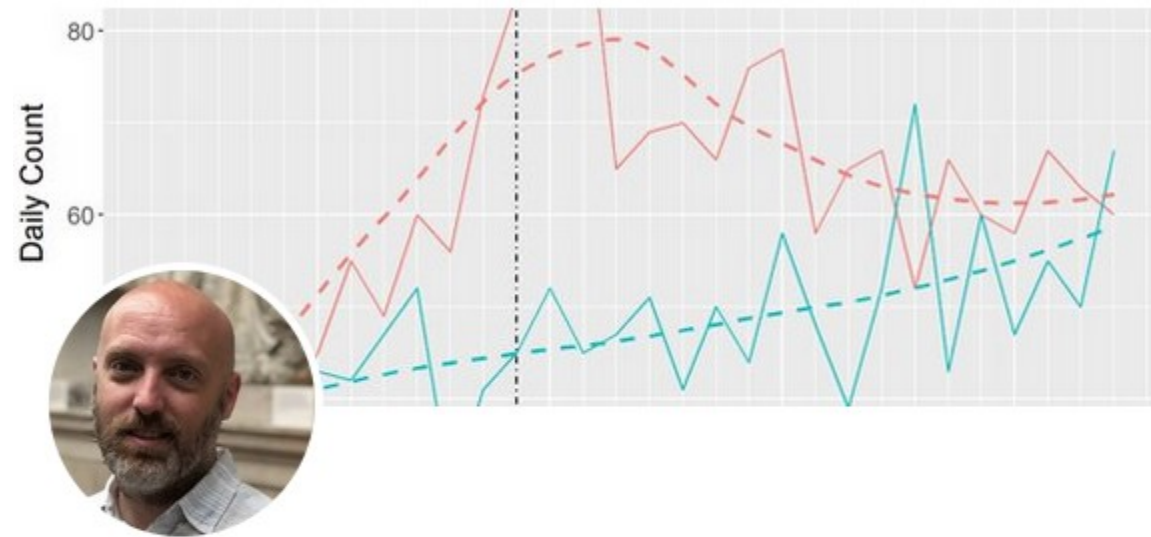


Studying Personally Expressed Political Identity at Scale

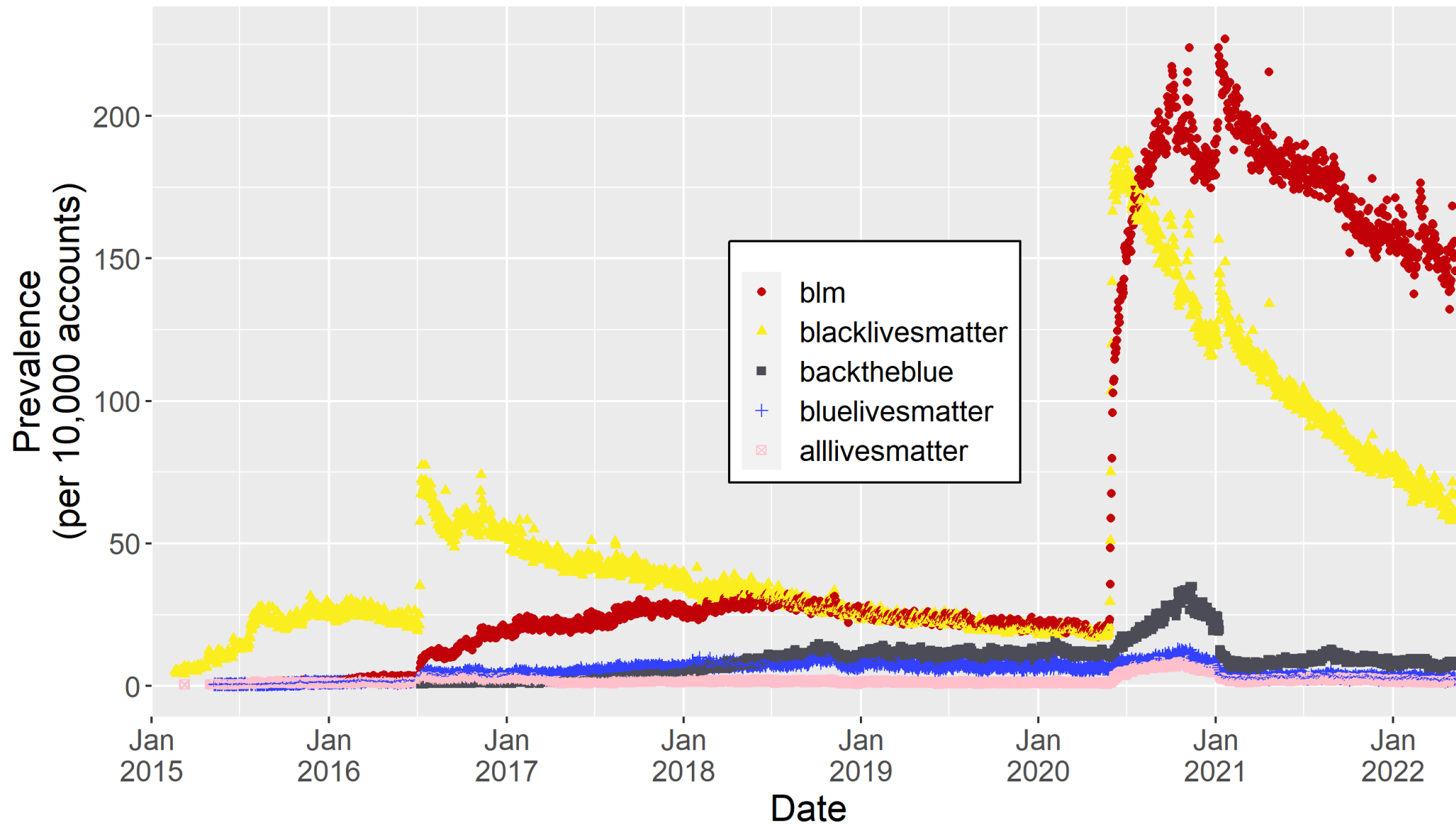


Jason Jeffrey Jones

@jasonjones_jjj

Computational Social Scientist. I mine datasets to see what they have to say about the human condition.

© Stony Brook, NY [jasonjones.ninja](https://www.jasonjones.ninja)



N ~ 200,000 unique US Twitter accounts per day.
Source: <https://jasonjones.ninja/ipseology-central/blog/black-lives-matter-daily-prevalence-2022.html>
© Jason Jeffrey Jones
You may share and adapt this work under terms of the CC BY 4.0 License.



Tweets **15.5K** Following **621K** Followers **104M** Likes **11** Lists **3**

[Follow](#)

Barack Obama ✓

@BarackObama

Dad, husband, President, citizen.

Washington, DC

obama.org

Joined March 2007

Born on August 4, 1961

[Tweet to Barack Obama](#)

Tweets **Tweets & replies** **Media**

Barack Obama Retweeted



Michelle Obama ✓ @MichelleObama · Jun 18

Sometimes truth transcends party.

Laura Bush ✓ @laurawbush

I live in a border state. I appreciate the need to enforce and protect our international boundaries, but this zero-tolerance policy is cruel. It is immoral. And it breaks my heart.


13K 129K 468K



Tweets **40.4K** Following **45** Followers **57.9M** Likes **7** Moments **6**

Donald J. Trump

@realDonaldTrump

45th President of the United States of America 

 Washington, DC

 [Instagram.com/realDonaldTrump](https://www.instagram.com/realDonaldTrump)

 Joined March 2009

[Tweet to Donald J. Trump](#)

Tweets **Tweets & replies** Media



Donald J. Trump  @realDonaldTrump · 7h 

....a source of potential danger and conflict. They are testing Rockets (last week) and more, and are coming very close to the edge. There economy is now crashing, which is the only thing holding them back. Be careful of Iran. Perhaps Intelligence should go back to school!

