

Data Collection Methods and Biases in Digital Trace Data

Ancsa Hannák

What is DATA?



Having collected your own data is power

Collecting, cleaning your data teaches you crucial things about the characteristics and limitations of your data set.

Data from Online Platforms

Larger and cheaper than surveys or field experiments

Allows to examine human interactions in their natural environments

We can see immediate feedback after external events

Common criticisms:

Big data doesn't mean more representative or better quality

Data driven analysis, or over simplified representation of a theory

External validity

OUTLINE

1. Ethical and Legal issues around data collection
2. Overview of tools available to collect data from online platforms

Make crawling less painful

3. Representation and bias in online data

Be more conscious about the limitations of your sampling method, population, and characteristics of data

Ethical and Legal Considerations

1. Am I harming the users?

Interference through experiments, collection of personal or sensitive information

2. Am I harming the site?

Click fraud, interference with algorithms on the site

3. Overcoming limitations of the Platform

Rate limits, robots.txt, terms of service

Am I Harming the Users?

IRB/ERB (Ethical Review Board)

Has to approve research to protect the rights and welfare of human research subjects

- Interference through experiments

Try to minimize the effect on users

- Personal vs Sensitive Information

Anonymize data and make sure to not share it publicly, especially if there is a danger of fingerprinting

Facebook reveals news feed experiment to control emotions

Protests over secret study involving 689,000 users in which they were moved to influence moods

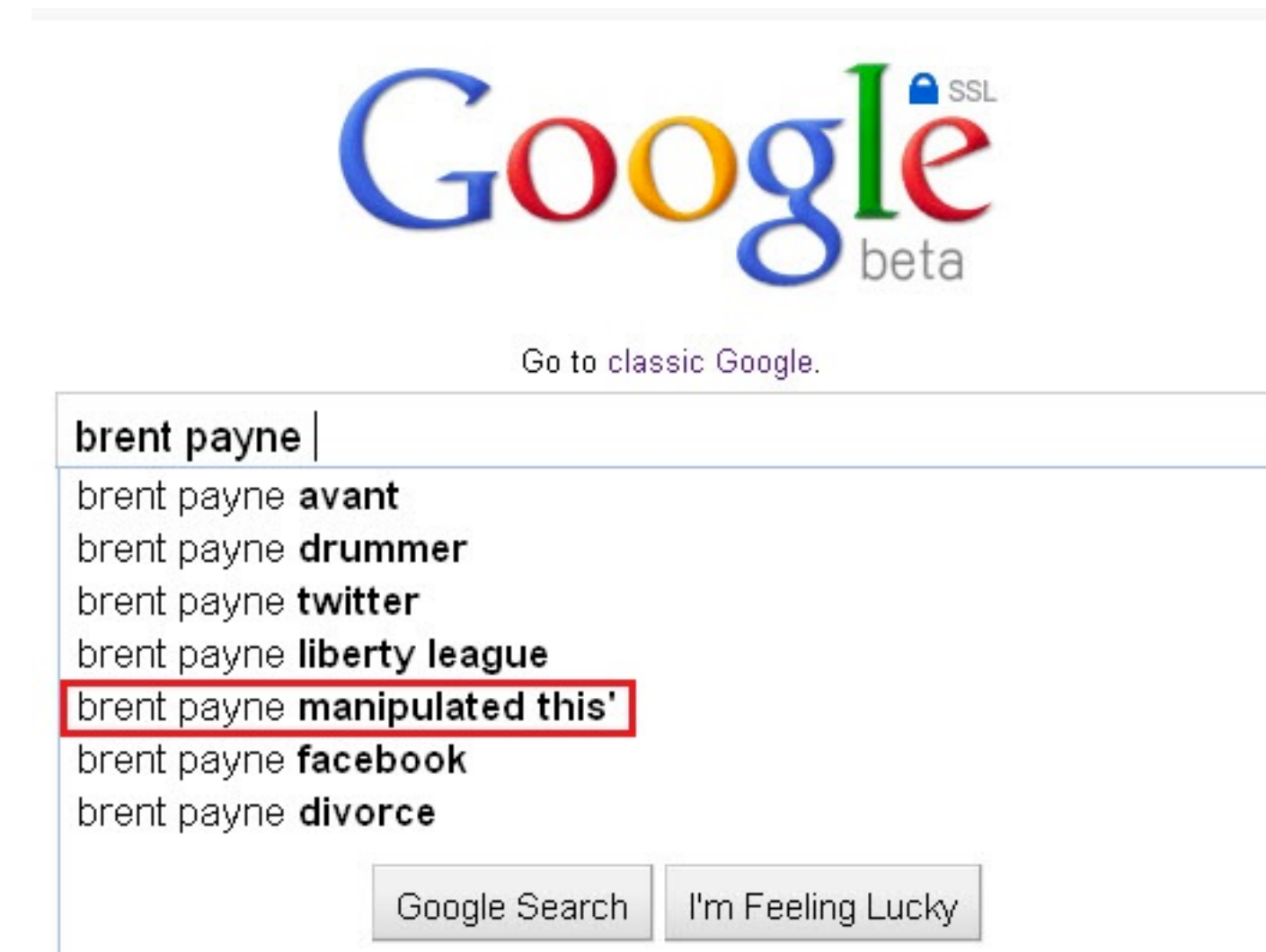


Harming Platforms

Click Fraud: keep in mind that advertisers pay for every click or impression

Interference with algorithms on the site by clicking or searching

AdFisher may have cost advertisers a small sum of money. AdFisher never clicked on any ads to avoid per click fees, which can run over \$4 [34]. Its experiments may have caused per-impression fees, which run about \$0.00069 [35]. In the billion dollar ad industry, its total effect was about \$400.



Am I going to jail?

Terms of Service



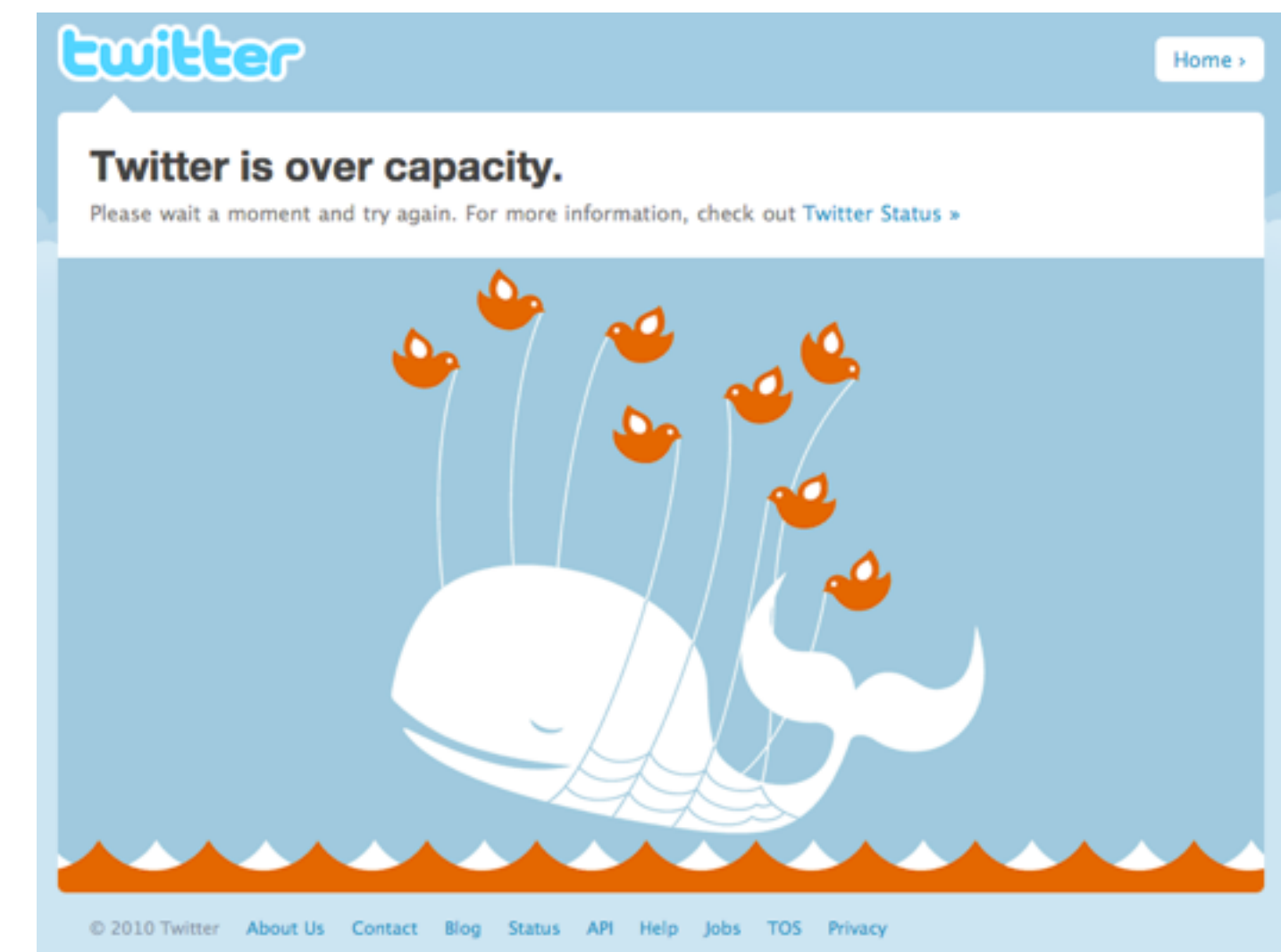
1. You will not provide any false personal information on Facebook, or create an account for anyone other than yourself without permission.
2. You will not create more than one personal account.
7. If you collect information from users, you will: obtain their consent, make it clear you (and not Facebook) are the one collecting their information, and post a privacy policy explaining what information you collect and how you will use it.

Robots.txt

```
https://varvy.com/robots.txt  
User-agent: *  
Disallow: /folder/  
Disallow: /file.html  
Disallow: /image.png
```

If you really care, time to switch fields

Rate limits



CFAA Lawsuit

External audits are crucial in identifying pernicious business practices such as discrimination or redlining under the Computer Fraud and Abuse Act

We're suing the federal government to be free to do our research

March 28, 2017 3.40am BST

The screenshot shows the LinkedIn Terms of Service page. At the top, the LinkedIn logo is on the left, and 'Sign in' and 'Join now' buttons are on the right. The main heading is '8. LinkedIn "DOs" and "DON'Ts."'.

8.1. Dos. You agree that you will:

- Comply with all applicable laws, including, without limitation, privacy laws, intellectual property laws, anti-spam laws, export control laws, tax laws, and regulatory requirements;
- Provide accurate information to us and keep it updated;
- Use your real name on your profile;
- Use the Services in a professional manner.

8.2. Don'ts. You agree that you will not:

- Send spam or other unwelcomed communications to others;
- Scrape or copy profiles and information of others through any means (including crawlers, browser plugins and add-ons, and any other technology or manual work);

A red arrow points from the text 'Prevents research' to the 'Don'ts' section.

Authors



Christo Wilson
Assistant Professor of Computer and Information Science, Northeastern University



Alan Mislove
Associate Professor of Computer Science, Northeastern University

Sharing your data publicly

First step, make sure to anonymize users

Even then people can be fingerprinted if:

sample size is small, there are outliers or minorities among the population, it can be merged with other available data sets, etc

K-anonymization: Given person-specific field-structured data, produce a release of the data with scientific guarantees that the individuals who are the subjects of the data cannot be re-identified while the data remain practically useful.

