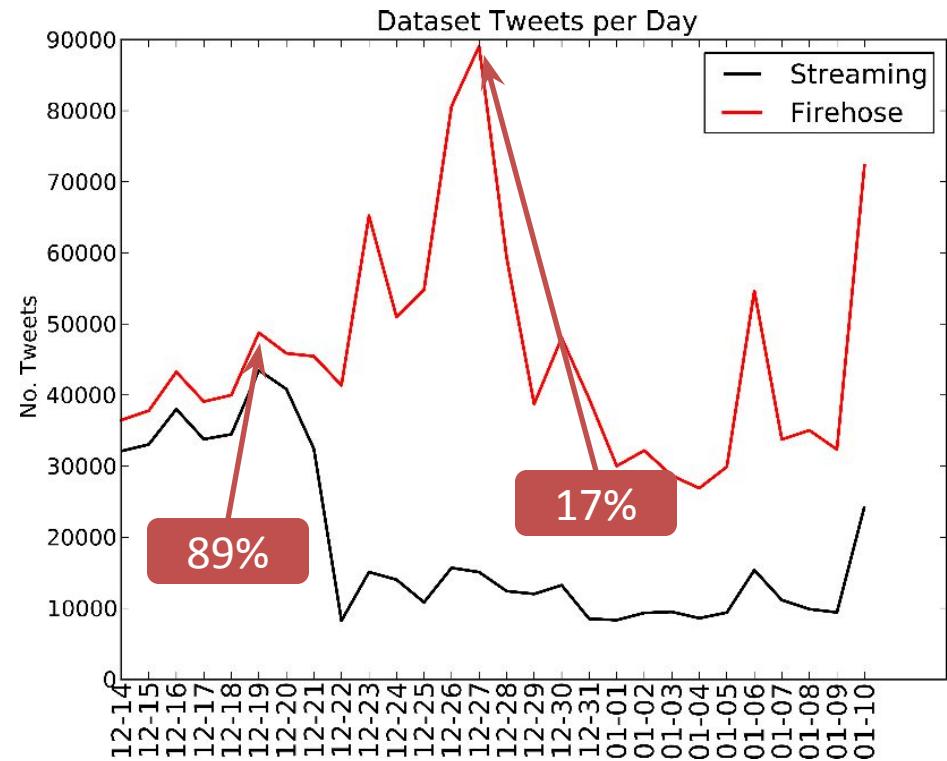
Platforms Do Not Capture All Relevant Data

- We know what people write, but not what they see.
- Having access to this may reveal different insights.
- What is exposure?
- Public / private issue.