 32K  14K  56K 



Donald J. Trump  @realDonaldTrump · 7h 

The Intelligence people seem to be extremely passive and naive when it comes to the dangers of Iran. They are wrong! When I became President Iran was making

Personally Expressed Identity



Barack Obama ✓

@BarackObama

Dad, husband, President, citizen.

Bio

📍 Washington, DC

🔗 obama.org

📅 Joined March 2007

🎂 Born on August 4, 1961

- Twitter bios are the best-ever source of data for *personally expressed identity*.
 - Personal
 - The individual is describing themselves.
 - Expressed
 - The individual emits words where others might see them.
 - Identity
 - The purpose of the text is description of the author.
- *Describe yourself in 160 characters or less.*



Malala ✓

@Malala

Advocate for girls' education & women's equality | UN Messenger of Peace | Nobel laureate 2014 | Founder @MalalaFund

malala.org Born July 12 Joined November 2

536 Following 1.9M Followers



Sean Hannity ✓

@seanhannity

TV Host Fox News Channel 9 PM EST. Nationally Syndicated Radio Host 3-6 PM EST. Hannity.com Retweets, Follows NOT endorsements! Due to hackings, no DM's!

New York, USA hannity.com Born December 30 Joined May 2009

6 Following 5.4M Followers



olivia ✓

@oliviamunn

Proud Asian American she/her Typing...

Los Angeles gofundme.com/esea

1,157 Following 842.5K Followers



Al Yankovic ✓

@alyankovic

You know... the weird one.



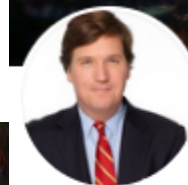
Tom Hanks ✓

@tomhanks

I'm that actor in some of the movies you liked and some you didn't. Sometimes I'm in pretty good shape, other times I'm not. Hey, you gotta live, you know?

Los Angeles playtone.com Joined June 2009

25 Following 16.4M Followers



Tucker Carlson ✓

@TuckerCarlson

Host of "Tucker Carlson Tonight", weeknights at 8 PM ET @FoxNews. My new book is out now at TuckerCarlson.com Re-tweets are emphatic endorsements.

Washington, DC TuckerCarlson.com Joined March 2009

96 Following 4.6M Followers



christina applegate ✓

@1capplegate

homeschool mom. failing mom. she will be ok .carry the damn one. wtf with new math??!!!!!!!

here twitter.com/1capplegate Joined May 2009

321 Following 1.4M Followers



Tom Brady ✓

@TomBrady

Family and Football



Greta Thunberg ✓

@GretaThunberg

Climate- and environmental activist with Asperger's Born at 375 ppm

Sverige ClimateEmergencyEU.org Joined June 2018

2,619 Following 5M Followers



Venus Williams ✓

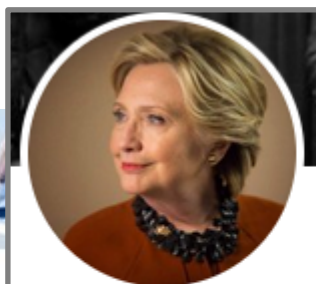
@Venuseswilliams

Tennis player, big sister, grown up girl. Double Tap! ❤️ Be Well ❤️ #CoachVenus

[@elevenbyvenus](https://www.elevenbyvenus.com) workouts @ link in bio

Palm Beach Gardens, FL linktr.ee/VenusWilliams Joined June 2009

257 Following 1.7M Followers



Hillary Clinton ✓

@HillaryClinton

2016 Democratic Nominee, SecState, Senator, hair icon. Mom, Wife, Grandma x3, lawyer, advocate, fan of walks in the woods & standing up for our democracy.



Danny DeVito ✓

@DannyDeVito

I'm an actor, director and producer.

Los Angeles [imdb.com/name/nm000](https://www.imdb.com/name/nm000)

21 Following 4.2M Followers



Kamala Harris ✓

@KamalaHarris

United States government official

Fighting for the people. Wife, Momala, Auntie. She/her. Official account is @VP

Washington, DC [joebiden.com](https://www.joebiden.com) Joined April 2009

734 Following 18.7M Followers

Who am I? Measures

- Traditional state-of-the-art measurement of self-described identity comprises ***Who-am-I instruments***.
 - Verbal, pencil-and-paper instruments (survey-like).
 - So-called because the latent question to be answered is Who am I?
 - There are a lot of them (too many).
- Let's focus on one: the ***Twenty Statements Test***



Twenty Statements Test

Please write twenty answers to the simple question 'Who am I?'.
Just give twenty different answers to this question.