Do not share copyright content

Just because you can download it, it is still someone's intellectual property

OUTLINE

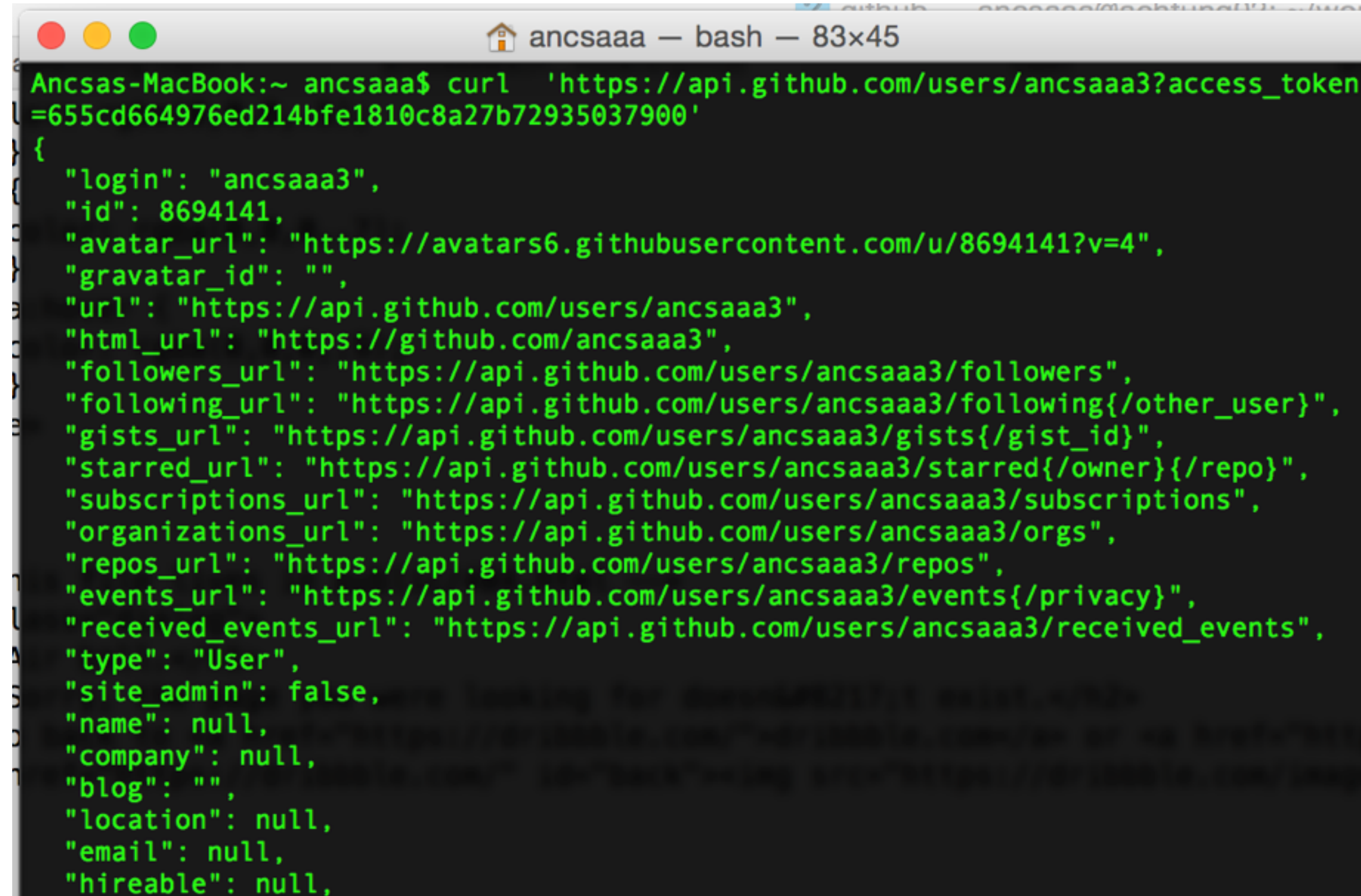
1. Ethical and Legal issues around data collection
2. Overview of tools available to collect data from online platforms
3. Representation and bias in online data

Data Collection

	Tools (examples)	Pros	Cons
API		ToS compliant, easy to use	possible bias, incompleteness Auth and rate limits
scraping static pages	Curl, python requests	easy to use, parallelizable	no ajax, no images, no javascript you have to parse data
Automated Browser	Selenium	mimics real humans, possible to log-in, design flow of events	not possible to parallelize, unpredictable bugs (pop-ups, ads) you have to parse data
Headless Browser Implementation	phantomJS, selenium	fast, parallelizable	hard to debug since there is no physical browser window you have to parse data

APIs

```
curl 'https://api.github.com/users/ancsaaa3?access_token=655cd664976ed214bfe1810c8a27b72935037900'
```



```
ancsaaa — bash — 83x45
Ancsas-MacBook:~ ancsaaa$ curl 'https://api.github.com/users/ancsaaa3?access_token=655cd664976ed214bfe1810c8a27b72935037900'
{
  "login": "ancsaaa3",
  "id": 8694141,
  "avatar_url": "https://avatars6.githubusercontent.com/u/8694141?v=4",
  "gravatar_id": "",
  "url": "https://api.github.com/users/ancsaaa3",
  "html_url": "https://github.com/ancsaaa3",
  "followers_url": "https://api.github.com/users/ancsaaa3/followers",
  "following_url": "https://api.github.com/users/ancsaaa3/following{/other_user}",
  "gists_url": "https://api.github.com/users/ancsaaa3/gists{/gist_id}",
  "starred_url": "https://api.github.com/users/ancsaaa3/starred{/owner}/{/repo}",
  "subscriptions_url": "https://api.github.com/users/ancsaaa3/subscriptions",
  "organizations_url": "https://api.github.com/users/ancsaaa3/orgs",
  "repos_url": "https://api.github.com/users/ancsaaa3/repos",
  "events_url": "https://api.github.com/users/ancsaaa3/events{/privacy}",
  "received_events_url": "https://api.github.com/users/ancsaaa3/received_events",
  "type": "User",
  "site_admin": false,
  "name": null,
  "company": null,
  "blog": "",
  "location": null,
  "email": null,
  "hireable": null,
```

Terms of Service compliant
Easy to use
Data is in nice format

Auth and rate limits
Possible incompleteness of data
Biases are unknown

APIs - rate limits



Auth and rate limits

E.g.: X request/day/auth
or X% of data

APIs - completeness of data

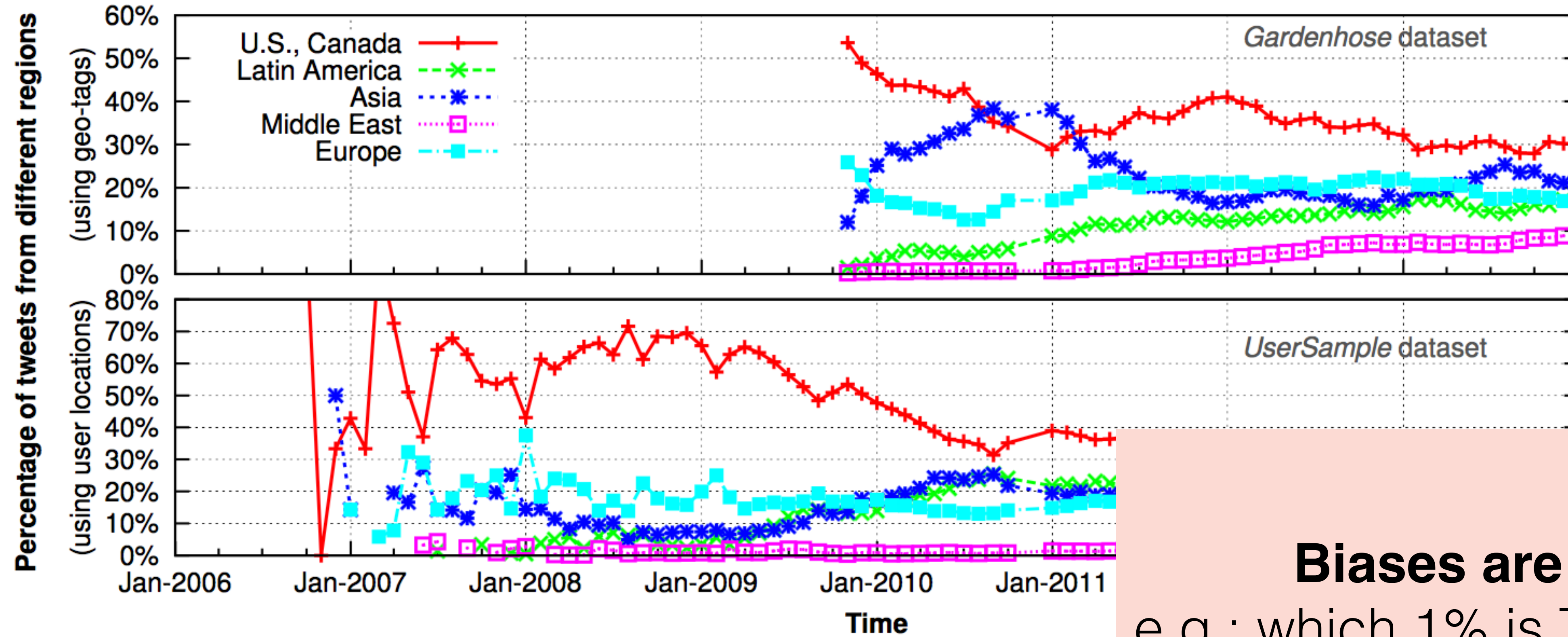
```
curl 'https://api.dribbble.com/v1/users/justintran/shots?access_token=f3f9df1f093c81071cf59df03428870d46a7c9f8460276600778872af294be09'
```

```
ancsaaa — bash — 179x48
Ancsas-MacBook:~ ancsaaa$ curl 'https://api.dribbble.com/v1/users/justintran/shots?access_token=f3f9df1f093c81071cf59df03428870d46a7c9f8460276600778872af294be09'
{
  "id": 3631505,
  "title": "Prepare for work like an athlete",
  "description": "<p>for the <a href=\"http://blogs.dropbox.com/dropbox/2017/07/work-tips-from-pro-athletes/\" rel=\"nofollow noreferrer\">Dropbox Blog</a></p>\n\n<p><a href=\"https://dribbble.com/shots/3631505-Untitled-1/attachments/810946\" rel=\"nofollow noreferrer\">larger</a></p>",
  "width": 400,
  "height": 300,
  "images": {
    "hidpi": "https://cdn.dribbble.com/users/237814/screenshots/3631505/untitled-1.jpg",
    "normal": "https://cdn.dribbble.com/users/237814/screenshots/3631505/untitled-1_1x.jpg",
    "teaser": "https://cdn.dribbble.com/users/237814/screenshots/3631505/untitled-1_teaser.jpg"
  },
  "views_count": 5141,
  "likes_count": 310,
  "comments_count": 9,
  "attachments_count": 1,
  "rebounds_count": 0,
  "buckets_count": 15,
  "created_at": "2017-07-05T16:46:18Z",
  "updated_at": "2017-07-05T16:47:32Z",
  "html_url": "https://dribbble.com/shots/3631505-Prepare-for-work-like-an-athlete",
  "attachments_url": "https://api.dribbble.com/v1/shots/3631505/attachments",
  "buckets_url": "https://api.dribbble.com/v1/shots/3631505/buckets",
  "comments_url": "https://api.dribbble.com/v1/shots/3631505/comments",
  "likes_url": "https://api.dribbble.com/v1/shots/3631505/likes",
  "projects_url": "https://api.dribbble.com/v1/shots/3631505/projects",
  "rebounds_url": "https://api.dribbble.com/v1/shots/3631505/rebounds",
  "animated": false,
  "tags": [
    "dropbox",
    "illustration",
    "sports",
    "work"
  ],
  "team": {
    "id": 148611,
    "name": "Dropbox",
    "username": "dropbox",
    "html_url": "https://dribbble.com/dropbox",
    "avatar_url": "https://cdn.dribbble.com/users/148611/avatars/normal/1f8c32a9c8864e161ceca7d41e6b16be.jpg?1480441349",
    "bio": "Simplifying the way people work together · Join us! <a href=\"http://dropbox.com/jobs/design\" rel=\"nofollow noreferrer\">dropbox.com/jobs/design</a>",
    "location": "SF, NY",
    "links": {
      "web": "http://dropbox.com",
      "twitter": "https://twitter.com/dropboxdesign"
    }
  }
}
```

Incompleteness of data

- 1) Missing info such as images
- 2) Can not measure exact user experience

APIs - unknown biases

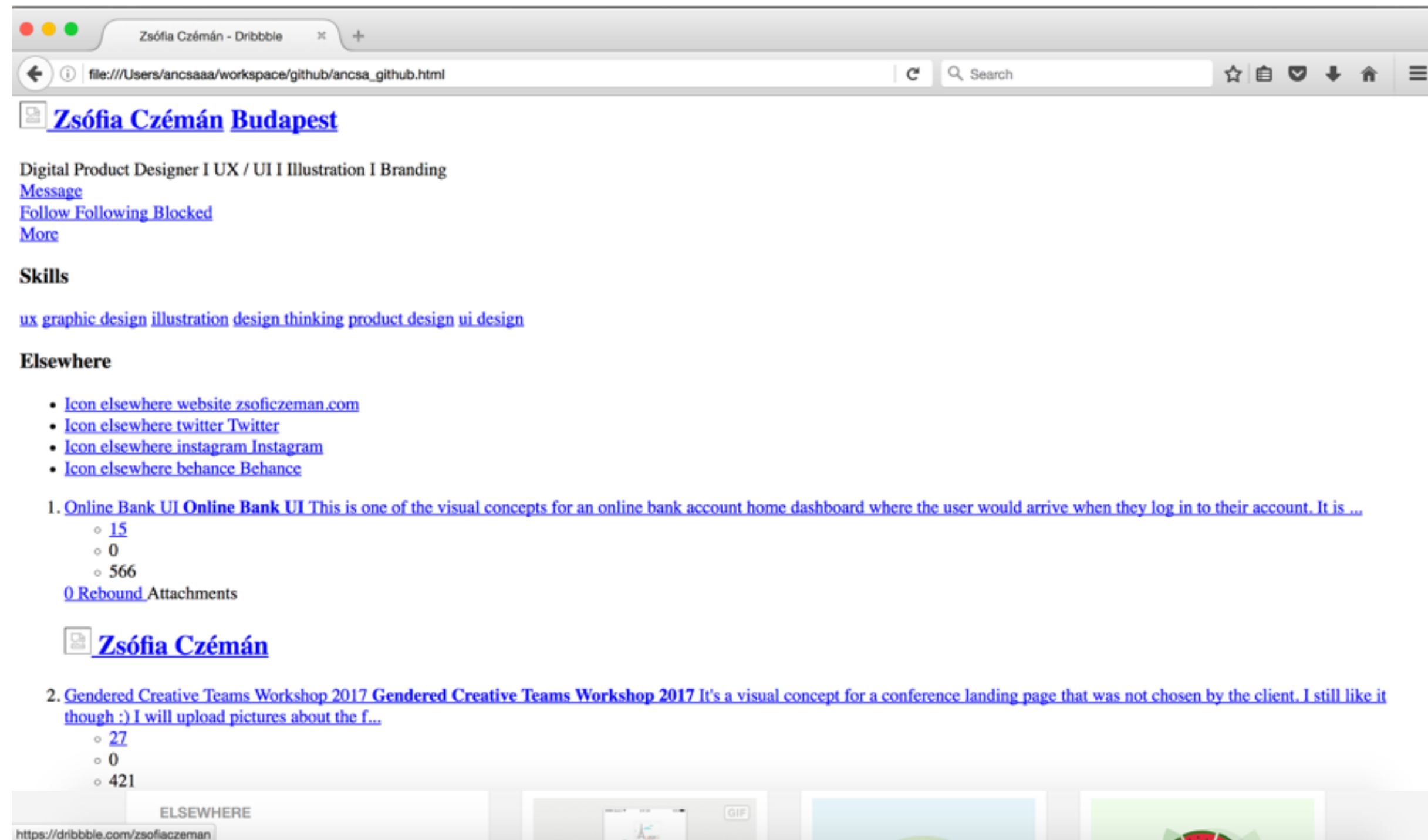


Biases are unknown
e.g.: which 1% is Twitter giving us?

