Political Analysis of Social Media Data

Representativeness

Instructor: Gregory Eady
Office: 18.2.10
Office hours: Fridays 13-15

Social media data: What are they good for?

The potential uses of social media data are massive

1. **Measurement opportunities**
    - Public opinion and political behavior voting
        - Vote choice
        - Anti-immigrant sentiment
        - ISIS sympathies
        - Protest turnout
        - Elite communications
        - Network behavior
        - Conflict dynamics

The potential uses of social media data are massive

2. Social media opportunities

- Examine effects of social media itself
  - Polarization
  - Incivility
  - Political campaigning
  - Political knowledge
  - Censorship
  - Advertising
  - Radicalization
- Differences across platforms
- Effects of interventions (e.g. labeling fake news, banning types of speech, increasing the character limit)
- Different behavior toward some groups than others (e.g. by partisanship, age, gender, race, religion)

Benefits of social media data

Benefits of social media data

1. **Unobtrusive data collection**
   - Minimizes changes in behaviors from being observed (Hawthorne effects)
   - Minimizes social desirability bias
   - Collection is done at scale
   - Can capture the network structure

Benefits of social media data

2. Homogeneous content and format

- Similar text lengths
- Similar use of language and meta data
- Consistent norms in usage (e.g. hashtags)
- Many actors interact in similar ways (e.g. retweets, sharing)

Benefits of social media data

3. **Data granularity**
   - Across space
   - Across time
   - Across topic
   - Across individuals

Benefits of social media data

4. Widespread use

- Vast majority of politicians use social media (especially in Western democracies)
- Majority of adult population uses social media

Benefits of social media data

5. **Offline attitudes and behaviors have online analogs**
    - Ideology can be capturde online and offline
    - Attitudinal homophily
    - Diurnal activity
    - Personality can be measured from, for example, Facebook likes

And yet...

04-24-19

Study confirms: Twitter is not real life

This survey from the Pew Research Center uses hard data to explain some stark differences between the extremely online and those who tune Twitter out.

Twitter Isn't Real Life (if You're a Democrat)

The online left doesn't like Joe Biden. Voters seem to.

By Michelle Goldberg
Opinion Columnist

May 13, 2019

1675

TECHNOLOGY

Twitter Is Not America

A new Pew study finds a gulf between the general population and Twitter users.

ALEXIS C. MADRIGAL APRIL 24, 2019

Or is it...

SOCIAL MEDIA | APR. 12, 2019

For Democrats, Twitter Is Not Real Life. But It Could Be.

By Eric Levitz *@EricLevitz*

Maybe Twitter *Is* Real Life

If anything, political Twitter is underrated

Nicholas Grossman Follow

Oct 11, 2019 · 6 min read ★

Opinion

Twitter Is Real Life

Elites just pretend it's not.

By Charlie Warzel

Mr. Warzel is an Opinion writer at large.

Feb. 19, 2020

334

Is social media real life?

Is social media representative?

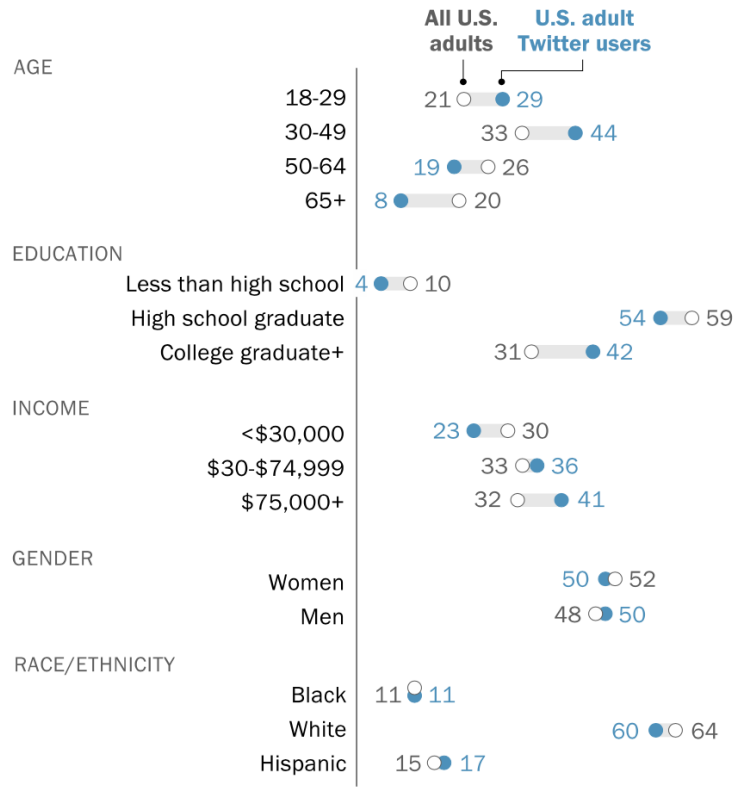
Representativeness concern 1 (the obvious one):

Representativeness concern 1 (the obvious one):

1. Those who use social media are not representative of the general population
   - Estimates of public opinion will be skewed
   - Estimates of prevalence of political behaviors will be skewed
   - Is the case both between groups and within them
2. Larger samples do *not* somehow drown out bias

Twitter users are younger, more highly educated and wealthier than general public