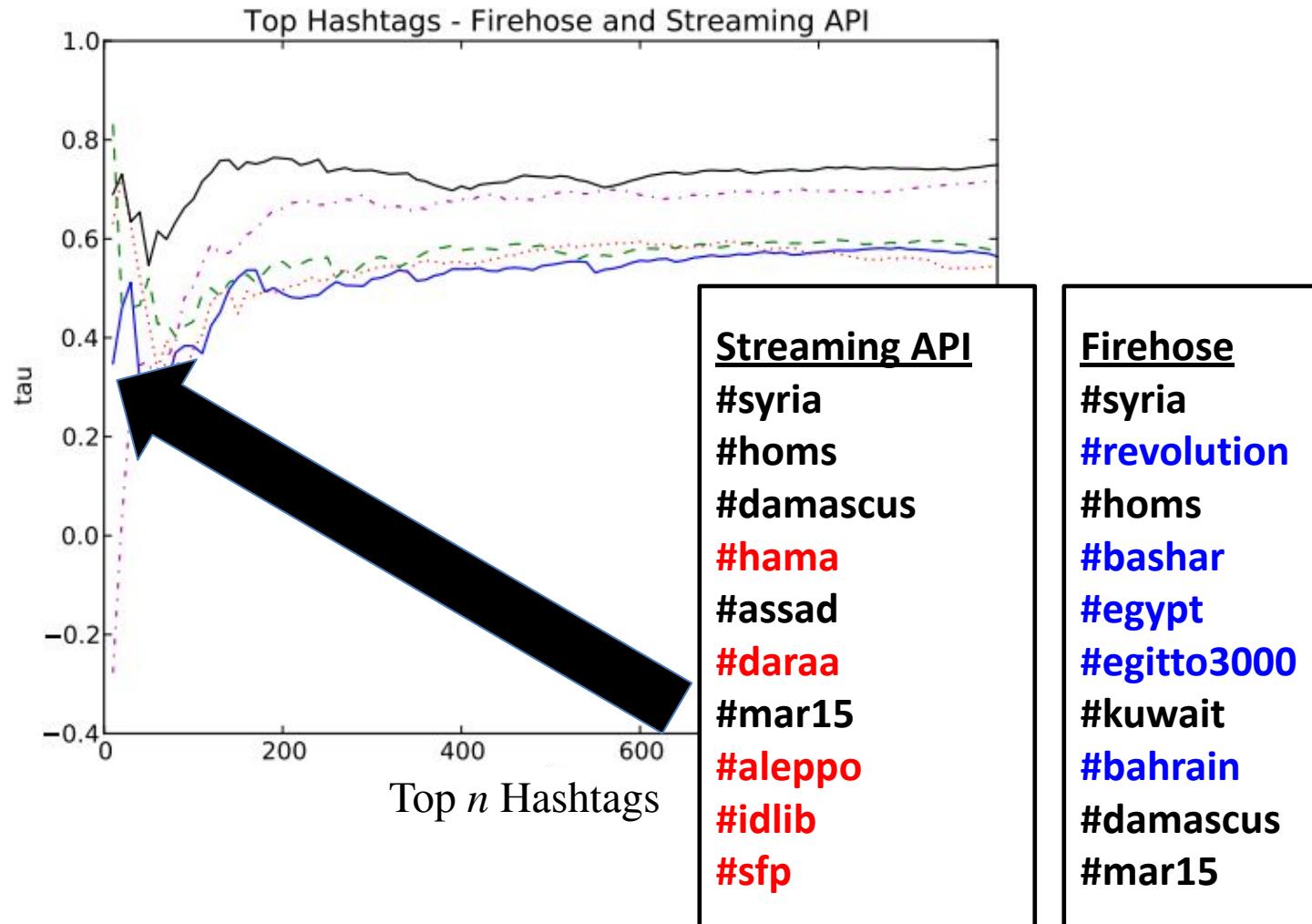
Platforms' Sampling is Often Opaque

- Twitter shares its data.
- “Firehose” - 100% - costly.
- “Sample API” - 1% - free.
 - Takes no parameters from users.
 - Returns a random 1% sample.
- “Streaming API” - 1% - free.
 - Takes query parameters from user.
 - Returns tweets matching query.
 - Samples data when volume reaches 1%.

Dataset



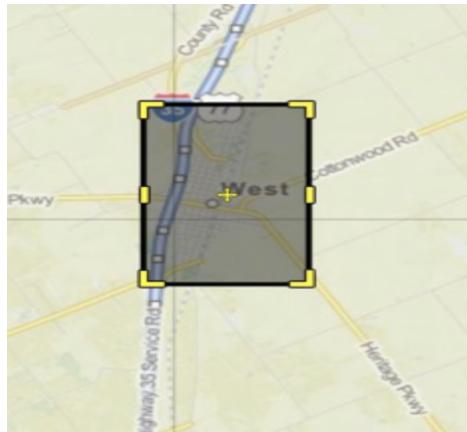
Top Hashtags



APIs Have Limited Expressiveness

- Often take just a few parameter types:
 - Twitter: keywords, geolocation, users.
 - Everything an OR?
- Common Data Querying

This flooding is crazy! Hoping my fellow Albertans and Calgarians stay safe! [#abflood](#) [#yycflood](#)