Answer as if you were giving the answers to yourself, not to somebody else.

Write the answers in the order that they occur to you. Don't worry about logic or importance.

Go along fairly fast, for time is limited.

LOPS: A Method to Measure *Change* in Personally Expressed Identity

- **Longitudinal Online Profile Sampling**
- *Longitudinal*
 - Follow the same people over time (years)
- *Online*
 - For ease of collection and scale
- *Profile*
 - Wherein one describes oneself
- *Sampling*
 - Daily, monthly, annual checks for changes
- Rogers, N., & Jones, J. J. (2021). *Using twitter bios to measure changes in self-identity: Are Americans defining themselves more politically over time?* Journal of Social Computing, 2(1), 1-13. [Manuscript PDF](#)



LOPS: A Method to Measure *Change* in Personally Expressed Identity

1. Define a population.
2. Collect a sample of bios over time.
3. Tokenize bios on `\b` word boundaries.
 - Tokens are elements of identity.
4. Compute user ***prevalence*** for tokens of interest.

LOPS: A Method to Measure *Change* in Personally Expressed Identity

1. Define a population.
 - US users of Twitter
 - Filtered from the Twitter Streaming API 1% random sample of all tweets.
2. Collect a sample of bios over time.
3. Tokenize bios on `\b` word boundaries.
 - Tokens are elements of identity.
4. Compute user *prevalence* for tokens of interest.

US Users of Twitter

Daily: about 200,000 unique users

Annually: about 10 million unique users

Based on 'location':
e.g. *USA, , OH, Texas*

LOPS: A Method to Measure *Change* in Personally Expressed Identity

1. Define a population.
2. Collect a sample of bios over time.
 - Median days between profile edits is 150 days.
 - 2015 - Present
3. Tokenize bios on `\b` word boundaries.
 - Tokens are elements of identity.
4. Compute user *prevalence* for tokens of interest.



LOPS: A Method to Measure *Change* in Personally Expressed Identity

1. Define a population.
2. Collect a sample of bios over time.
3. Tokenize bios on `\b` word boundaries.
 - Tokens are elements of identity.
4. Compute user *prevalence* for tokens of interest.



↓
Tokenize

dad
husband
president
citizen

LOPS: A Method to Measure *Change* in Personally Expressed Identity

1. Define a population.
2. Collect a sample of bios over time.
3. Tokenize bios on \b word boundaries.
 - Tokens are elements of identity.
4. Compute user ***prevalence*** for tokens of interest.

$$\textit{Prevalence} = 10,000 * \textit{Count of Users with Token} / \textit{Total User Count}$$

Prevalence is incidence divided by population size (times 10,000).

Incidence is raw tally of users who use the token.

The ratio of users with the token to total users is multiplied by 10,000 because, of course, most bios do not contain most tokens, and it is much easier for humans to think in whole numbers than small fractions or decimals.

It is important to be clear that reported counts are always counts of users and never words. A bio that reads "token token token token" counts as one user whose bio contains "token," even though the bio text happens to contain "token" four times.

LOPS: A Method to Measure *Change* in Personally Expressed Identity

- One more note: It is usefully to study prevalence in both ***cross-sectional*** and ***longitudinal*** samples.
- Annual US User token datasets:
 - ***Cross-sectional*** sample: ~ 10 million unique US users observed in a given year
 - ***Longitudinal*** sample: ~ 1.5 million unique US users observed *each and every year*
 - Data: <https://osf.io/guah5/>