Scraping via requests or curl

BASH: `curl https://dribbble.com/zsofiaczeman > user.html`

Python Requests: `requests.get("https://dribbble.com/zsofiaczeman")`



Easy to use
Easy to parallelize

Not ToS compliant
No ajax, no images, no javascript
You have to parse content

Parsing the source code

Tools: BeautifulSoup, LXML, etc

```
63 <div id="logo">
64   <a href="/">Toggle navigation</a>
68
69 <ul id="nav" role="navigation">
70   <li id="t-search" role="search">
71     <form id="search" action="https://dribbble.com/search">
72       <input class="search-text" type="text" name="q" placeholder="Search " value="" />
73     </form>
74   </li>
75   <li id="t-signin">
76     <a href="https://dribbble.com/session/new?return_to=%2Fzsofiaczeman">
77       <span>Sign in</span>
78     </a>
79   </li>
80   <li id="t-signup">
81     <a href="https://dribbble.com/signup?return_to=%2Fzsofiaczeman">Sign up</a>
82   </li>
83   <li id="t-shots">
84     <a class="has-sub" href="/shots">Shots</a>
85     <ul class="tabs">
86       <li><a href="/shots">Popular</a></li>
87       <li><a href="/shots?sort=recent">Recent</a></li>
88       <li><a href="/shots?list=debuts">Debuts</a></li>
89       <li><a href="/shots?list=teams">Teams</a></li>
90       <li><a href="/shots?list=playoffs">Playoffs</a></li>
91     </ul>
92   </li>
93   <li class="separate"><a href="/highlights">Highlights</a></li>
94   <li><a href="/projects">Projects</a></li>
95   <li><a href="/goods">Goods by Designers</a></li>
96 </ul>
97
98 </li>
99 <li id="t-players">
```

Easy to use
Easy to parallelize

Not ToS compliant
No ajax, no images, no javascript
You have to parse content

Automated Browsing



SeleniumHQ
Browser Automation

[edit this page](#) search selenium: [Go](#)

[Projects](#) [Download](#) [Documentation](#) [Support](#) [About](#)

What is Selenium?

Selenium automates browsers. That's it! What you do with that power is entirely up to you. Primarily, it is for automating web applications for testing purposes, but is certainly not limited to just that. Boring web-based administration tasks can (and should!) also be automated as well.

Selenium has the support of some of the largest browser vendors who have taken (or are taking) steps to make Selenium a native part of their browser. It is also the core technology in countless other browser automation tools, APIs and frameworks.

Which part of Selenium is appropriate for me?

Selenium WebDriver



If you want to

- create robust, browser-based regression automation suites and tests
- scale and distribute scripts across many environments

Then you want to use [Selenium WebDriver](#); a collection of language specific bindings to drive a browser -- the way it is meant to be driven.

Selenium IDE



If you want to

- create quick bug reproduction scripts
- create scripts to aid in automation-aided exploratory testing

Then you want to use [Selenium IDE](#); a Firefox add-



Selenium is a suite of tools to automate web browsers across many platforms.

Selenium...

- runs in [many browsers](#) and [operating systems](#)
- can be controlled by many [programming languages](#) and [testing frameworks](#).

[Download Selenium](#)

Donate to Selenium with PayPal [Donate](#)

Mimics real human browsing
Loads ajax, images, etc
Design flow of events, e.g. log-in, search

Not ToS compliant
You have to parse content
Difficult to scale
Unpredictable bugs (e.g. pop-ups)

Headless browser

E.g.: PhantomJS



PhantomJS

**Headless webkit with
javascript API**

Fast

Easy to parallelize

You can design a flow of events

Hard to debug since there
is no physical browser window

You have to parse data

Ugly code

Not ToS compliant

Data Collection

	Tools (examples)	Pros	Cons
API		ToS compliant, easy to use	possible bias, incompleteness Auth and rate limits
scraping static pages	Curl, python requests	easy to use, parallelizable	no ajax, no images, no javascript you have to parse data
Automated Browser	Selenium	mimics real humans, possible to log-in, design flow of events	not possible to parallelize, unpredictable bugs (pop-ups, ads) you have to parse data
Headless Browser Implementation	phantomJS, selenium	fast, parallelizable	hard to debug since there is no physical browser window you have to parse data

What to crawl

Obtaining the list of URLs:

List of keywords (e.g.: Twitter, Google Search, Wikipedia)

All users, pictures, items of a site:

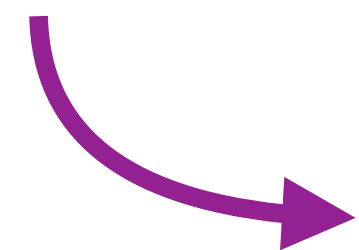
Sequential IDs

First search for all possible users, images, etc

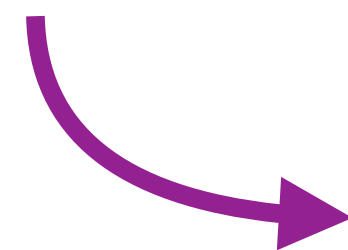
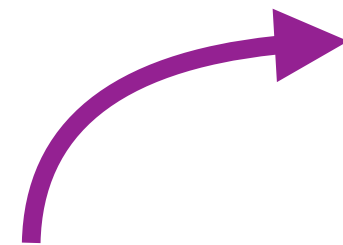
Search for all teams

Extract all userids

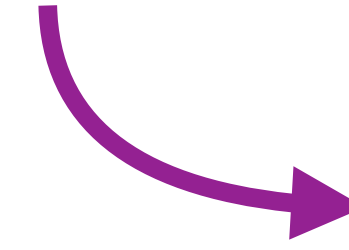
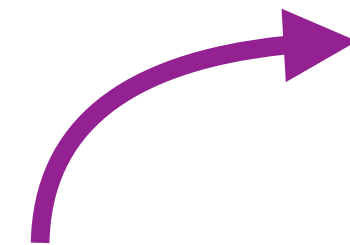
Extract all of their images



Crawl for all teams



Crawl all users



Crawl all images

Hacks and Tricks (I)

Overcoming Rate Limits:

Parallelization through multiple IPs

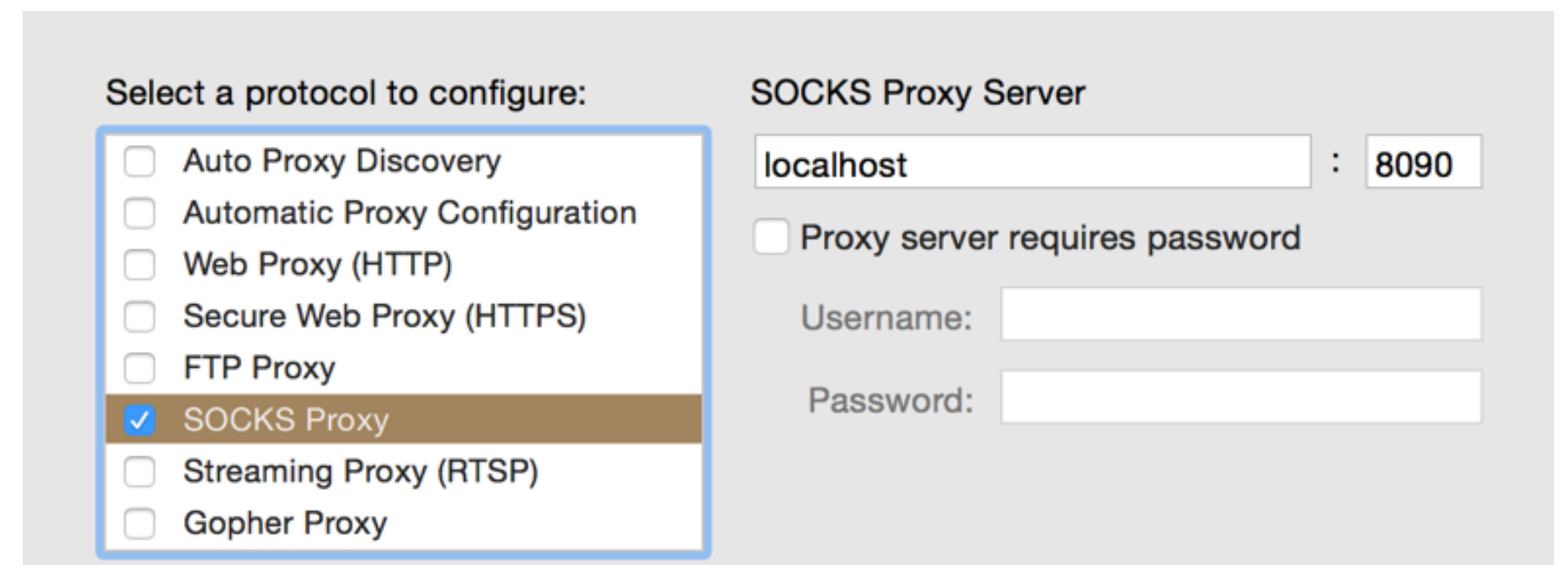
Changing IPs once limits exceeded

Breaking captchas



SSH tunnels:

```
Ancsas-Mac:~ancsaaa$ ssh -D 8090 ccs.neu.edu
```



Hacks and Tricks (II)