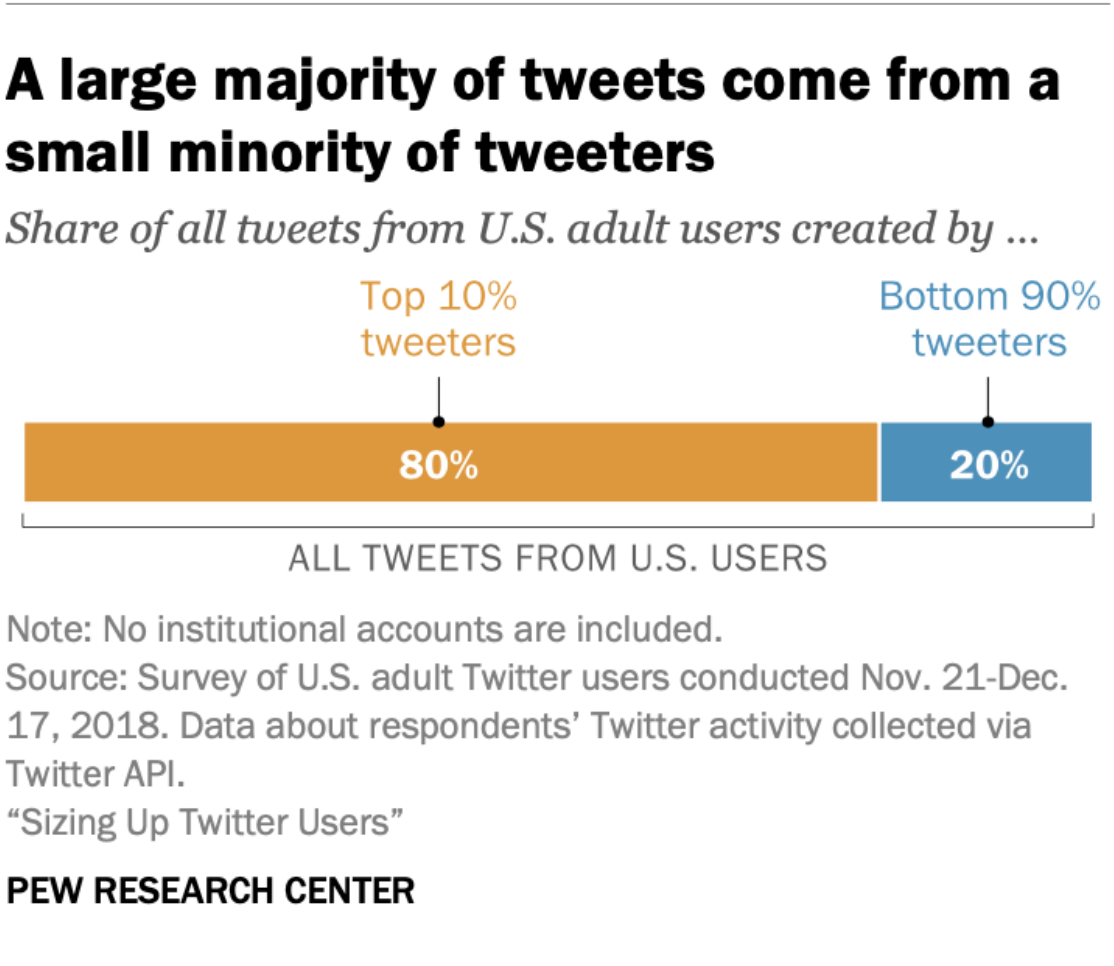
% of _____ who are ...



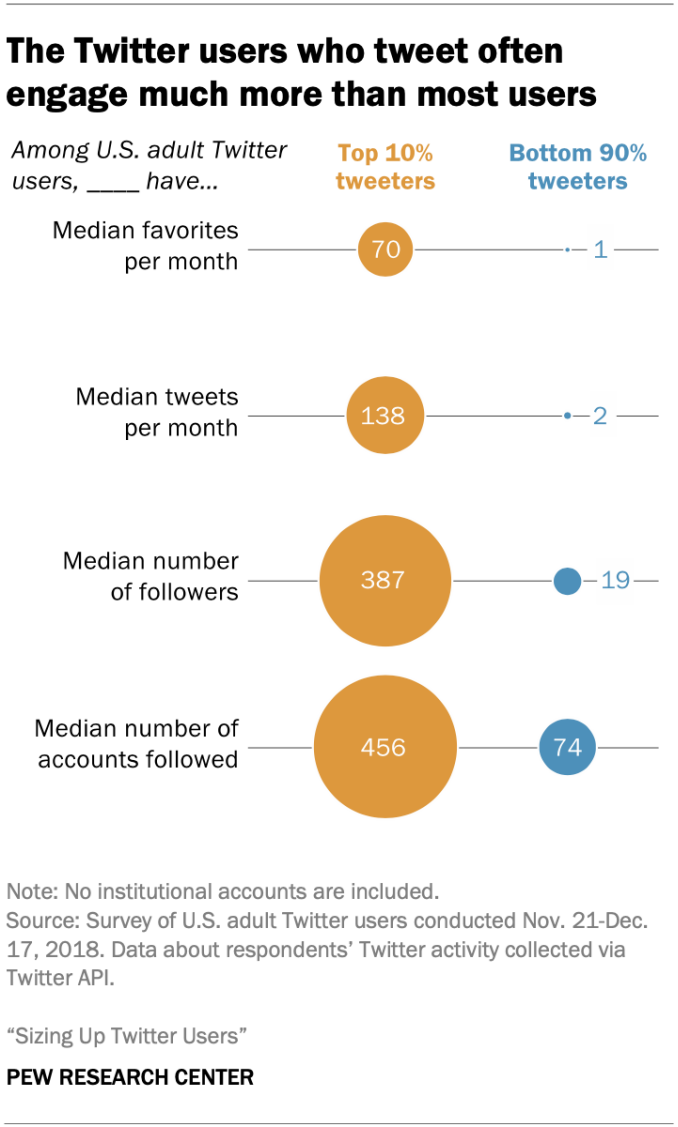
Note: Whites and blacks include only non-Hispanics. Hispanics are of any race.
Source: Survey of U.S. adult Twitter users conducted Nov. 21-Dec. 17, 2018, and survey of U.S. adults conducted Nov. 7-11, 2018.
"Sizing Up Twitter Users"

PEW RESEARCH CENTER

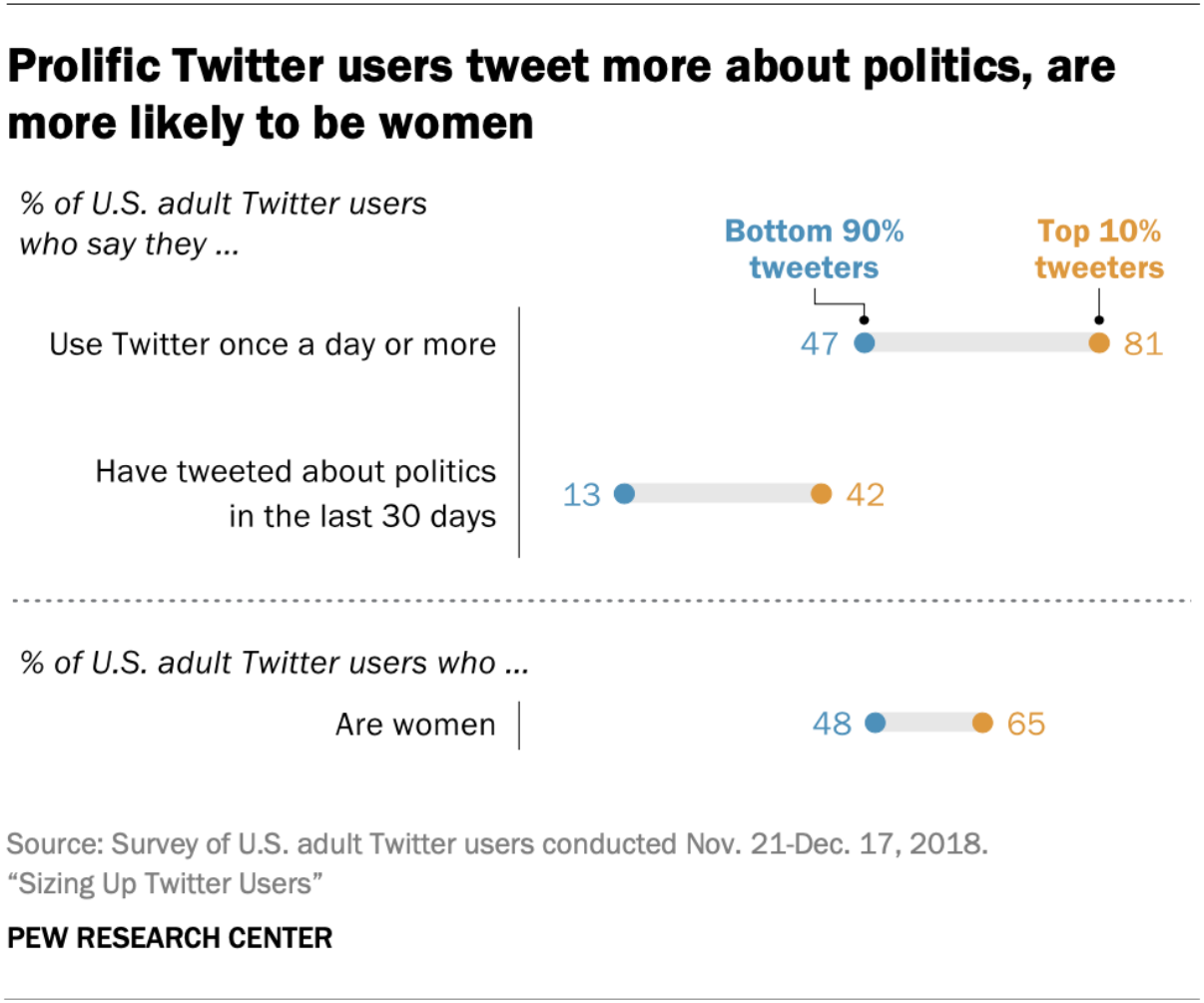
And this active population is skewed further:



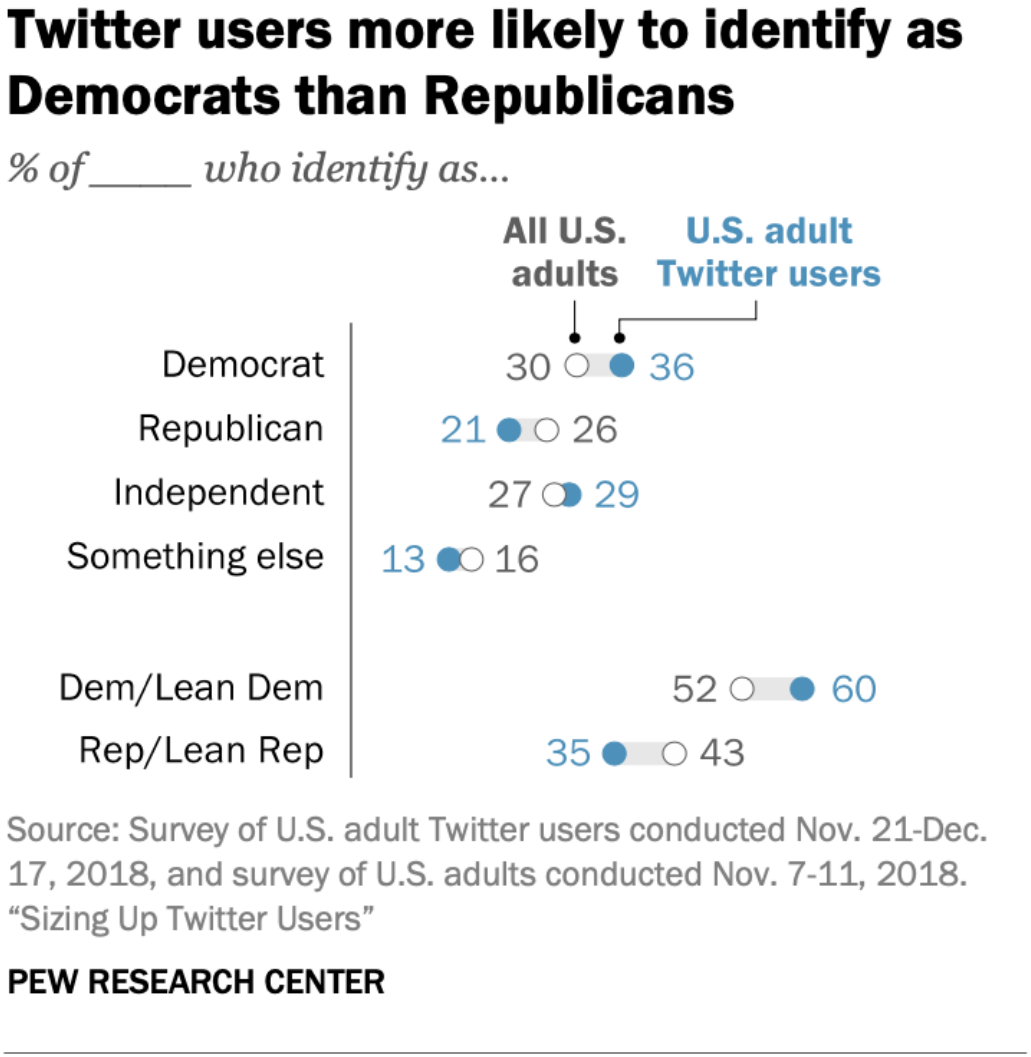
And this active population is skewed further:



And this active population is skewed further:



Moderate differences in overall partisan skew



But this can hide much within-group bias...

Democrats who do not post political content on social media sites:

Identify themselves as moderates or conservatives

29% of Democrats on social media

53% of other Democrats

Democrats who do not post political content on social media sites:

Say political correctness is a problem in the U.S.

48% of Democrats on social media

70% of other Democrats

Democrats who do not post political content on social media sites:

Say they don't follow the news much

27% of Democrats on social media

59% of other Democrats

Democrats who do not post political content on social media sites:

Be African-American

11% of Democrats on social media

24% of other Democrats

Democrats who do post political content on social media are more likely to:

Have a college degree

47% of Democrats on social media

33% of other Democrats

Democrats who do post political content on social media are more likely to:

Be white

71% of Democrats on social media

55% of other Democrats

Democrats who do post political content on social media are more likely to:

Say they have become more liberal in their lifetime

53% of Democrats on social media

30% of other Democrats

Democrats who do post political content on social media are more likely to:

Say they have attended a protest in the last year

28% of Democrats on social media

7% of other Democrats

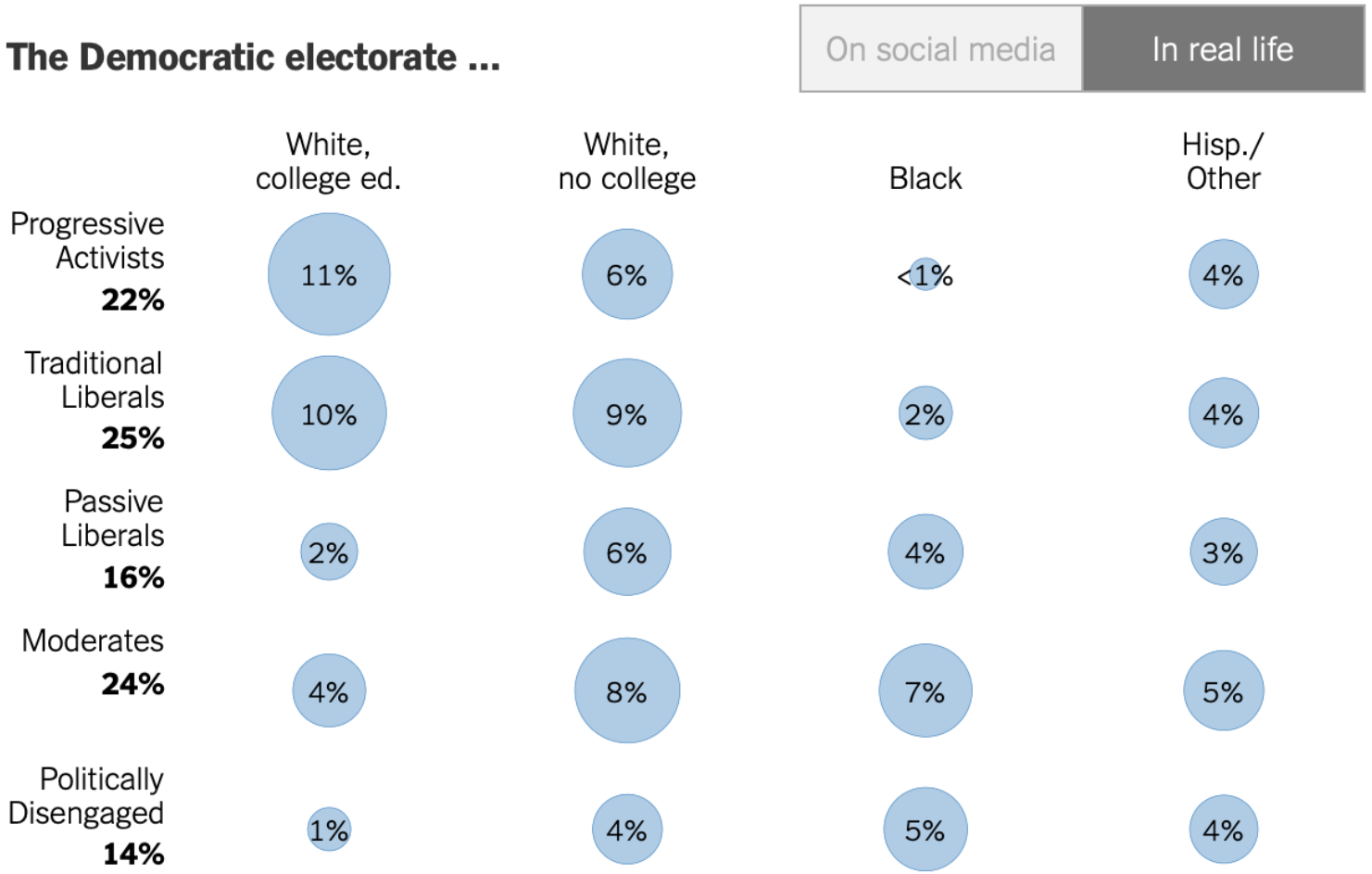
Democrats who do post political content on social media are more likely to:

Say they have donated to a political organization in the last year

45% of Democrats on social media

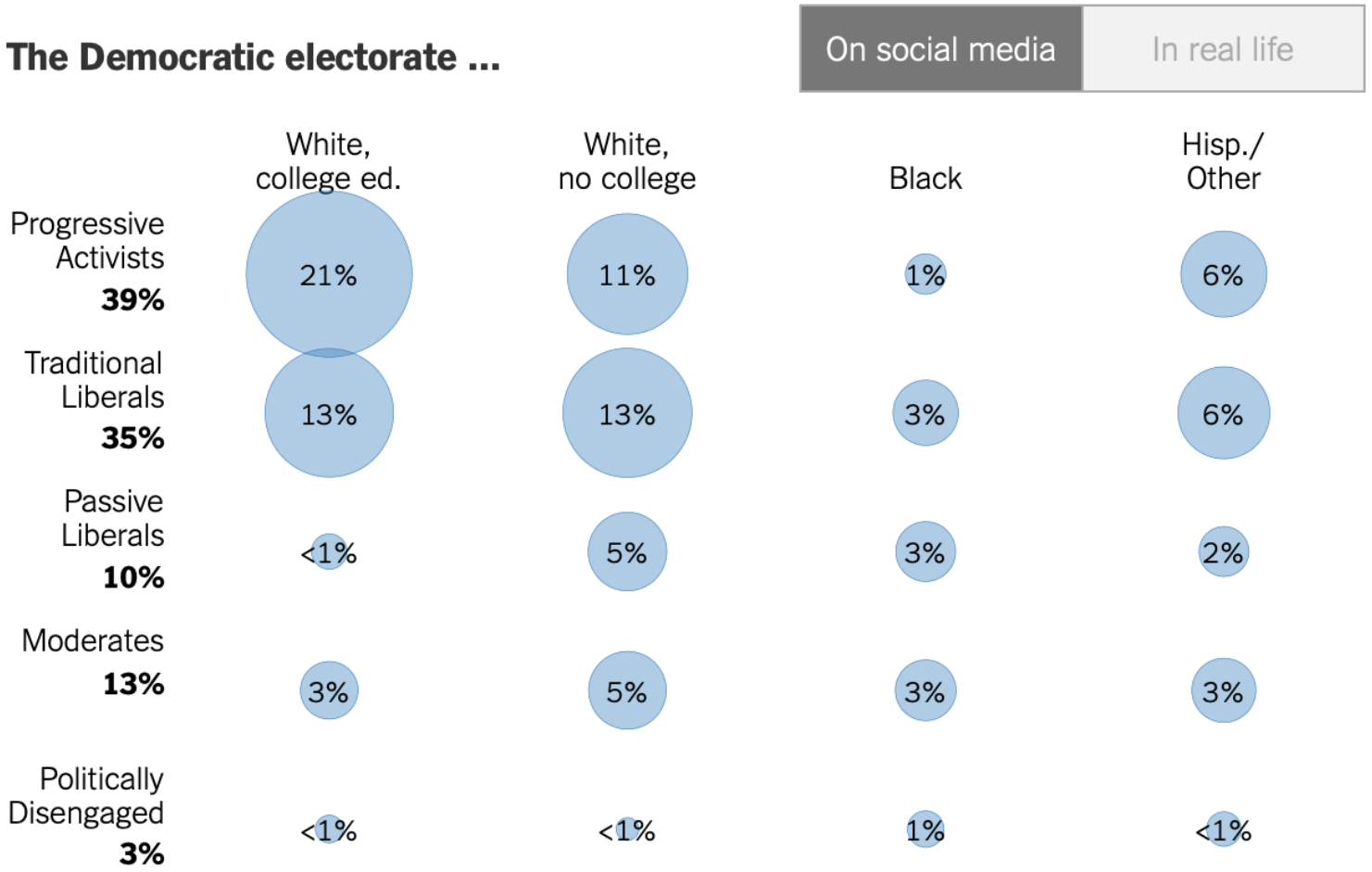
14% of other Democrats

The Democratic electorate in "real life":



Source: Upshot analysis of the Hidden Tribes project. "Moderate" group includes conservative tribes. Probabilistic likely voter weight based on self-reported turnout in 2016 and local elections, and self-reported intention to vote in the 2018 midterm election.