Cross-sectional

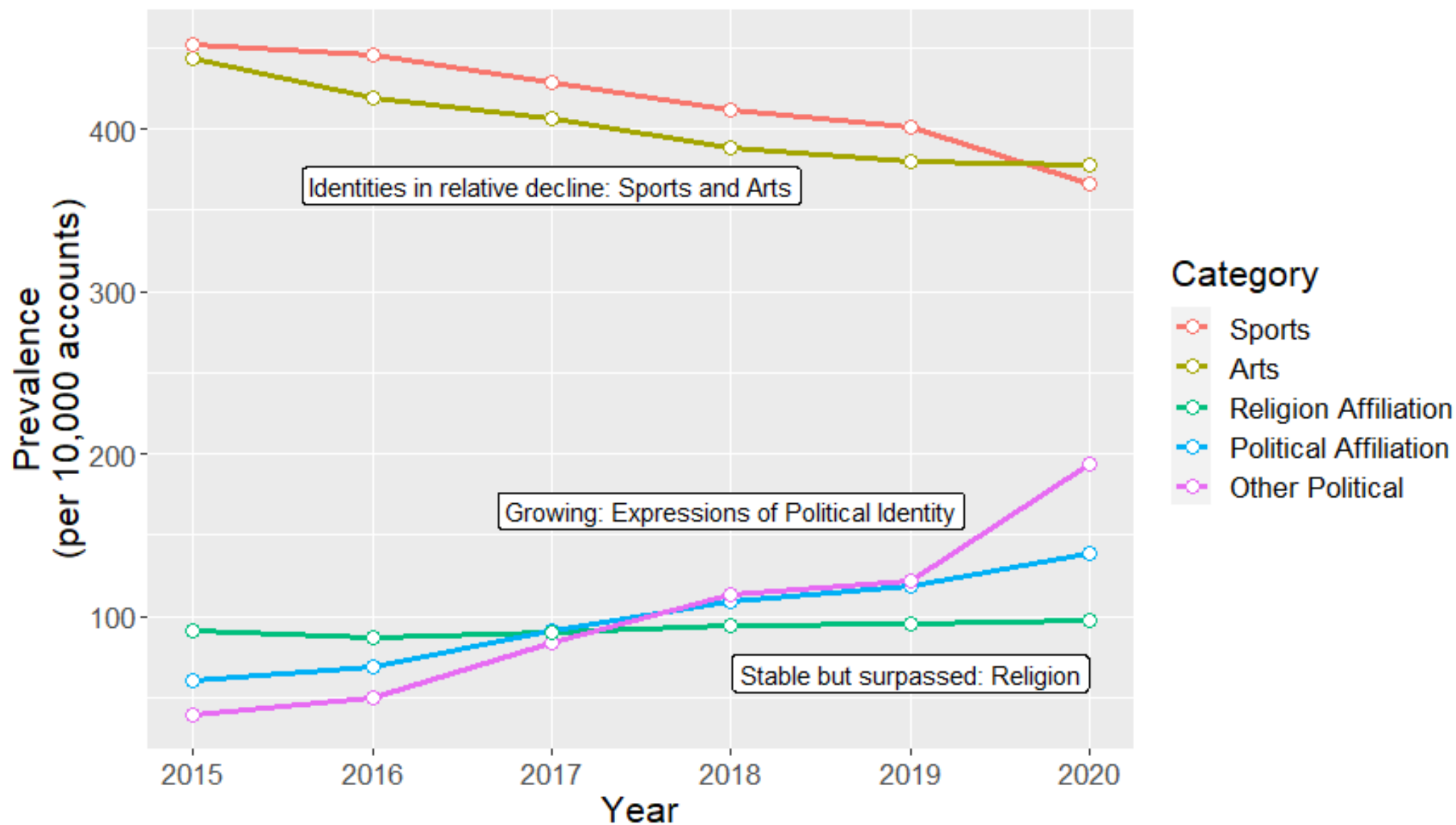
- Activity
- Visibility
- Population-level change