Overcoming personalization, localization effects

PlanetLab Machines, Amazon/Azure

HideMyAss

Hitting the same data centers

##

Host Database

##

127.0.0.1

localhost

255.255.255.255

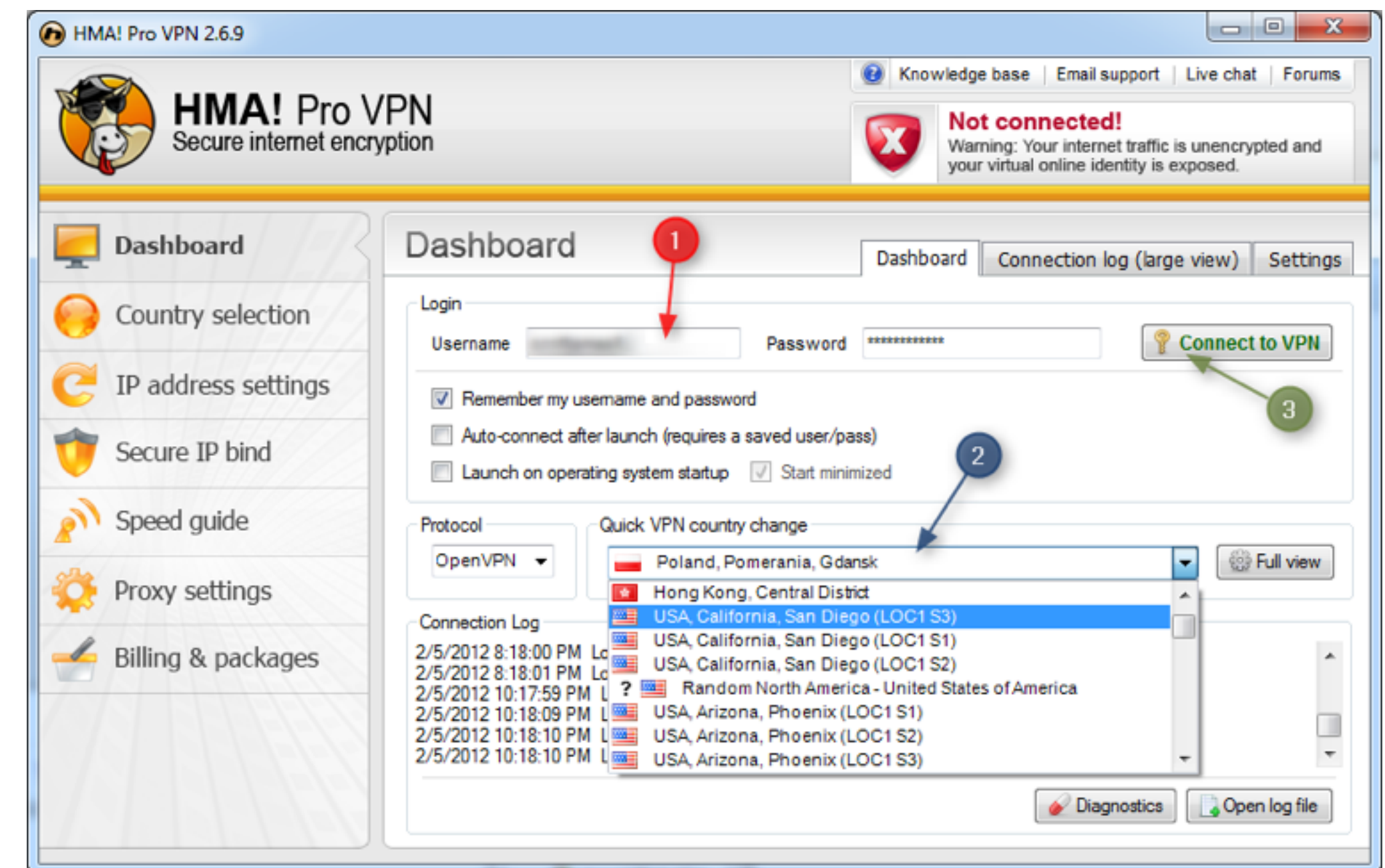
broadcasthost

:::1

localhost

172.217.18.78

www.google.com



Part II

Bias and representativeness
of digital trace data

Representativeness of your sample

Internal validity: “Internal Validity is the approximate truth about inferences regarding causal relationships”

Cross-platform validity

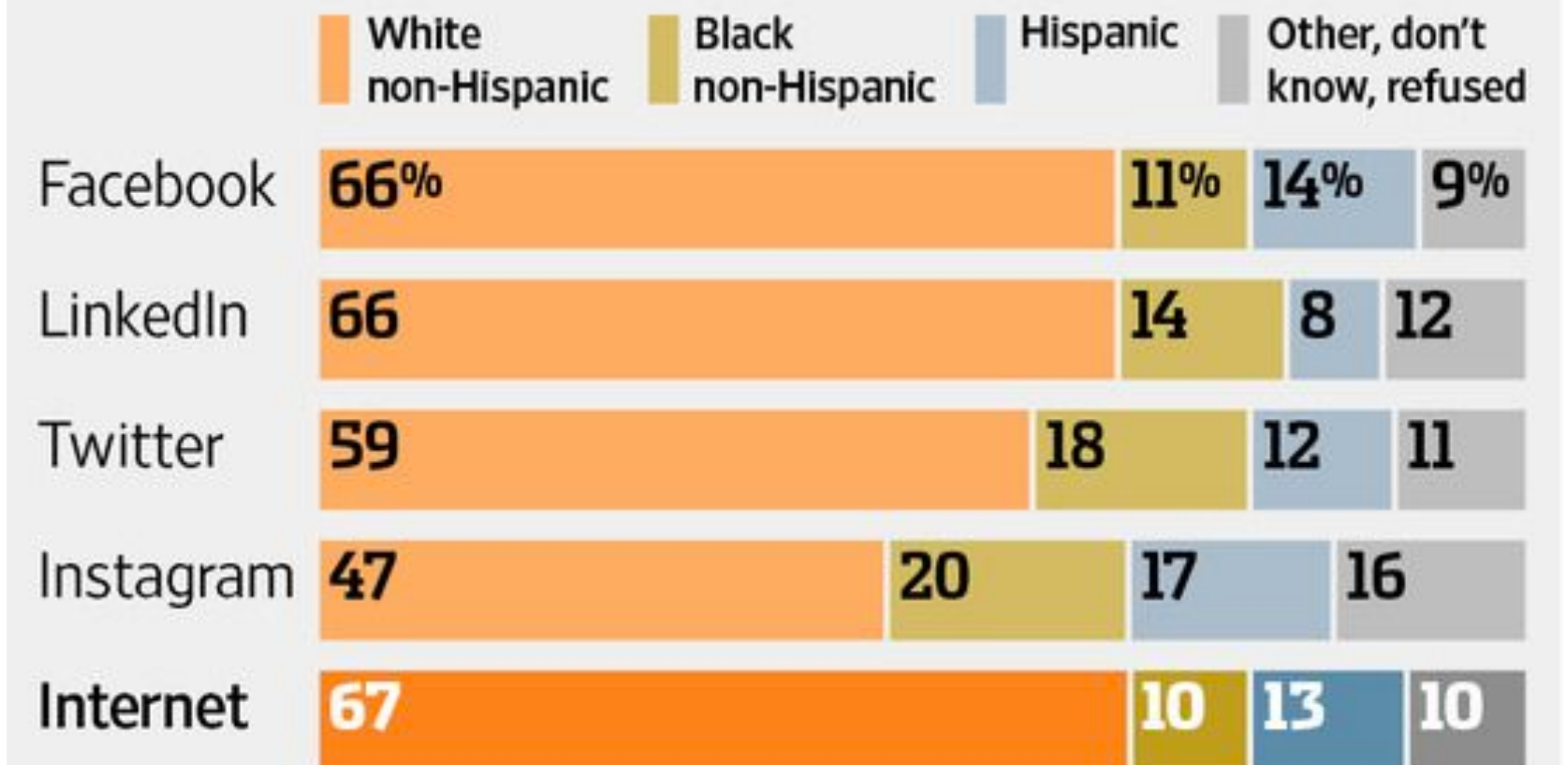
External validity

Who uses internet? Who uses social media? Who are the users of the platform you are looking at?

External Validity Problems



User demographics for social-media services compared with the overall U.S. Internet population



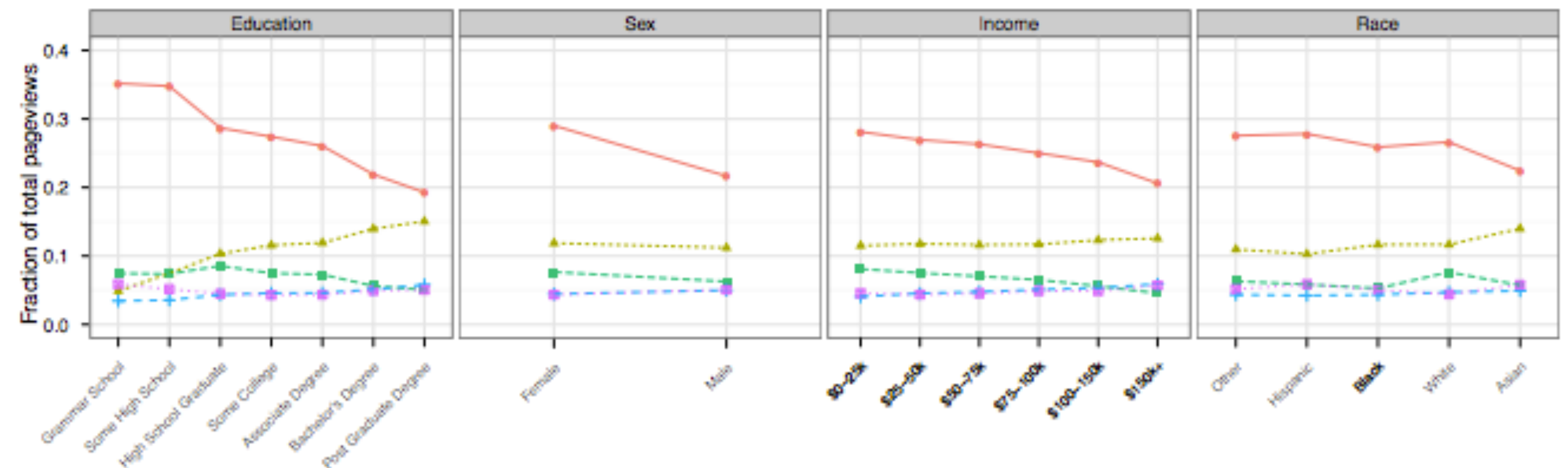
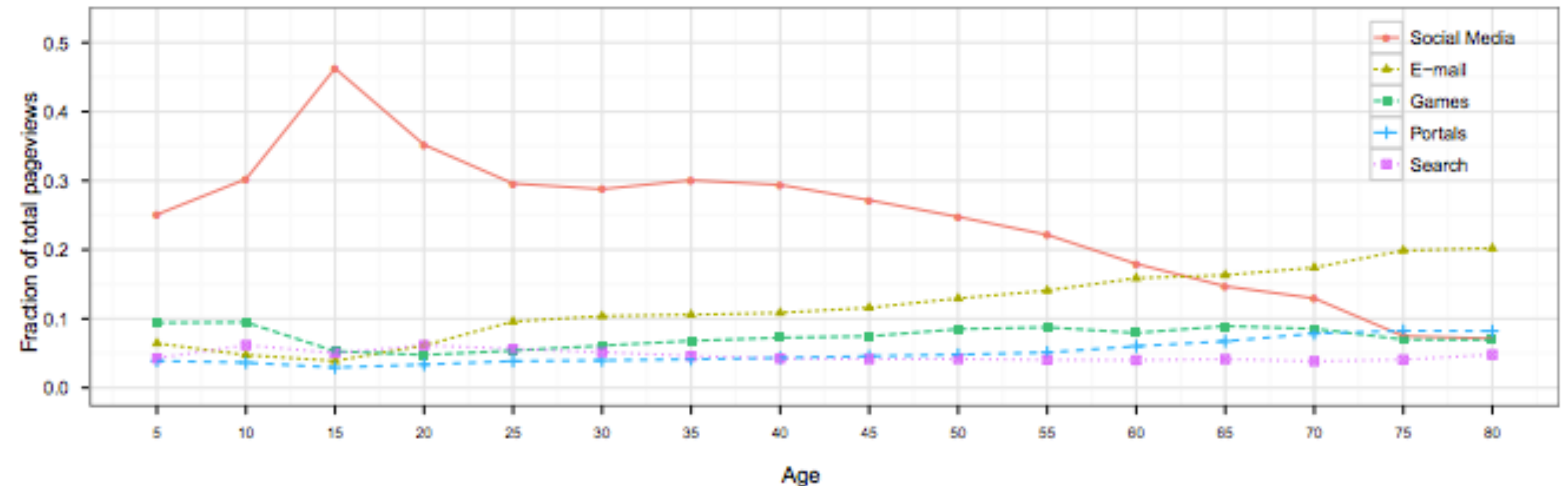
Source: Pew Research Center

The Wall Street Journal

Who creates what type of content?

[Goel et al. ICWSM'12]

Demographic features correlate highly with the amount of time spent in various type of online activities.