The Democratic electorate on social media:



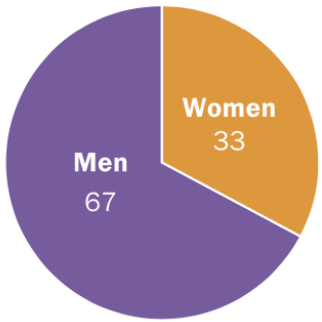
Source: Upshot analysis of the Hidden Tribes project. "Moderate" group includes conservative tribes. Probabilistic likely voter weight based on self-reported turnout in 2016 and local elections, and self-reported intention to vote in the 2018 midterm election.

Unrepresentativeness where you might or might not expect...

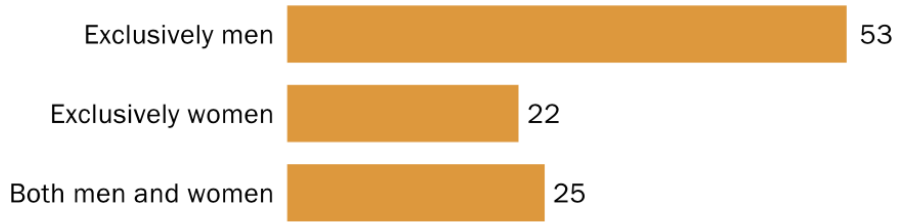
Men appear more than women in news images on Facebook when it comes to both individuals and groups of people

Among images on Facebook from 17 national news organizations ...

% of individuals shown who are ...



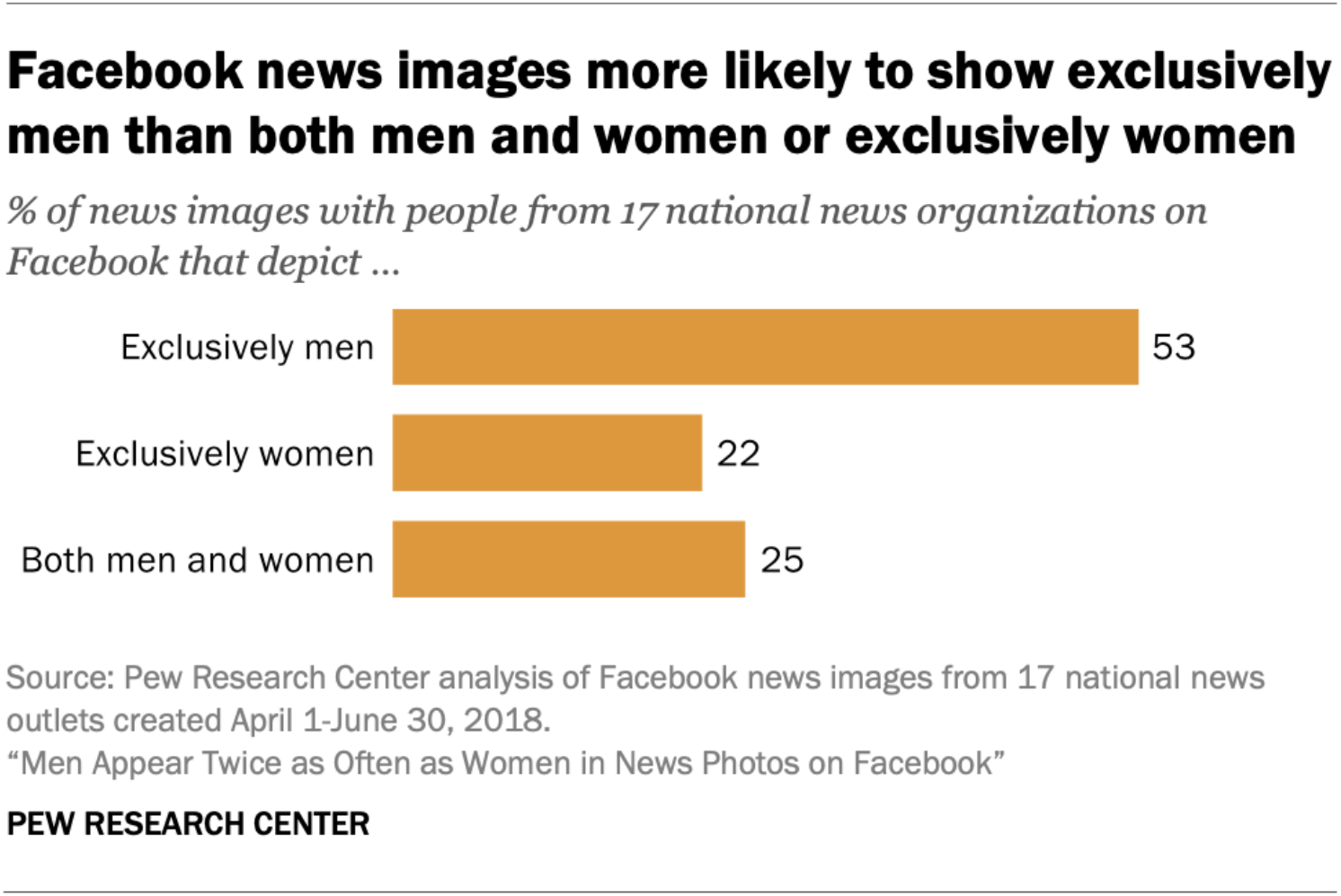
% of news images with people that depict ...



Source: Pew Research Center analysis of Facebook news images from 17 national news outlets created April 1–June 30, 2018.
"Men Appear Twice as Often as Women in News Photos on Facebook"

PEW RESEARCH CENTER

And unrepresentativeness itself can be politically interesting to study



And unrepresentativeness itself can be politically interesting to study

Parents mention sons more often than daughters on social media

Elizaveta Sivak[a,1] **and Ivan Smirnov**[a,1,2]

[a]Institute of Education, National Research University Higher School of Economics, Moscow 101000, Russia

Edited by Laura K. Nelson, Northeastern University, Boston, MA, and accepted by Editorial Board Member Mary C. Waters December 21, 2018 (received for review March 29, 2018)

Gender inequality starts early in life. Parents tend to prefer boys over girls, which is manifested in reproductive behavior, marital life, and parents' pastimes and investments in their children. While social media and sharing information about children (so-called "sharenting") have become an integral part of parenthood, whether and how gender preference shapes the online behavior of users are not well known. In this paper we use public posts made by 635,665 users from Saint Petersburg on a popular Russian social networking site, to investigate public mentions of daughters and sons on social media. We find that both men and women mention sons more often than daughters in their posts. We also find that posts featuring sons receive more "likes" on average. Our results indicate that girls are underrepresented in parents' digital narratives about their children, in a country with an above-average ranking on gender parity. This gender imbalance may send a message that girls are less important than boys or that they deserve less attention, thus reinforcing gender inequality from an early age.

on our sample and data collection). We then identified posts with mentions of children by examining posts that contained the words "daughter" and "son," along with their different forms, e.g., "dochenka" (daughterling) or "sooooooon" (see *Materials and Methods* for details). Common topics for such posts included celebrations of different achievements and important events (e.g., births and birthdays or starting and finishing school); expression of love, affection, and pride; and reports on spending time with the children (see Fig. 1 for illustrative examples and *SI Appendix* for more information about common topics).