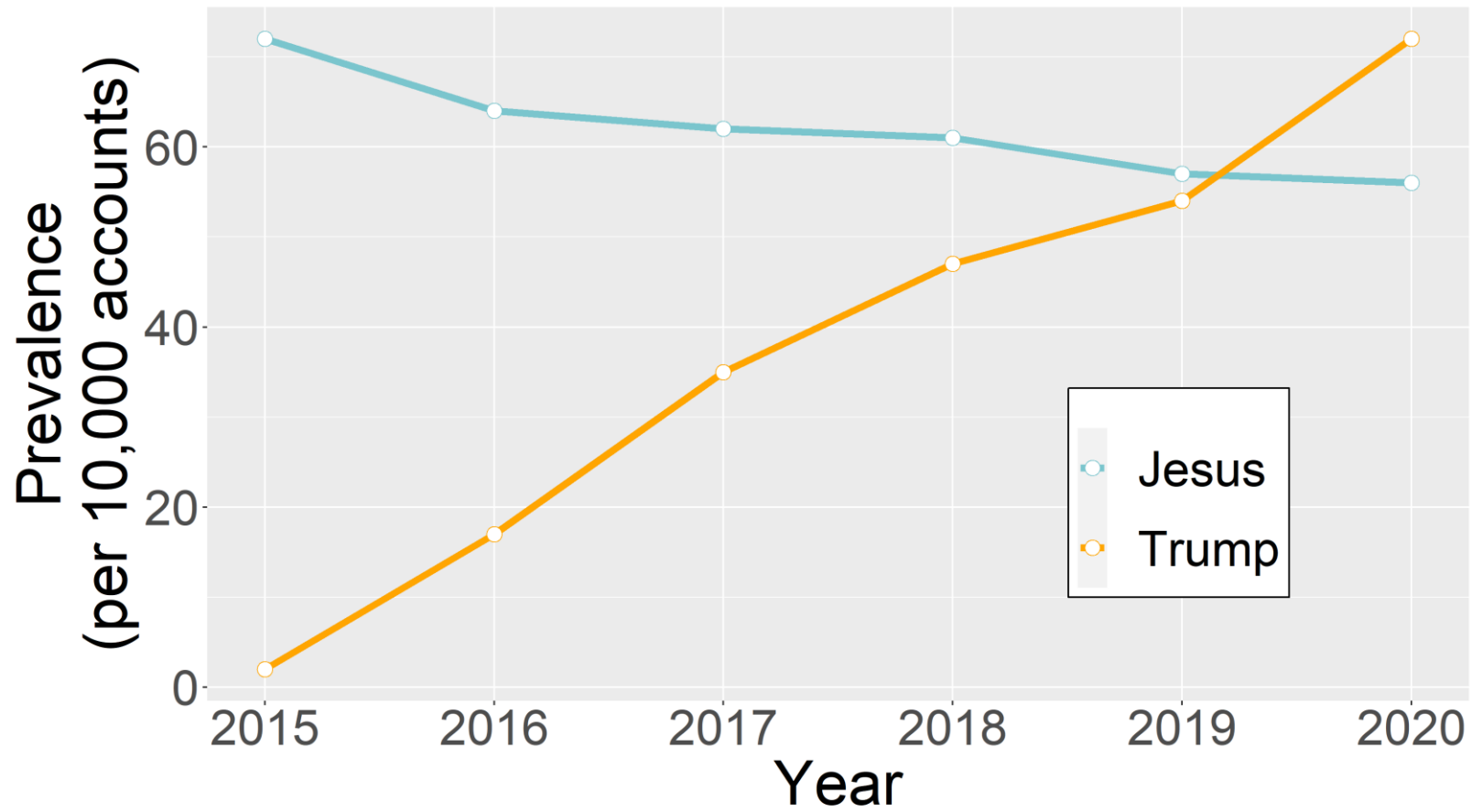
Longitudinal

- Adds
- Deletes
- Transitions
- Individual-level change


Prevalence of Users whose Bio Contains Category Identifier

Cross-sectional sample of American users of Twitter in each year 2015-2020






Note: Cross-sectional sample

 **The Washington Post**
Democracy Dies in Darkness