Predicting the German elections

Twitter is commonly used to predict “things”, especially elections
Multiple papers analyzing German politics and making predictions

Party	All mentions		Election	
	Number of tweets	Share of Twitter traffic	Election result*	Prediction error
CDU	30,886	30.1%	29.0%	1.0%
CSU	5,748	5.6%	6.9%	1.3%
SPD	27,356	26.6%	24.5%	2.2%
FDP	17,737	17.3%	15.5%	1.7%
LINKE	12,689	12.4%	12.7%	0.3%
Grüne	8,250	8.0%	11.4%	3.3%
			MAE:	1.65%

* Adjusted to reflect only the 6 main parties in our sample

[Tumasjan et al. ICWSM'10]

“the mere number of messages mentioning a party reflects the election result”

The

arty

*“I Wanted to Predict Elections with Twitter
and all I got was this Lousy Paper”*

In 2012 Jungh

tion

A Balanced Survey on Election Prediction using
Twitter Data

lication)

Party

Daniel Gayo-Avello

dani@uniovi.es

@PFCdgayo

Department of Computer Science - University of Oviedo (Spain)

May 1, 2012

CDU

CSU

SPD

FDP

Linke

Grüne

Piraten

2.1

34.8

Take-away: estimating elections from tweets suffers from sec-selection bias

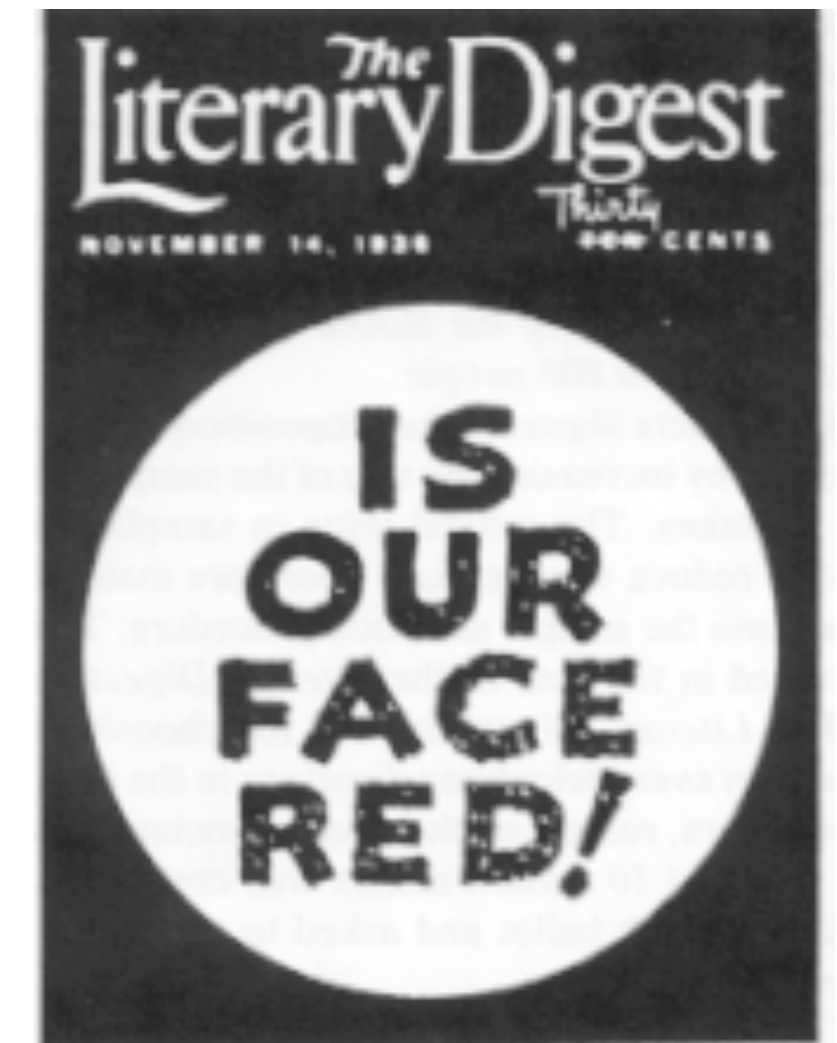
The 1936 Literary Digest Poll

Presidential Election of 1936: Alfred Landon against F. D. Roosevelt

Literary Digest successfully predicted elections since 1916

“Once again, [we are] asking more than ten million voters – one out of four, representing every county in the United States – to settle November's election in October.

Next week, the first answers from these ten million will begin the incoming tide of marked ballots, to be triple-checked, verified, five-times cross-classified and totaled. When the last figure has been totted and checked, if past experience is a criterion, the country will know to within a fraction of 1 percent the actual popular vote of forty million [voters].”



Predicted 57%-43% for Landon but Roosevelt wins with 62%

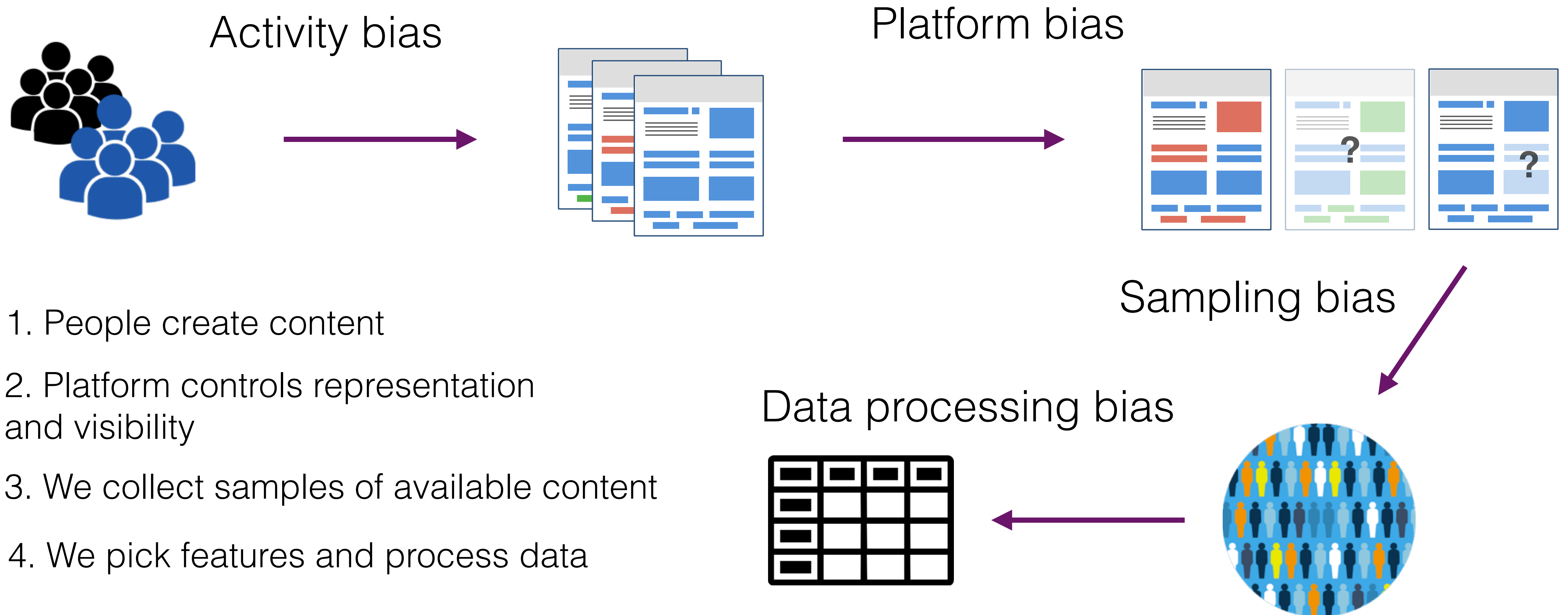
Selection bias as well as non-responsive bias

Bias

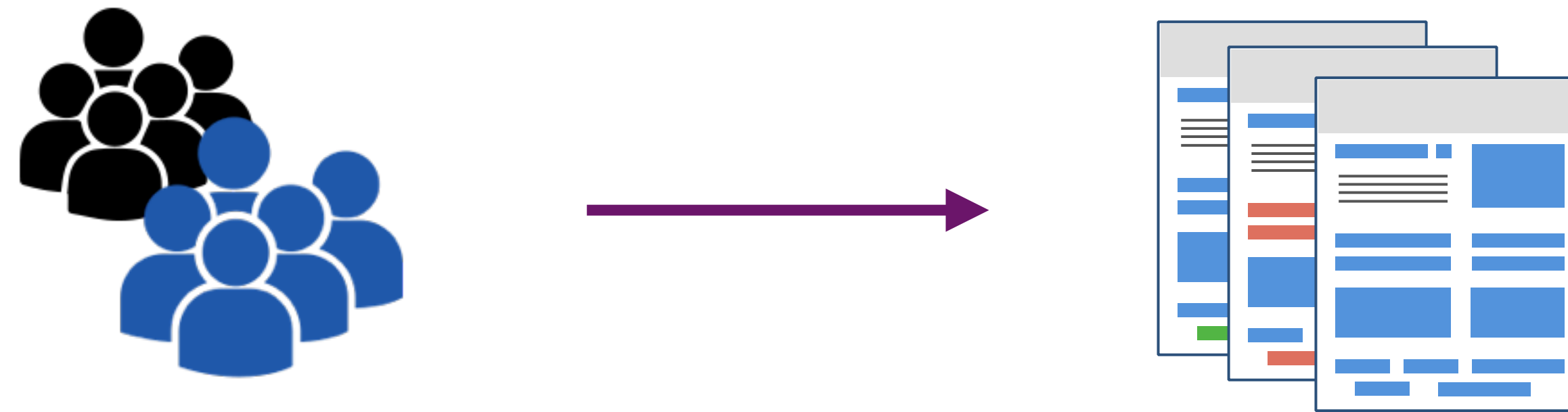
Bias is defined as any tendency which prevents unprejudiced consideration of a question. In research, bias occurs when “systematic error [is] introduced into sampling or testing by selecting or encouraging one outcome or answer over others”

We will overview biases related to sampling, feature selection and data cleaning, (not biases related to testing or analysis).

Bias in online data collection



1. Activity Bias



Examples:

Differences in the **rate** at which users create content

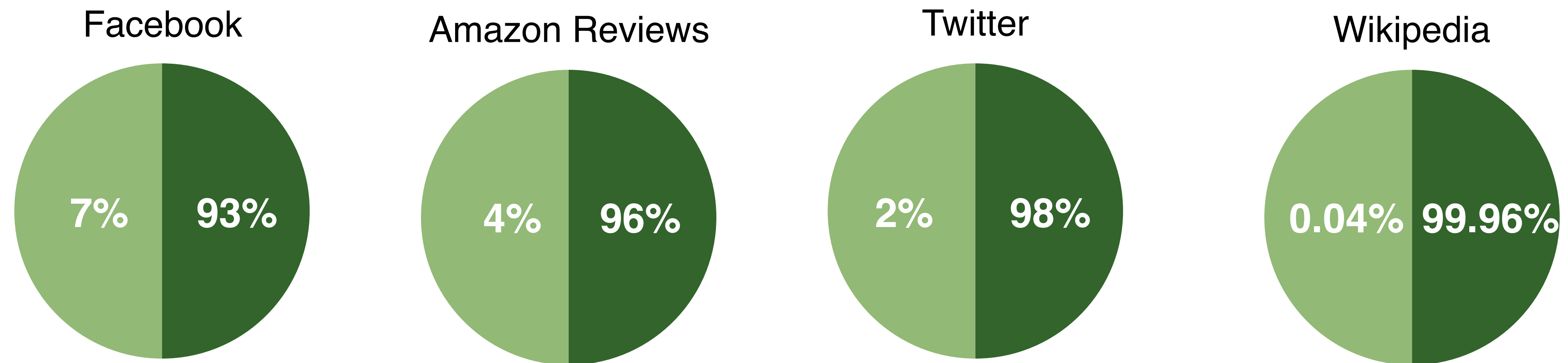
Type of content available depending on the **time of day, week, seasonal changes**

Gender, age, location, etc correlates with type of content created

Fake users, bots, deleted content

Wisdom of the few

Ratio of people creating 50% of content



[[Baza-Yates et.al. Hypertext'15](#)]

Large percent of online content is created by a small number of users.

Fake, spam, non-human?

Bots and organizations create large % of content but do not represent “normal” human behavior

One person, multiple accounts

Deleted accounts or content

[Petrovic et.al & Almuhimedi et.al]

~3% of tweets get deleted over time. Significant differences in the deleted set based on location, amount of reply triggered, sentiment, etc.

Different populations use platforms differently, e.g.:

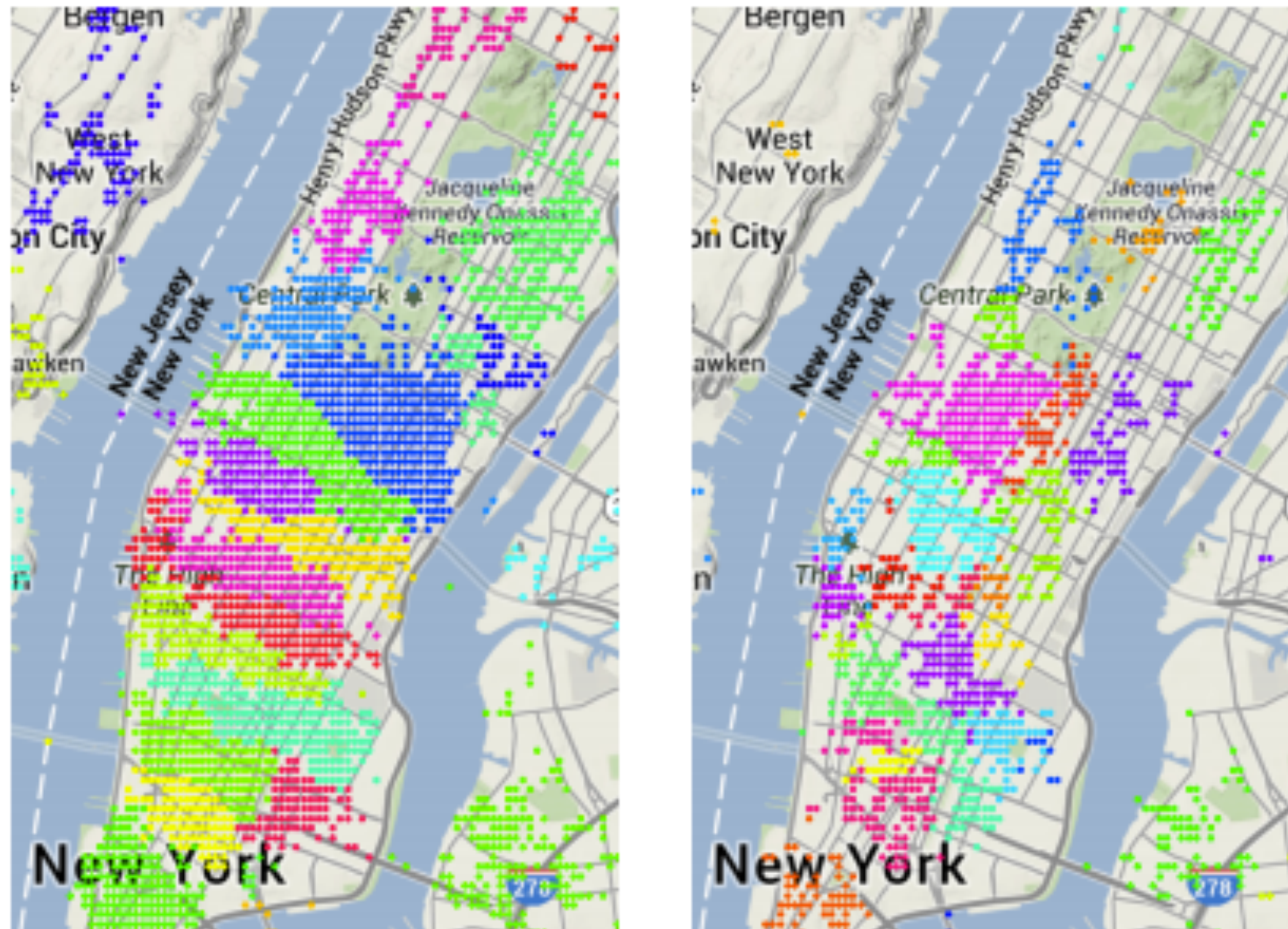
[Hong et al. ICWSM'11]

Users of different languages use Twitter differently:

Germans tend to include URLs and hashtags more often, while Koreans tend to reply to each other more often.