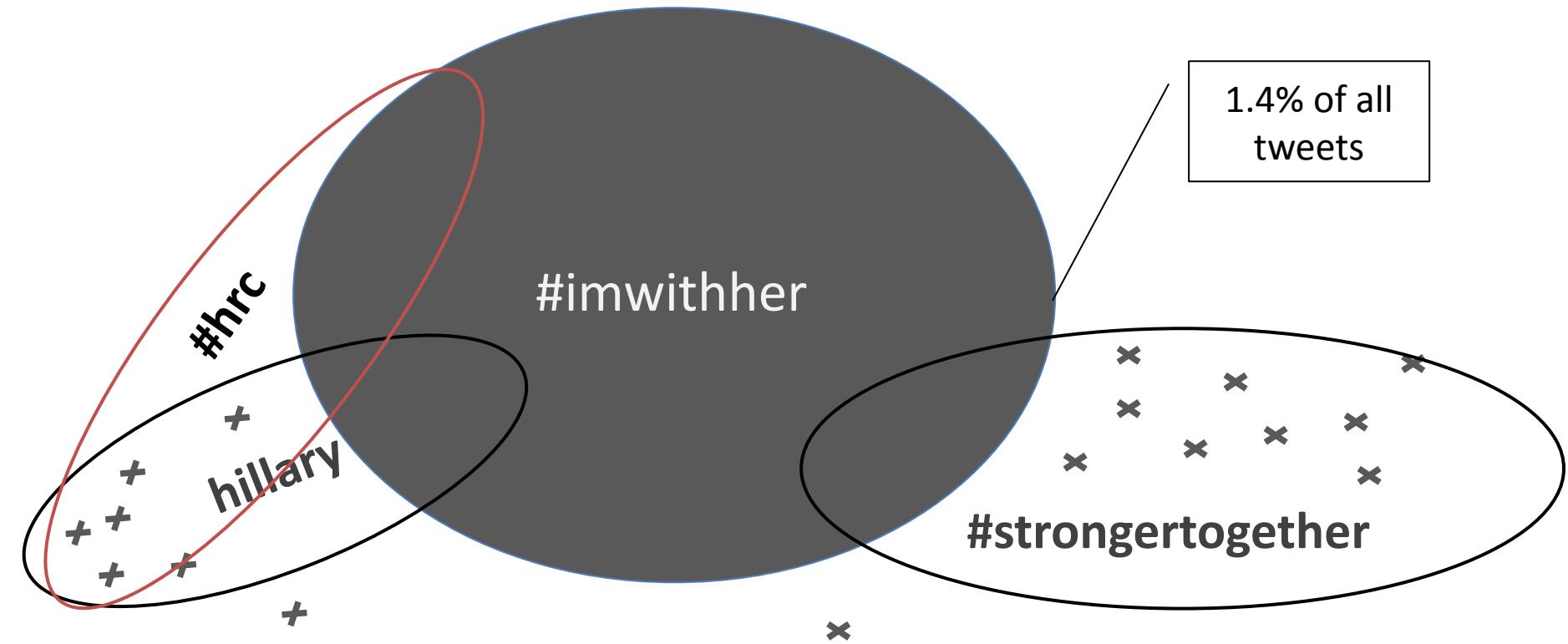


APIs != Populations

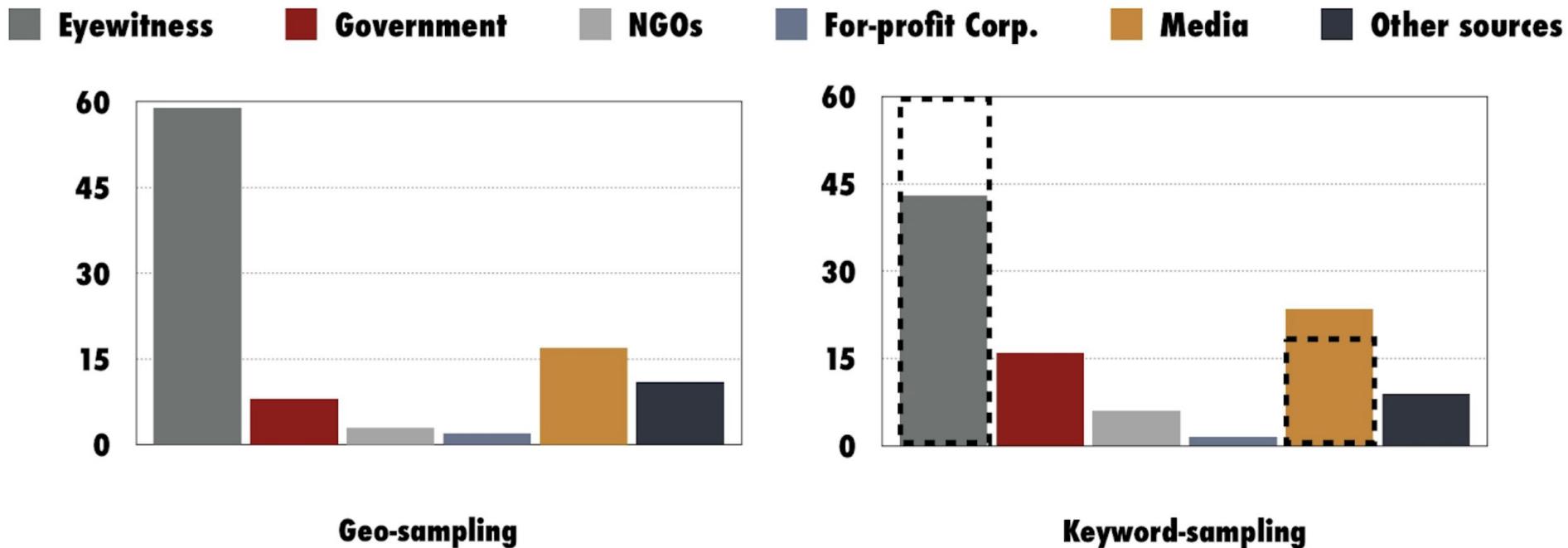
- User-selection criteria may include features held at lower rates by members of certain groups
 - E.g., unpopular keywords
- Use of proxy populations that fail to correctly capture populations under study
- Over or under-emphasis of certain types of users, such as active users or those perceived as relevant.