We computed the proportion of female and male users from each cohort who mentioned sons or daughters in their posts at least once, along with the average number of mentions of children for these users. In our analysis, we used various definitions for "mentions of children" to ensure that the results were not influenced by a specific choice of words (*Materials and Methods*). We also collected information about the num-

And unrepresentativeness itself can be politically interesting to study

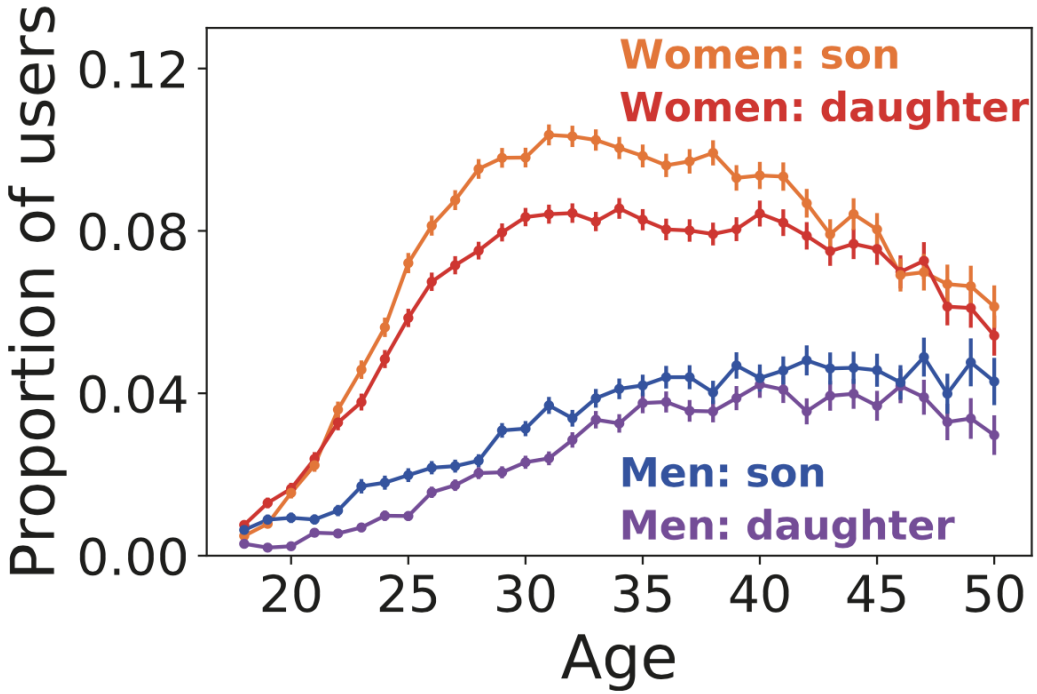


Fig. 2. The proportion of users who mentioned children in their public posts at least once in 2016. Sons are mentioned by a larger proportion of both men and women. Vertical bars are standard errors.

More generally, is measuring public opinion that is representative of the general population a lost cause?

- A heavily caveated "no"
- High degree of skepticism is warranted
- We're past the optimism of the early 2010s when some thought social media would be the new way to gauge general public opinion
- Typically lack necessary demographic/attitudinal data about users to weight to known population values (e.g. gender, age, partisanship)
- BUT, adjustment is still possible...

AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

Predicting and Interpolating State-Level Polls Using Twitter Textual Data

Nicholas Beauchamp Northeastern University

Abstract: *Spatially or temporally dense polling remains both difficult and expensive using existing survey methods. In response, there have been increasing efforts to approximate various survey measures using social media, but most of these approaches remain methodologically flawed. To remedy these flaws, this article combines 1,200 state-level polls during the 2012 presidential campaign with over 100 million state-located political tweets; models the polls as a function of the Twitter text using a new linear regularization feature-selection method; and shows via out-of-sample testing that when properly modeled, the Twitter-based measures track and to some degree predict opinion polls, and can be extended to unpolled states and potentially substate regions and subday timescales. An examination of the most predictive textual features reveals the topics and events associated with opinion shifts, sheds light on more general theories of partisan difference in attention and information processing, and may be of use for real-time campaign strategy.*

TABLE 1 Accuracy in Matching Out-of-Sample Text-Predicted Polls to True Polls

								Elastic Net[c]	
	M1	**M2**	**M3**	**M4**	**M5**	**Random Forest**	**SVM**	**$\lambda_1 =$ 0.001**	**$\lambda_1 =$ 0.1**
Twitter text	×		×		×	×	×	×	×
State fixed effects		×	×	×	×	×	×	×	×
Time trend				×	×	×	×	×	×
MAE (smoothed)[a]	1.91	0.60	0.53	0.54	**0.51**	1.53	3.53	0.88	3.76
MAE (real)[a]	2.16	1.38	1.32	1.30	**1.27**	1.81	2.76	1.53	3.21
R^2 Pooled[b]	0.77	0.98	0.98	0.98	**0.98**	0.90	0.19	0.95	0.01
R^2 Within[b]	0.03	0.19	0.36	0.37	**0.40**	0.09	0.07	0.08	0.22

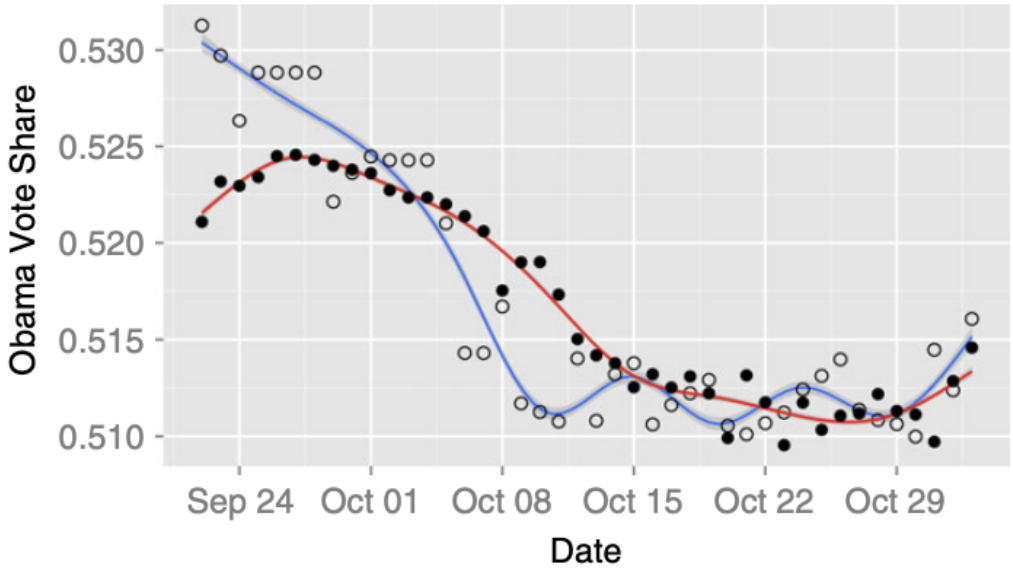
Note: N = 24 states × 42 days = 1008. × = variable included in model. All variables in M1–M5 are significant at $p < .00001$ (cluster-robust standard errors). Best scores are in bold.