Seasonal differences

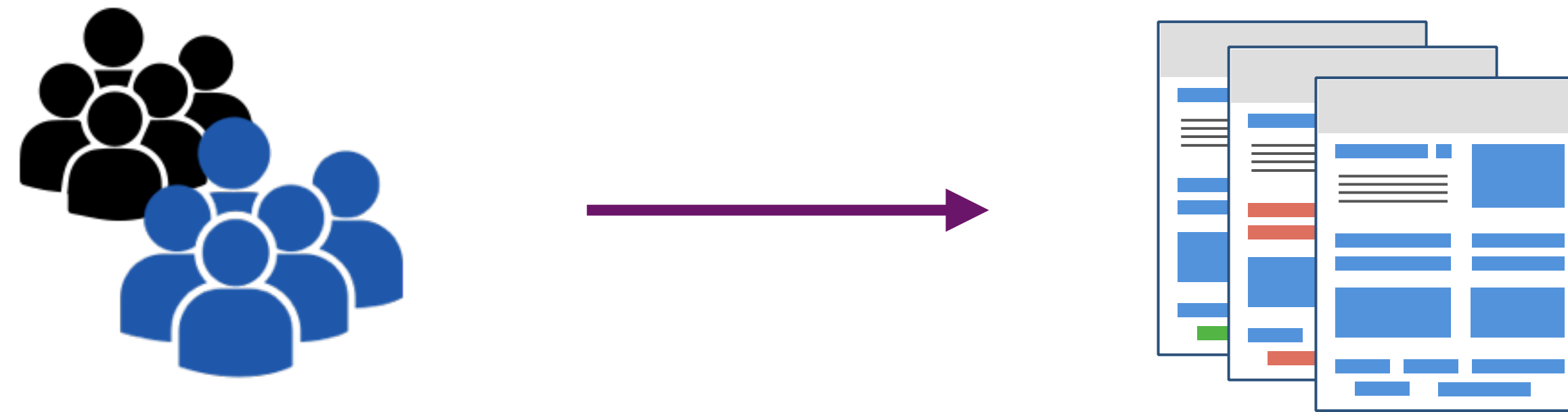
Who talks when about what topics?



Weekdays vs weekends

[Kiciman et al. ICWSM'14]
Neighborhoods Inferred from social media conversations differ depending on context such as time day/night, weekend/weekday.

1. Activity Bias



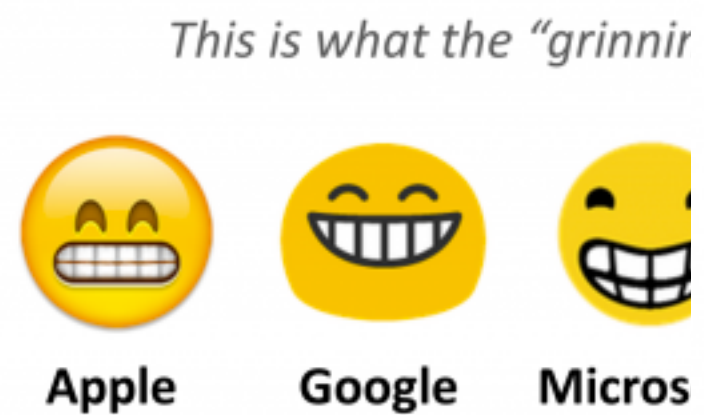
Take-aways:

Explorative stats can be important to discover activity patterns

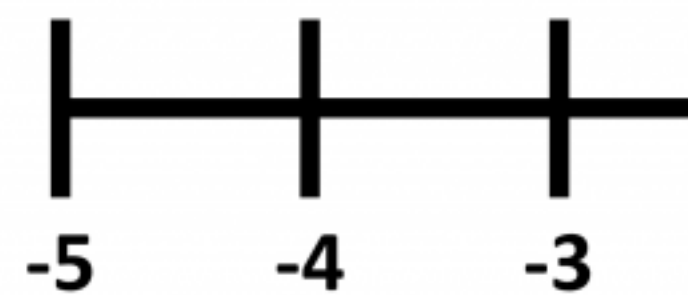
Control for variables that correlate with activity differences

2. Platform bias

Biases rela'

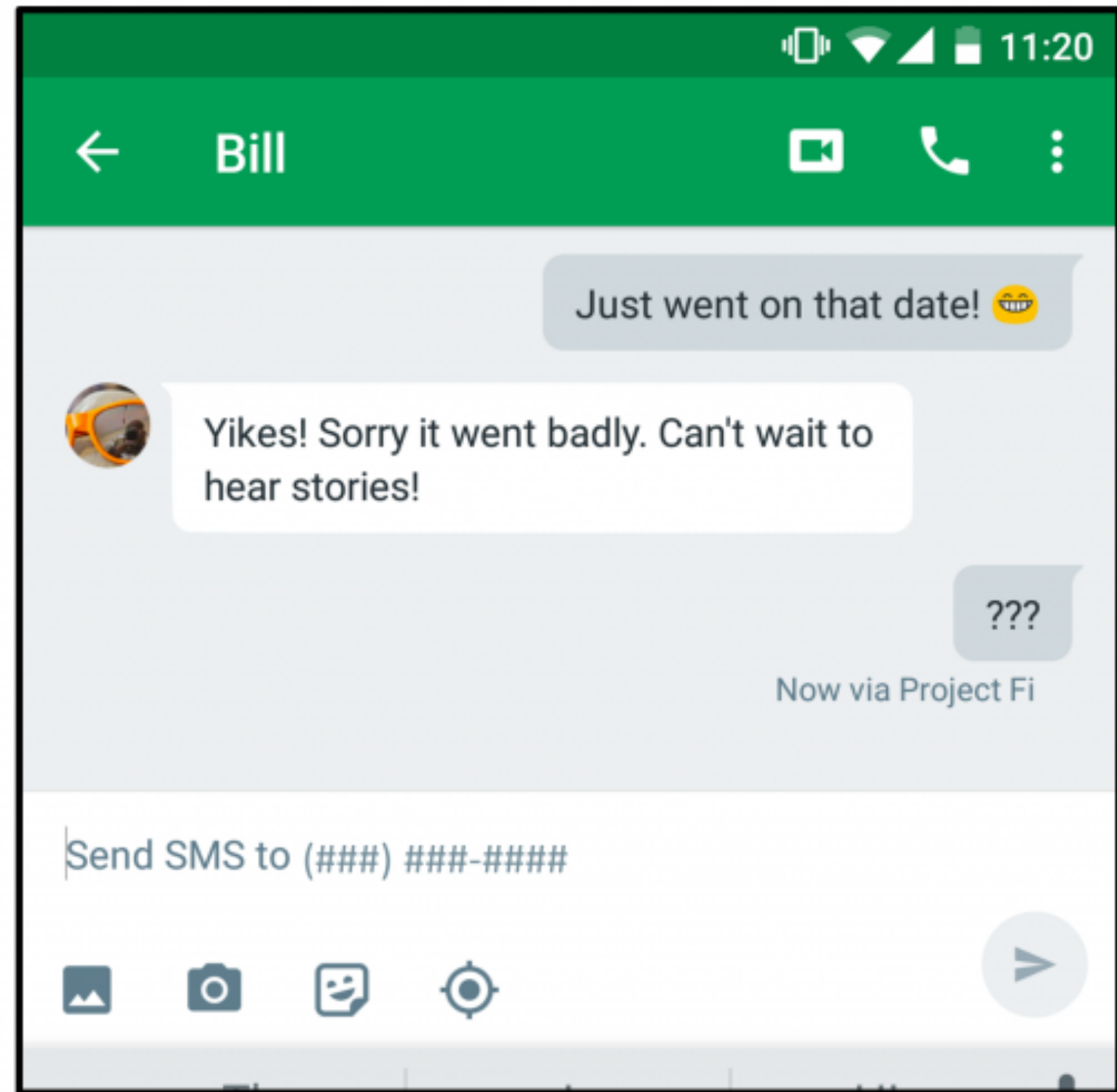


Same Emoji + Dif
For example, if you send th



Negative
(e.g., sad, mad)

Abby using a Google Nexus, texting Bill:



Neutral

Bill using an iPhone, texting Abby:



Positive

(e.g., happy, excited)

ent

[NSM'16]

across
ms can lead
cation.
rings are
different

Algorithms

Platforms continuously change, improve their systems

Helpful features such as autocorrect or recommendation may lead to over-representation of signals, e.g measuring

popularity of a product while it was recommended to you

search behavior while autocorrect influences how people search

traffic patterns while people rely on Google Maps



frogs are

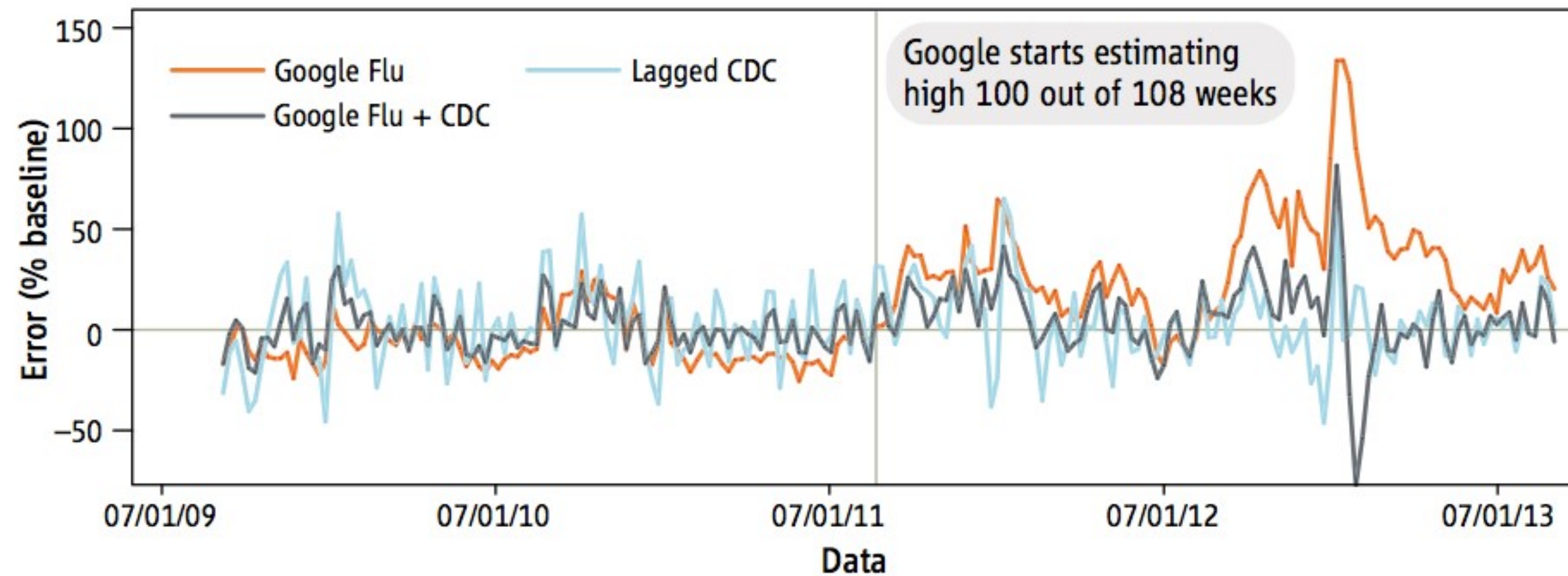
frogs are **gay**

frogs are

frogs are **reptiles**

frogs are **green**

Press Enter to search.