Choice of keywords is important!

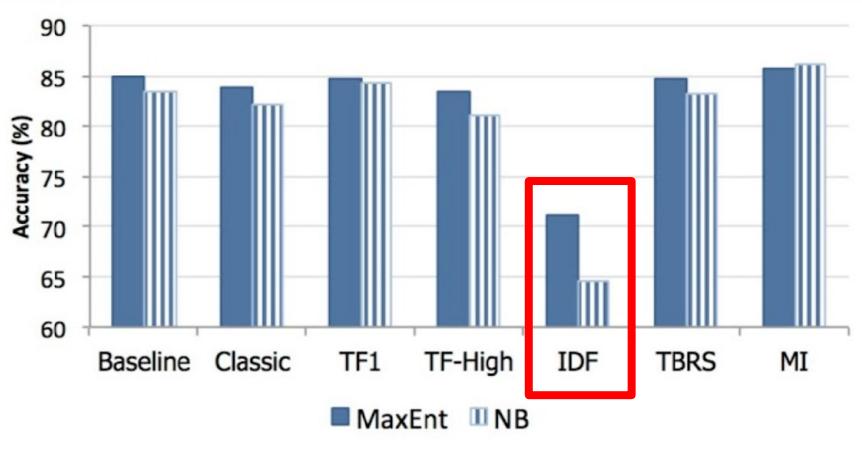
- Still far from perfect (misspellings)
- Hashtags associated with certain political frameworks
- Keyword based collections only capture a fraction of the relevant data.