[a]MAE = mean absolute error (percentage points) between polls (real or smoothed) and predictions.

[b] R^2 Pooled = variance across all observations; Within = variance within states.

[c]Elastic net performance optimal at $\lambda_2 = 0$ for all λ_1.

FIGURE 3 Predicted and Actual Polling for Ohio



Note: Open circles indicate polls; filled circles indicate text-based predictions.

The upshot

- Is a bit of room to use social media data to approximate general public opinion
- But, in general, be clear about your population of interest and the limitations of your data
  - e.g. Simply interested in opinion on a specific social media platform
  - e.g. Changes across time or geography on social media will approximate shifts among the general public (for instance, Alrababa'h et al. 2019)
- Insufficient critique to simply say that social media data are unrepresentative
  - Provide a reason why unrepresentativeness itself might explain the variation that is being explained, and why that is a problem

Representativeness concern 2: APIs

2. Social media APIs don't necessarily return a representative sample of social media data *themselves*
    - E.g. the streaming Twitter API does not return a representative sample of your search query
        - The API is a black box with respect to representativeness
    - Further, generating a random sample of users (or those specific to a given, say, country) is extremely challenging

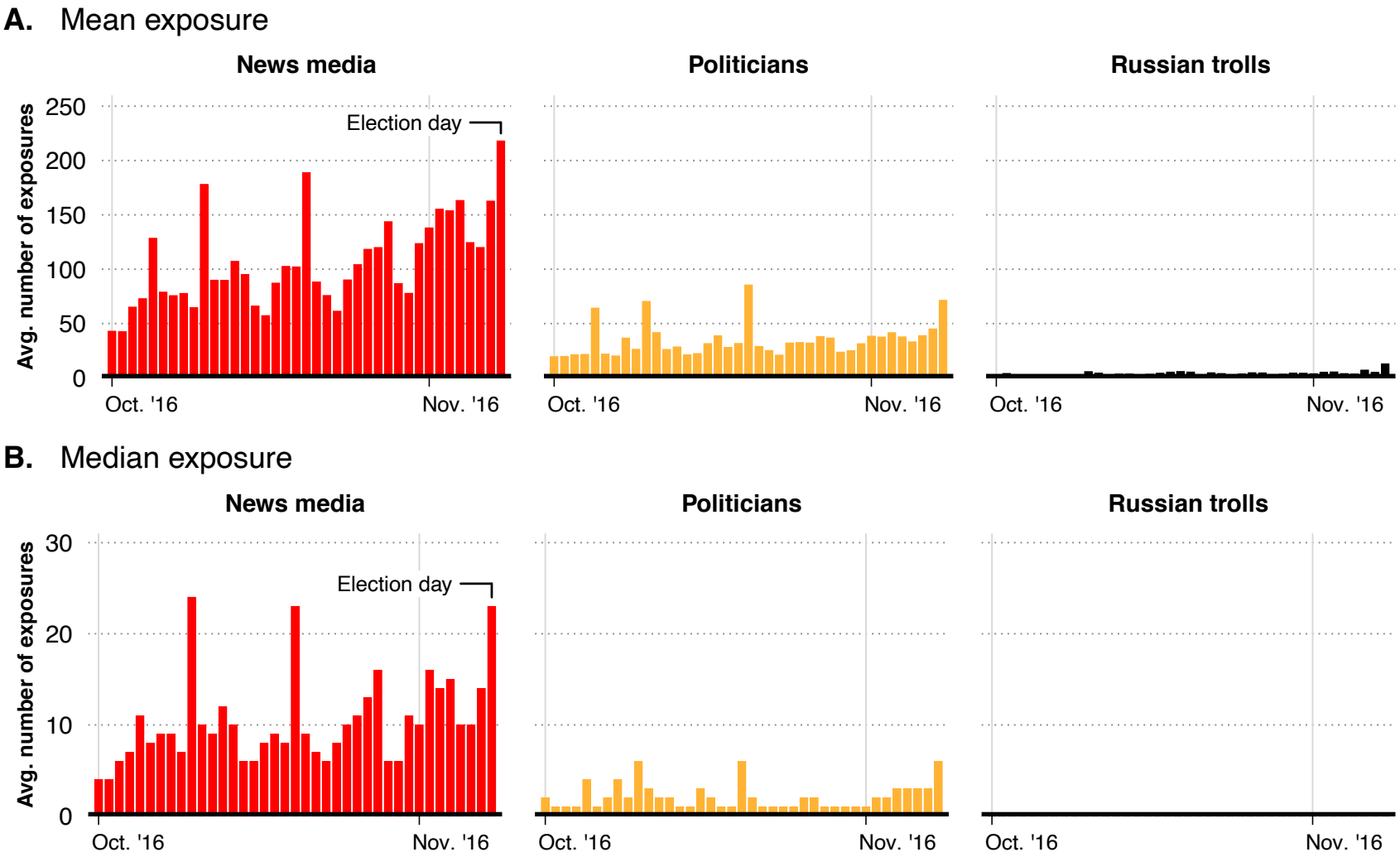
Possible solutions?