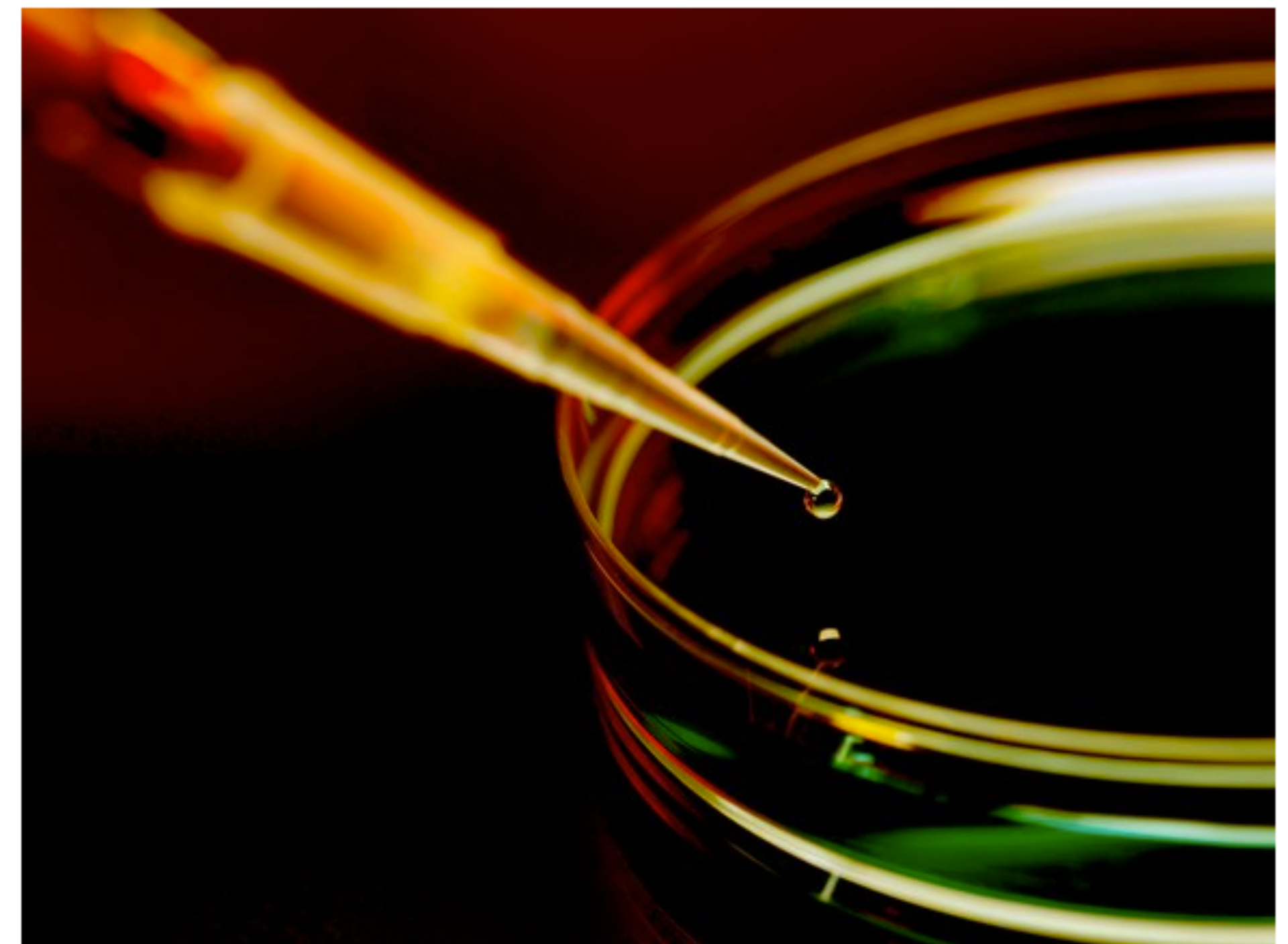


DAVID LAZER AND RYAN KENNEDY SCIENCE 10.01.15 7:00 AM

WHAT WE CAN LEARN FROM THE EPIC FAILURE OF GOOGLE FLU TRENDS

[Lazer et al. Science'13]

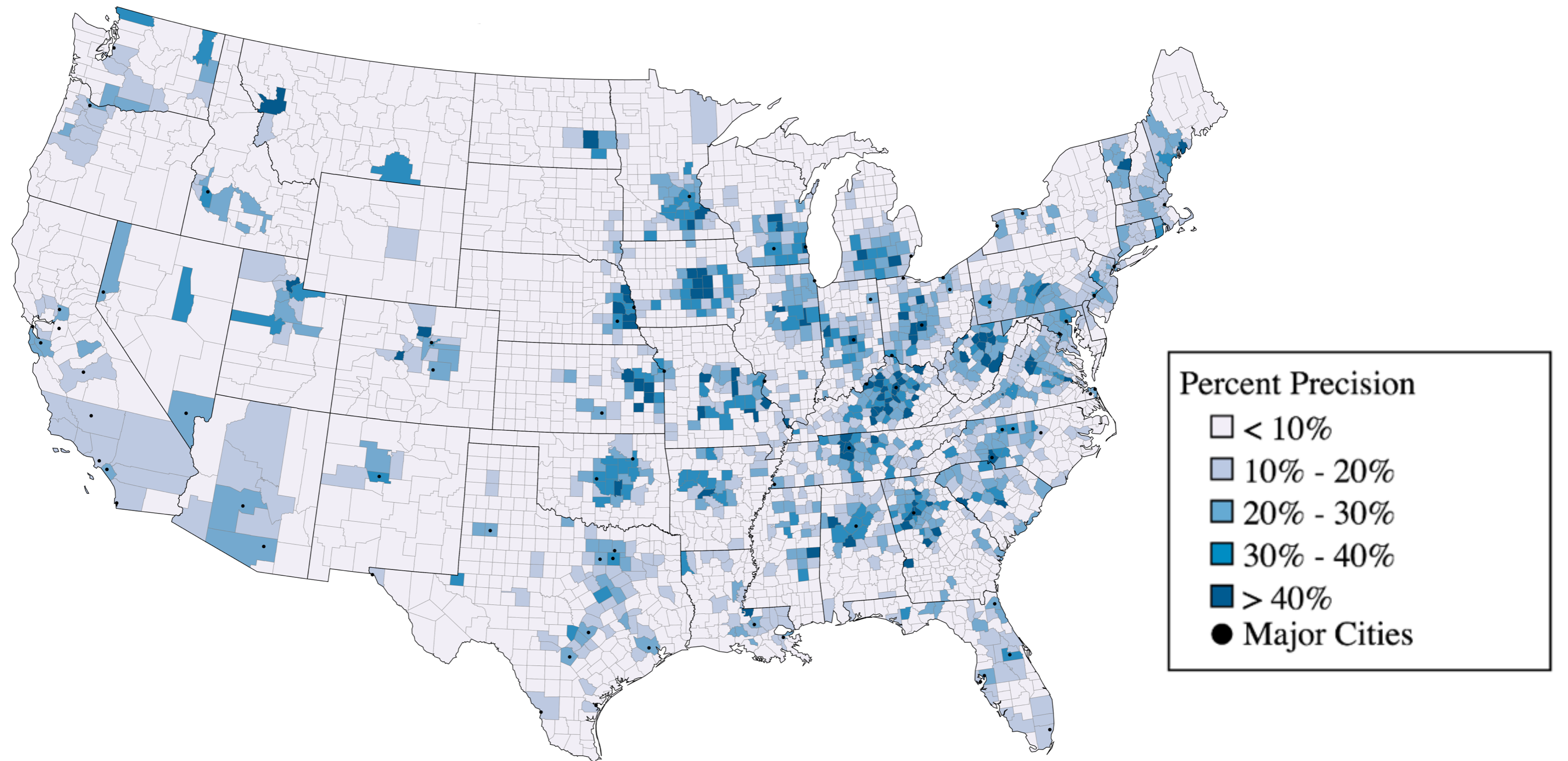
“GFT bakes in an assumption that relative search volume for certain terms is statistically related to external events, but search behavior is not just exogenously determined, it is also endogenously cultivated by the service provider.”



Biased Algorithms

[[Johnson et al. CHI'17](#)]

Big data algorithms trained on social data from online platforms perform significantly worse for underrepresented populations.



Text-based geolocation precision by county

3. Sampling bias

Due to the sampling tool itself API, keyword search, filtering on certain features such as location

Self-selection, loud people are overrepresented

Snowball-sampling might miss small clusters in a graph

Survey vs phone vs social media targets different demographics



Sampling tool

API: Firehose vs Streaming API

[Morstatter et al. ICWSM'13]

Different sampling methods for acquiring Twitter data result in significantly different properties



Keyword search: e.g.: bias due to differences hashtag usage

Proxy populations: real population of Sardinia vs who set their location to

Self-selection

People who willingly participate are already a biased sample

Psychological factors: being particularly interested in the topic, being generally open, trusting, willing to help