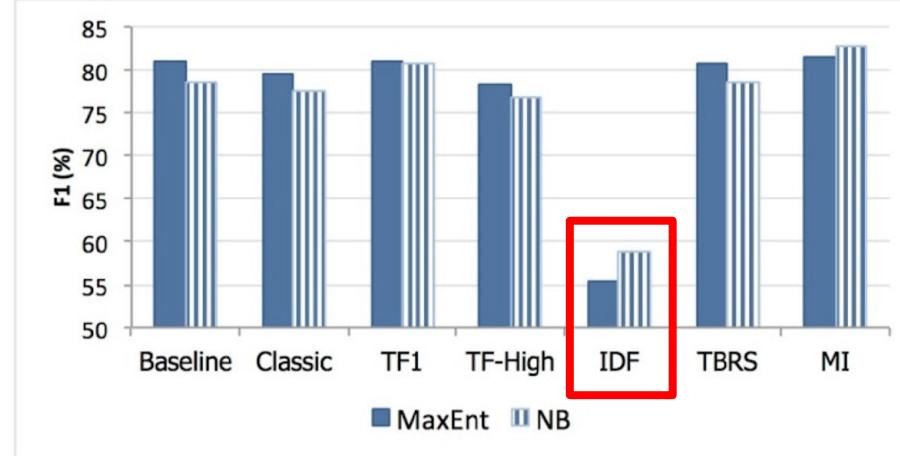
Keywords Overrepresent Media



Data Cleaning Decisions Matter



(a) Average Accuracy



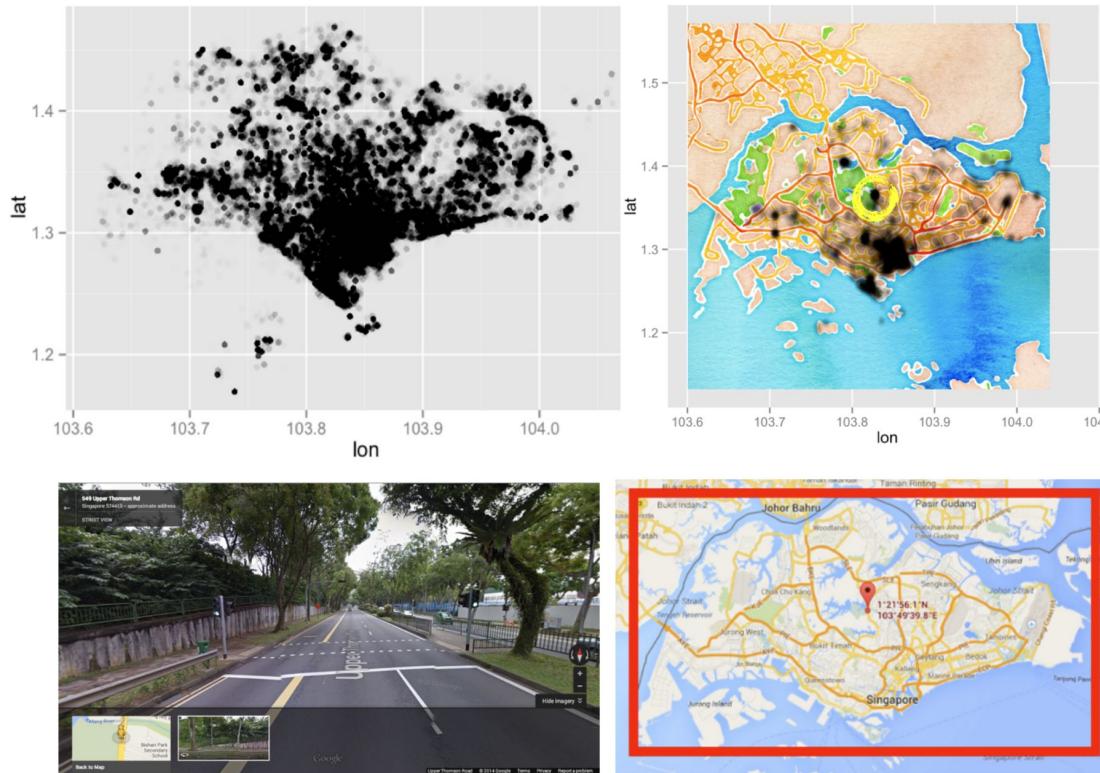
(b) Average F-measure

Two different stoplists – two sets of results!
(also note that metrics change results)

Data Representation Decisions Matter

- How to represent missing values?
- Do you interpret missing values as zeros?

Where are Flickr pictures geo-located in Singapore?



What about Annotation?

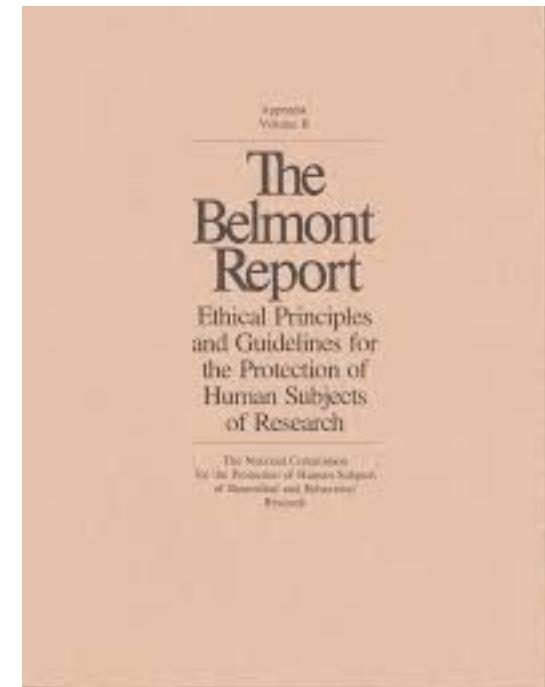
- Factors that affect quality:
 - Unreliable annotations
 - Poor annotation guidelines
 - Category design
 - Insufficient information
 - Ethical concerns – what can you show to the annotators?
 - Skewed difficulty of category assignment (babies vs people in their 40s)

Ethical Boundaries to Research

- Concepts and Principles
- Individual Autonomy
- Beneficence and Nonmaleficence 仁慈与不伤害
- Justice

Concepts and Principles

- Commissioned report in the US
- 3 principles:
 - Autonomy/Respect for persons: Respect for individual autonomy, protect diminished persons.
 - Beneficence and Nonmaleficence: maximize benefits; minimize harms.
 - Justice: benefits and the burdens should be divided justly.