- Little is possible with respect to the representativeness offered by an API
- Need a representative samples of users? Re-think your population of interest
  - Population of minimally politically interested users?
    - Collect followers of all politicians in a given country
    - Collect followers of all political news media accounts

Representativeness concern 3: Base rates

3. Absolute numbers of interactions and behaviors on social media can be deceptive
    - We almost never know how many people *saw* a post
    - Problematic because we often can't easily infer influence from absolute numbers
        - Topics of global interest can generate many clicks outside of a country of interest, even if there is narrow interest within the country
        - Tweets by Russian trolls were seen by 1.4 million users in the US. Seems like a lot!
    - What to do: Put numbers in relative context where possible

Exposure to Russian trolls in context

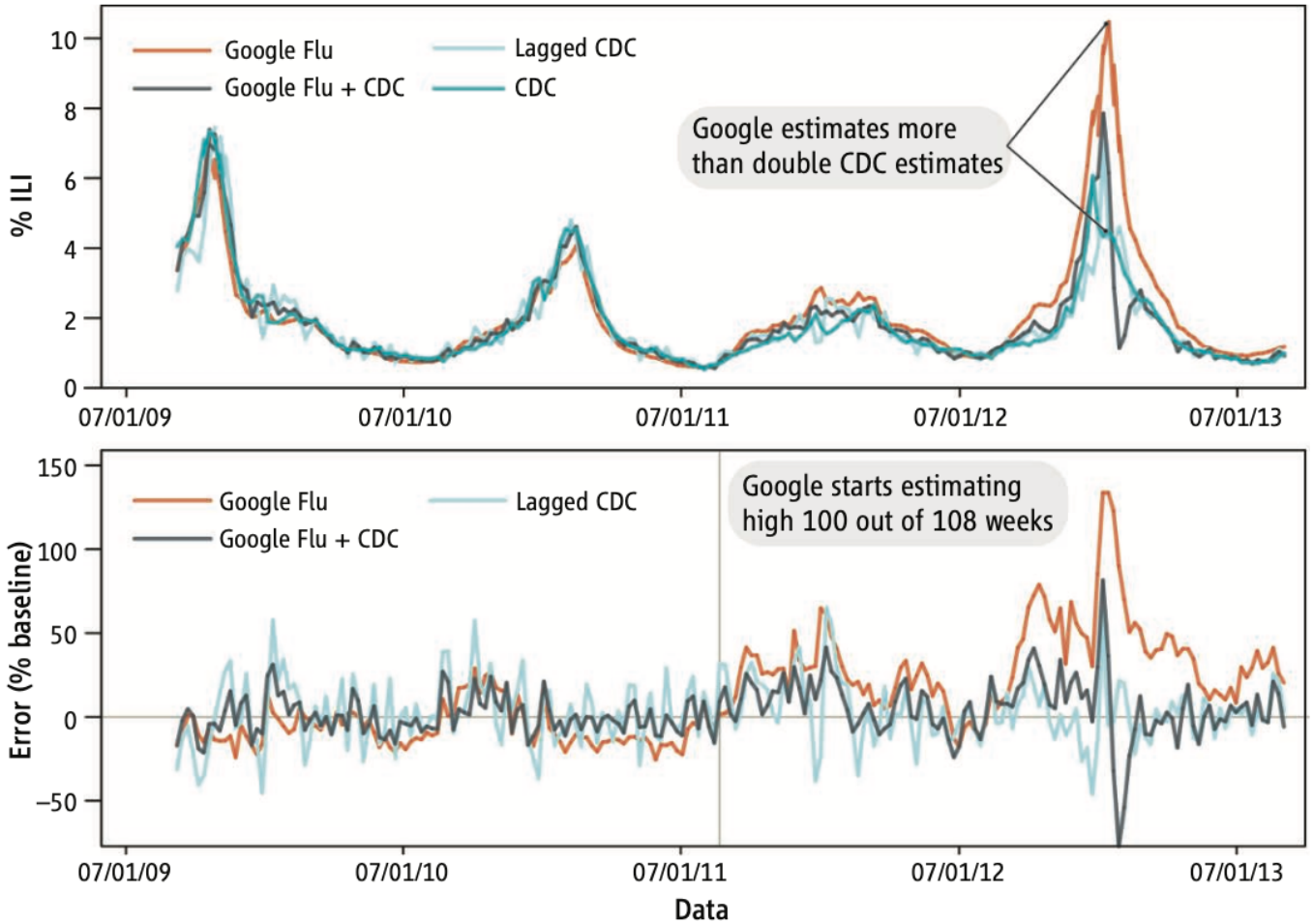


Representativeness concern 4: Algorithms

4. Algorithms are consistently changing, and affect behavior without our knowing
    - Changes in behavior over time that we think are of theoretical interest may simply be due to changes in the algorithms used by social media companies
    - These changes is typically unannounced
    - This can lead to unintended consequences for researchers...

Google Flu Trends

- Use Google Search data to predict trends in flu
- But the algorithm that runs the engine is constantly changed
- One change, for example, recommended search terms to users, thus increasing their frequency...



GFT overestimation. GFT overestimated the prevalence of flu in the 2012–2013 season and overshot the actual level in 2011–2012 by more than 50%. From 21 August 2011 to 1 September 2013, GFT reported overly high flu prevalence 100 out of 108 weeks. (**Top**) Estimates of doctor visits for ILI. "Lagged CDC" incorporates 52-week seasonality variables with lagged CDC data. "Google Flu + CDC" combines GFT, lagged CDC estimates, lagged error of GFT estimates, and 52-week seasonality variables. (**Bottom**) Error [as a percentage {[Non-CDC estmate)−(CDC estimate)]/(CDC) estimate)}. Both alternative models have much less error than GFT alone. Mean absolute error (MAE) during the out-of-sample period is 0.486 for GFT, 0.311 for lagged CDC, and 0.232

YouTube radicalization

- A number of studies have examined whether YouTube's algorithm sends users down a "rabbit hole"
- These data are collected through algorithm "audit studies" that start with a seed channel and follow recommendations
- These data will not be representative, however, without capturing the personalization of the algorithm to actual users
- Further, difficult to answer if minimal effects currently are representative of the algorithm in earlier years

Potential solutions:

- Be aware that algorithmic changes will affect the composition of your data
  - Compositional changes are a general issue independent of algorithms as well
- Be explicit about the drawbacks to data if such changes potentially exist
  - If so, conduct robustness checks of your results on different periods that may be affected

Representativeness concern 5: Platforms

- Are likely large differences in behavior across social media platforms
  - Behavior on Facebook will differ from that on YouTube from that on Reddit from that on Twitter
- Yet the literature focuses heavily on Twitter
- Twitter is not representative of the mechanisms on social media generally
  - Short messages
  - Rapid turnout in user base
  - Quick reaction times
  - Is a directed network (unlike Facebook)
  - Allow more interaction with elite accounts
  - Elite-dominated (media, celebs, politicians)

- Whereas Facebook has:
  - Longer messages
  - Longer reaction time
  - Much slow turnover
  - Is an undirected networked (friending is two-way)
  - Conversations evolve more slowly over time
- Upshot: Recognize that the social media platforms we can more easily study might be substantially different than those we cannot
- How might social media differ across platforms with respect to political behavior?

- Other concerns
  - Longer messages
  - Longer reaction time
  - Much slow turnover
  - Is an undirect networked (friending is two-way)
  - Conversations evolve more slowly over time
- Upshot: Recognize that the social media platforms we can more easily study might be substantially different than those we cannot

Summary

- Unrepresentative of the general public
- APIs don't necessarily generate representative data
- Statistics about social media interactions can be deceiving
- Algorithms affect behavior in unknown ways
- Are differences across platforms, but Twitter is over-studied

Next class

- Data collection
- Basic text cleaning & analysis