Economic factors: having more time, being more exposed to surveys

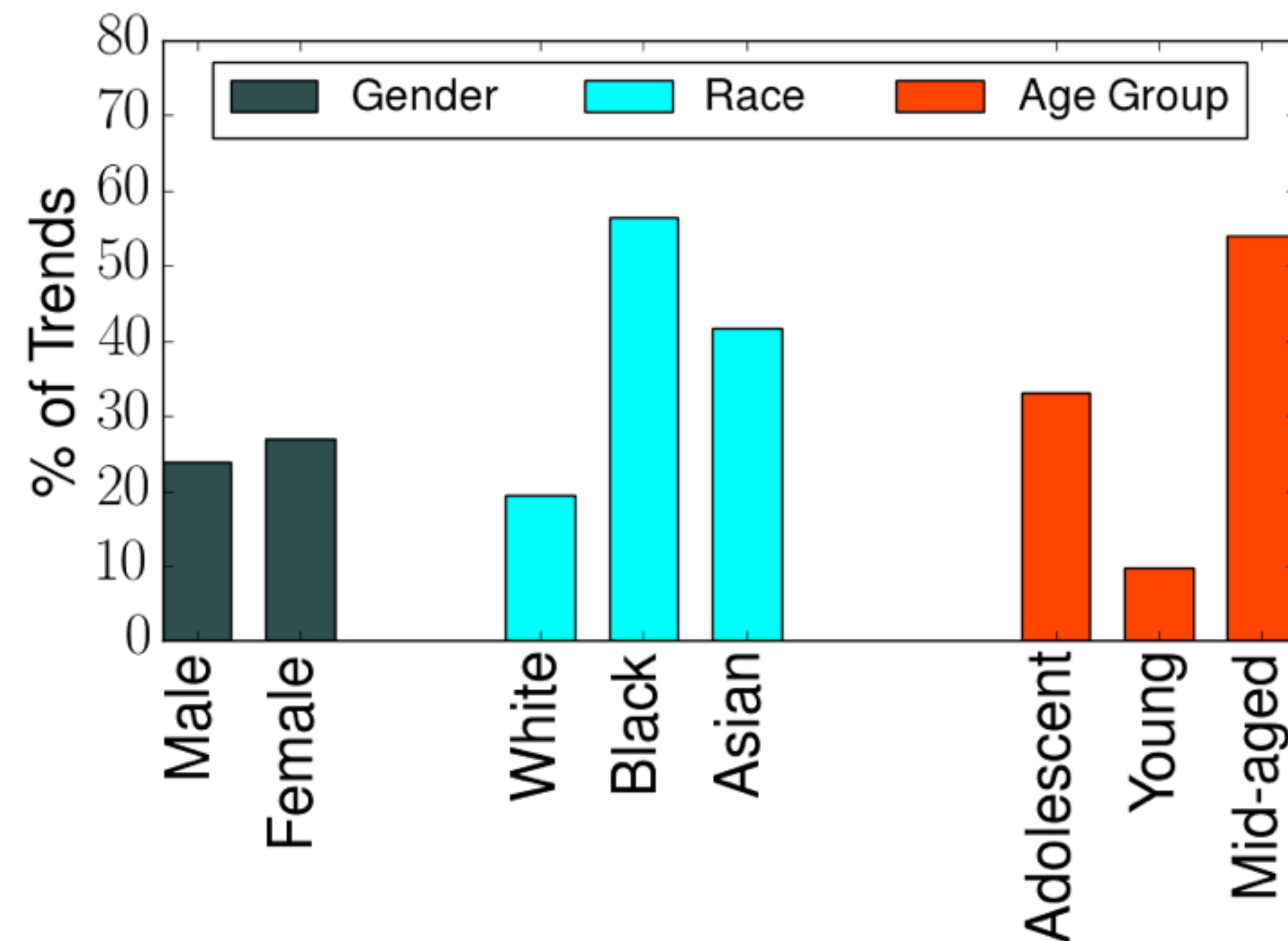
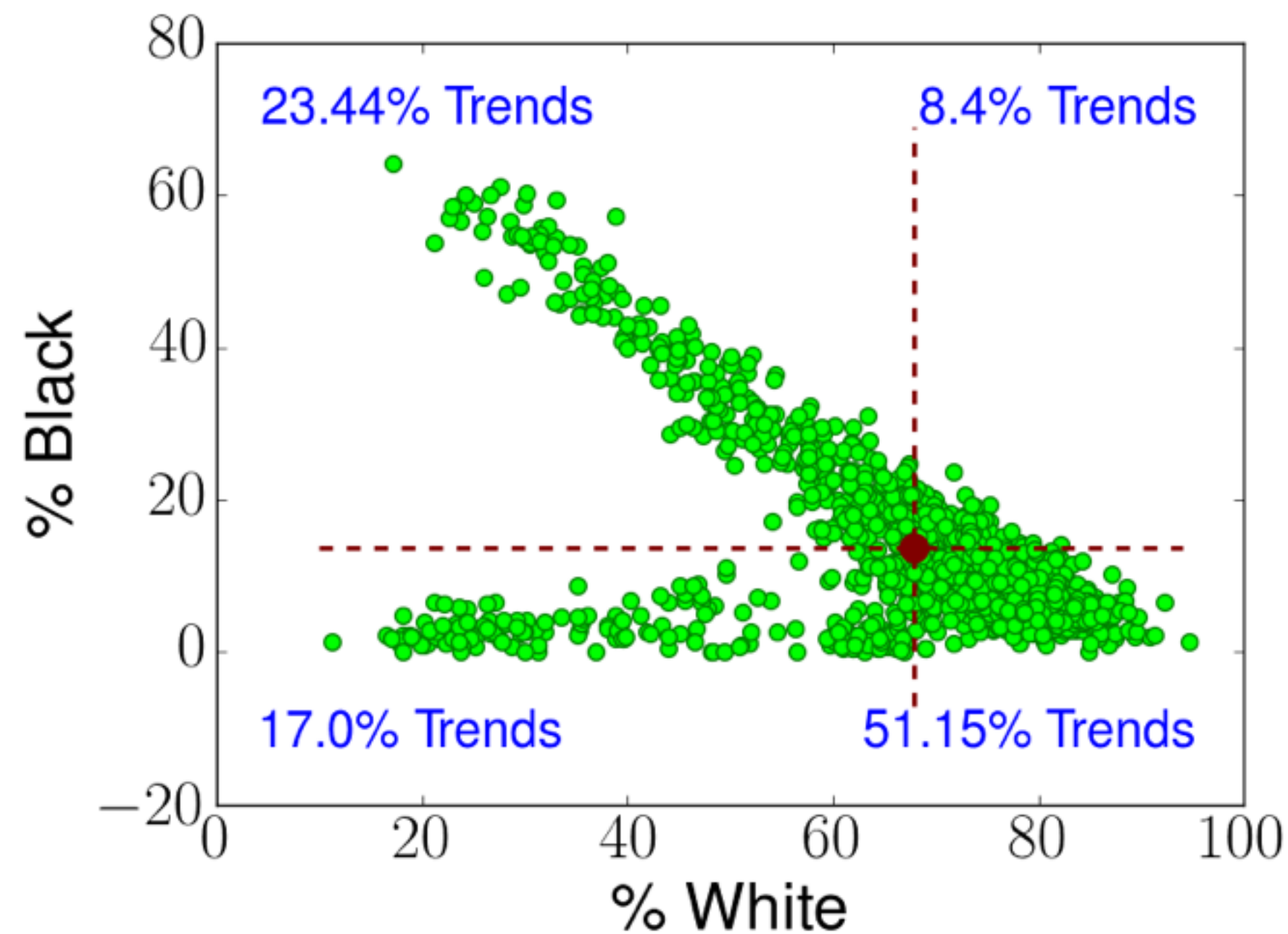
Self-selection is a bitch: non-participatory people are unavailable, thus hard to determine differences in populations

[Yasser et. al. 2014]: More proficient and more involved gamers of WoW are more likely to fill out online surveys.

[J Borjas 1987]: Immigrants' earnings can not be directly compared to US earnings since they are a selected sample, "more able and highly motivated"

The loud crowd

Chakraborty et.al. 2017: Demographics of Twitter users determining trending topics are significantly different from the overall population.



Sampling bias

Take-aways:

Really think about who your population is especially relative to who you are trying to make statements about.

Compare demographic features and other characteristics of the group you end up with to the whole population

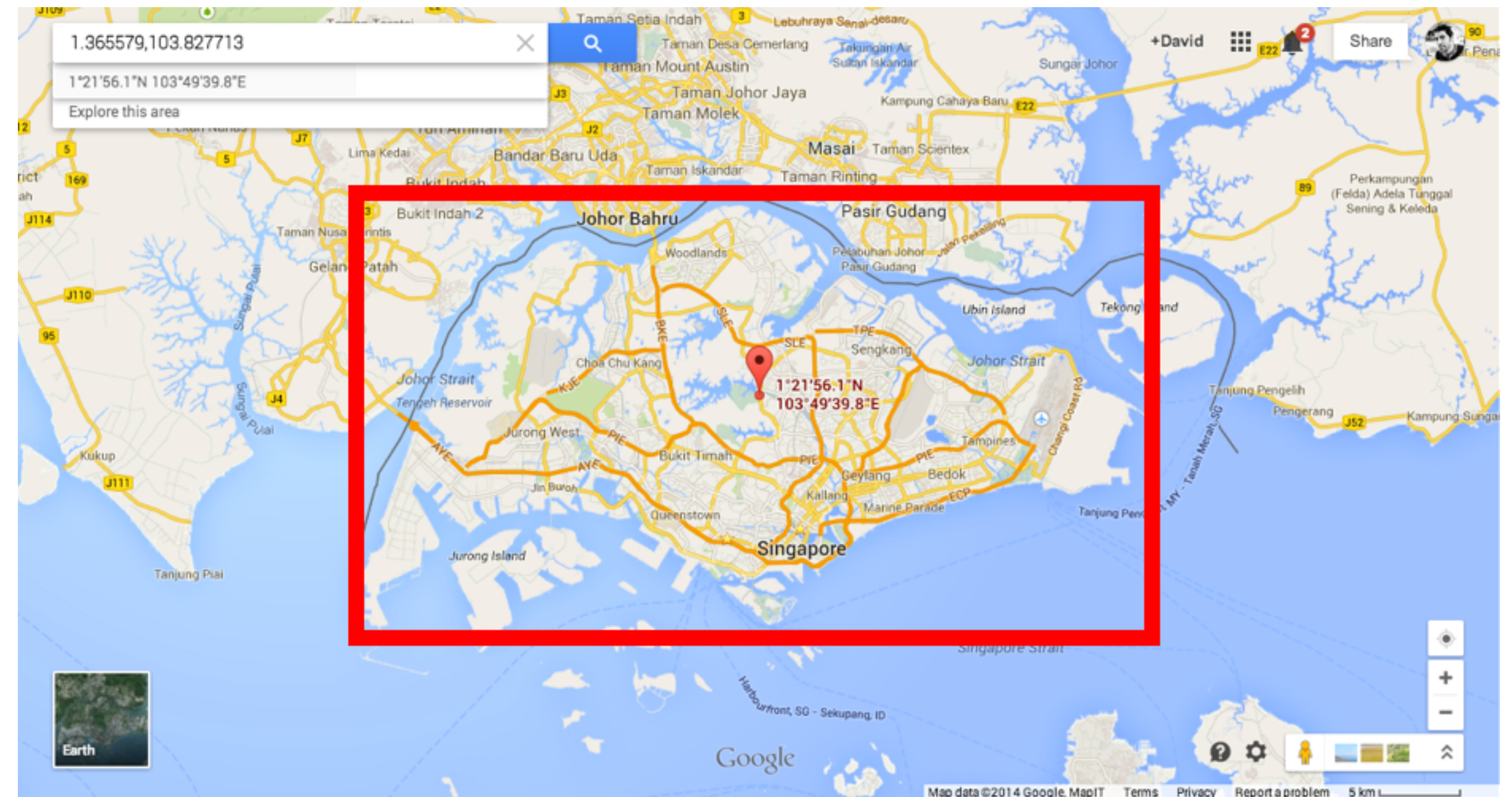
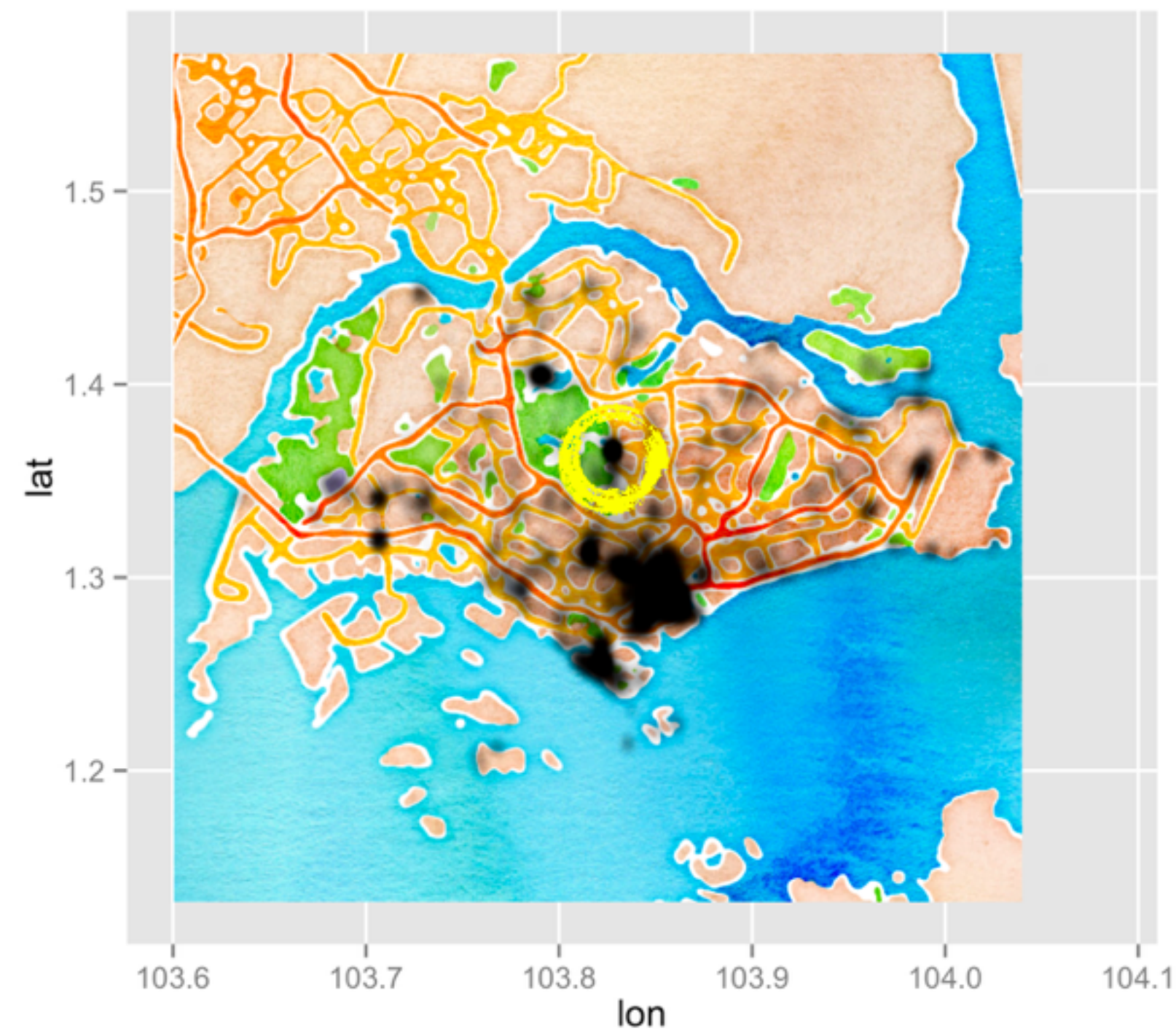
If you have a choice, avoid sampling techniques that require special attitude toward social media



4. Representation bias

Predefined features when filling out profiles, default values, min-max

Geographic bounding boxes



4. Representation bias

Predefined features when filling out profiles, default values, min-max

Geographic bounding boxes

The Guardian logo, featuring the word "theguardian" in a white, lowercase, sans-serif font on a dark blue rectangular background.

Kansas family sues mapping company for years of 'digital hell'

Geolocation company's glitch sent police and angry businesses to a remote Kansas farm looking for criminals, and now the residents want compensation

Data processing

Filtering the data:

removing highly active or inactive users

removing location that can not be recognized

removing languages, characters hard to parse

Annotation bias:

some characteristics are easier to recognize, gender vs race

usernames vs real names

Summary

Big data is great
but handle it with caution :)



"I'm too busy recommending things to experience them myself."