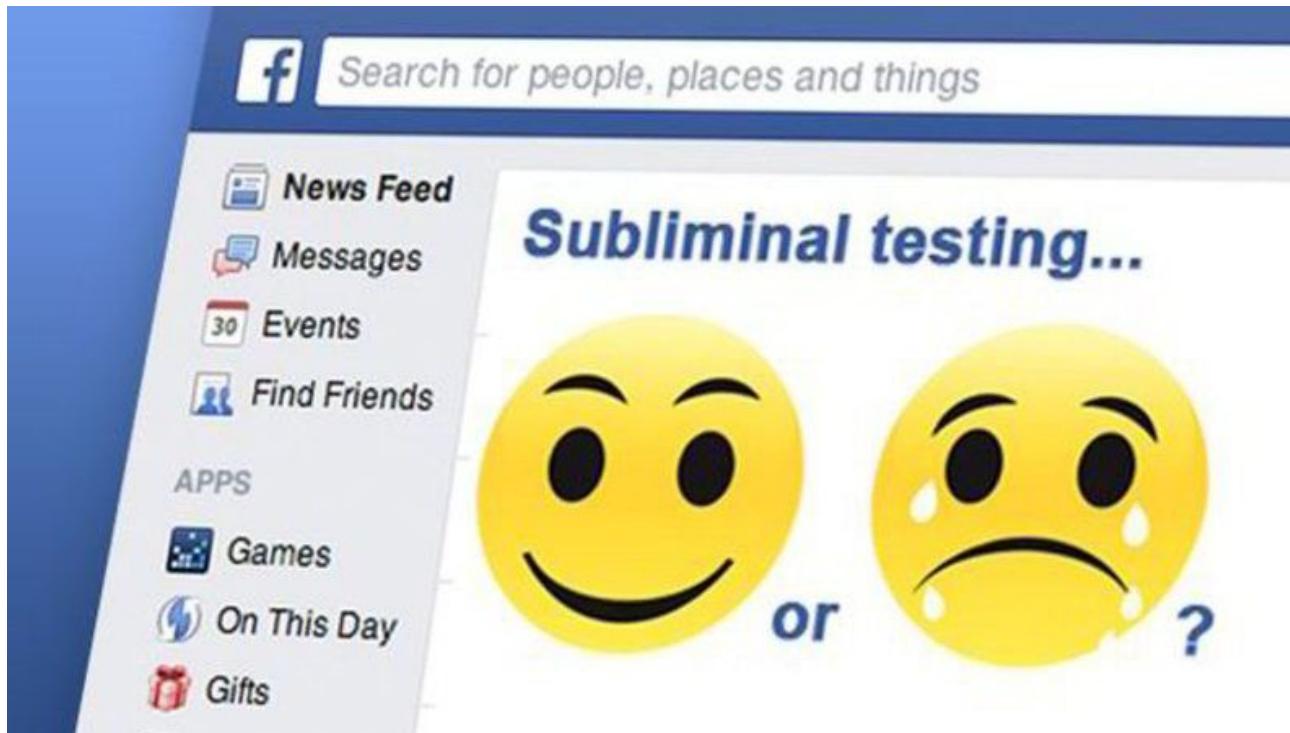


Four Common Types of Research

	Awareness	No Awareness
Manipulation	[Type I] User study, annotation	[Type III] A/B tests
No Manipulation	[Type II] User diaries	[Type IV] Observational studies

Individual Autonomy

- Facebook Contagion Experiment

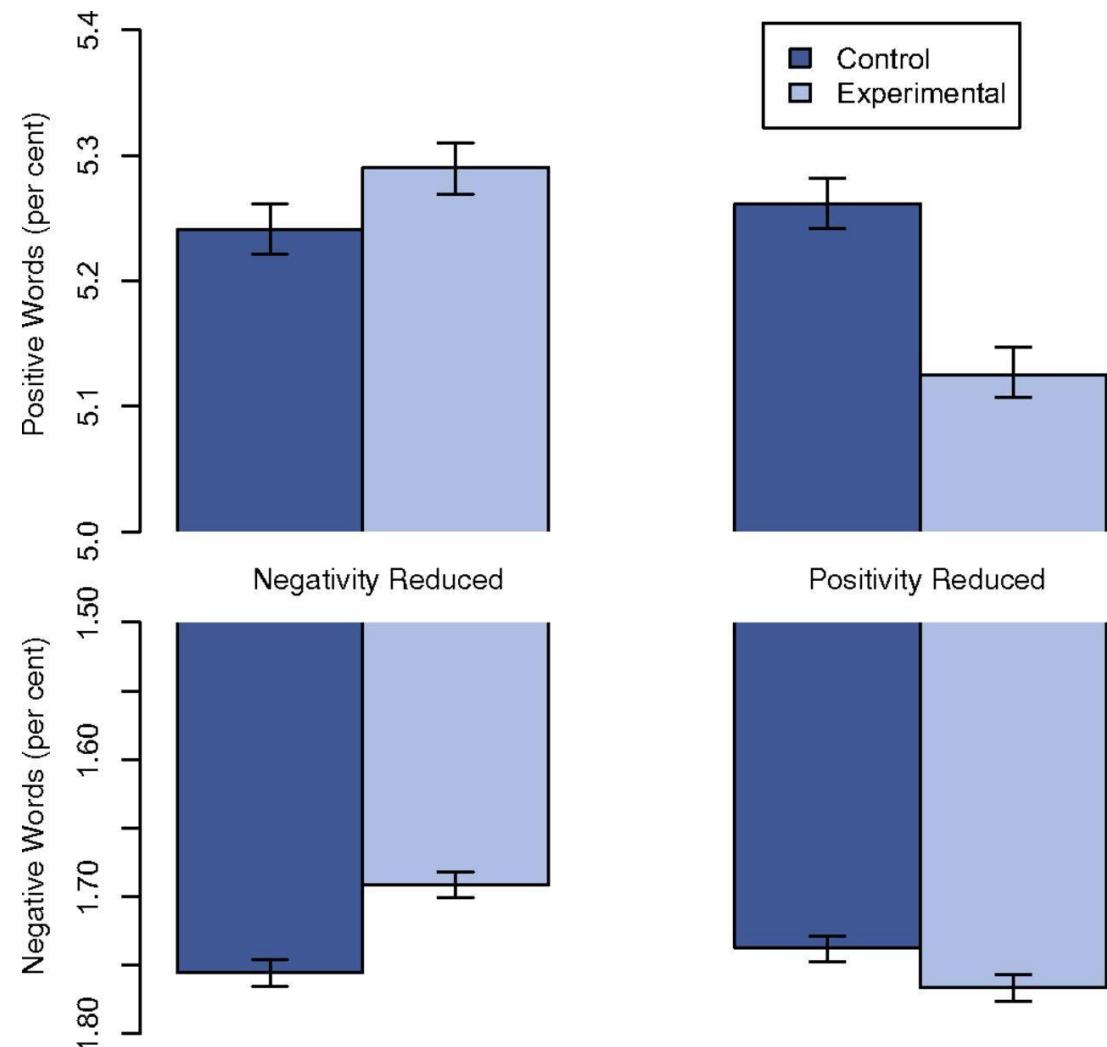


- Paypal's TOS text is 50k words
- Shakespeare's Hamlet is 30k

Facebook Contagion Experiment

Manipulated what people saw and shared without their knowledge

(rights to do experiment within large TOS doc)



Beneficence and Nonmaleficence

The Joy of Tech™



© 2013 Geek Culture

by Nitrozac & Snaggy



joyoftech.com

- Publicly available content wasn't meant to be consumed by just anyone (also issue with Common Crawl and LLMs). [boyd and Crawford 2010]
- “The process of evaluating the research ethics cannot be ignored simply because the data are seemingly public.”

Research outcomes can do harm

- **Aggregation:** A person buys a book about cancer, and a wig.
- **Secondary use:** “Mission/function” creep.
- **Collect embarrassing information:** NSA studying web browsing patterns of target.
- **Distortion:** Writer buys book on manufacturing meth to write a novel.

Conclusion

Issues other people have identified and addressed in research designs

Issues that are yet not identified in research designs

Issues we expect in our research design

Known Knowns

Known Unknowns

Issues we don't expect in our research design

Unknown Knowns

Unknown Unknowns