Monkey Cage

Analysis

On Twitter, Trump is more popular than Jesus

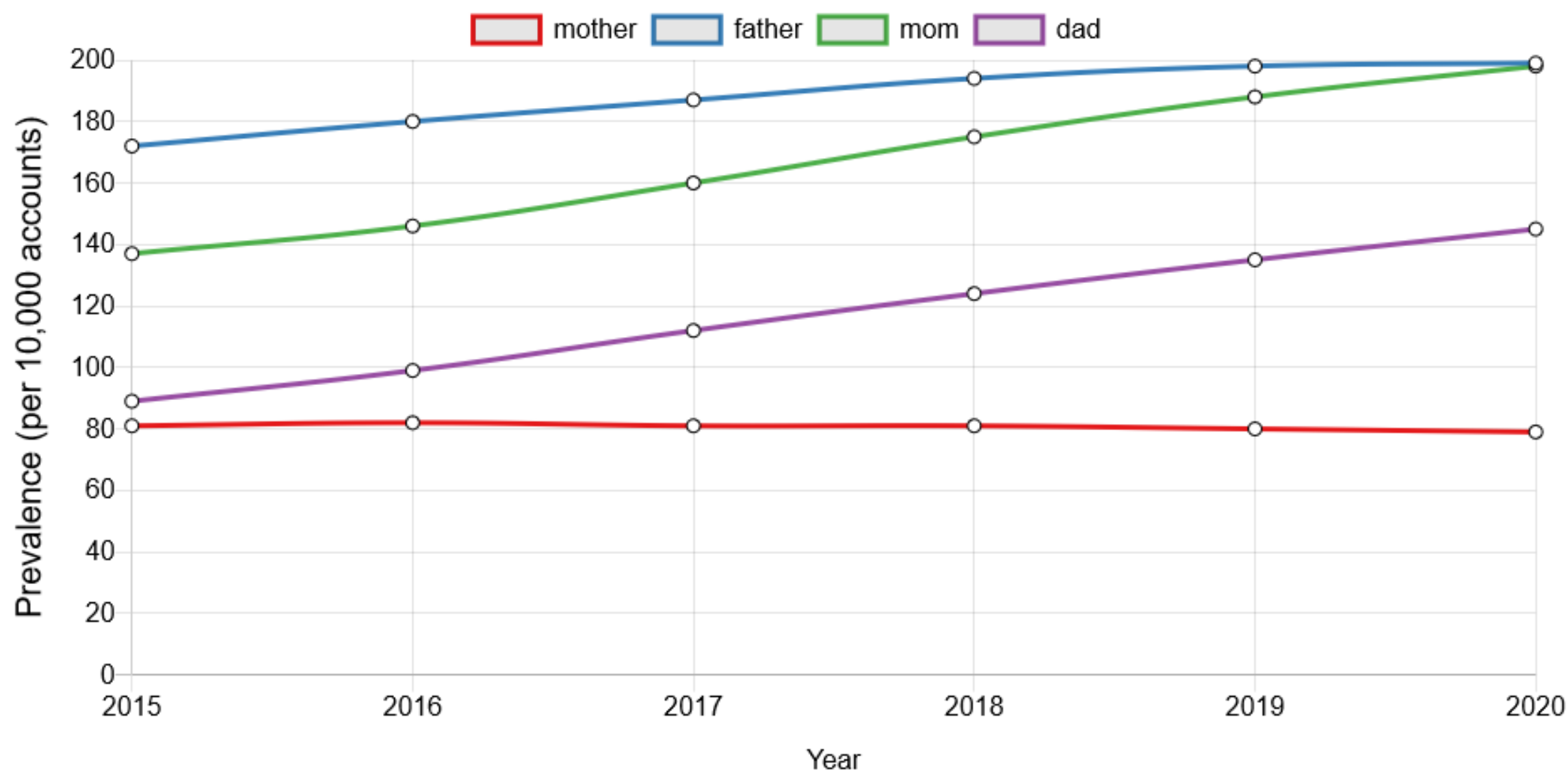


Twitter users' bios suggest they now identify more with politics than religion.

Nick Rogers and Jason J. Jones · 1 hour ago

Query: mother, father, mom, dad

Sample: Longitudinal sample



Identity Trends

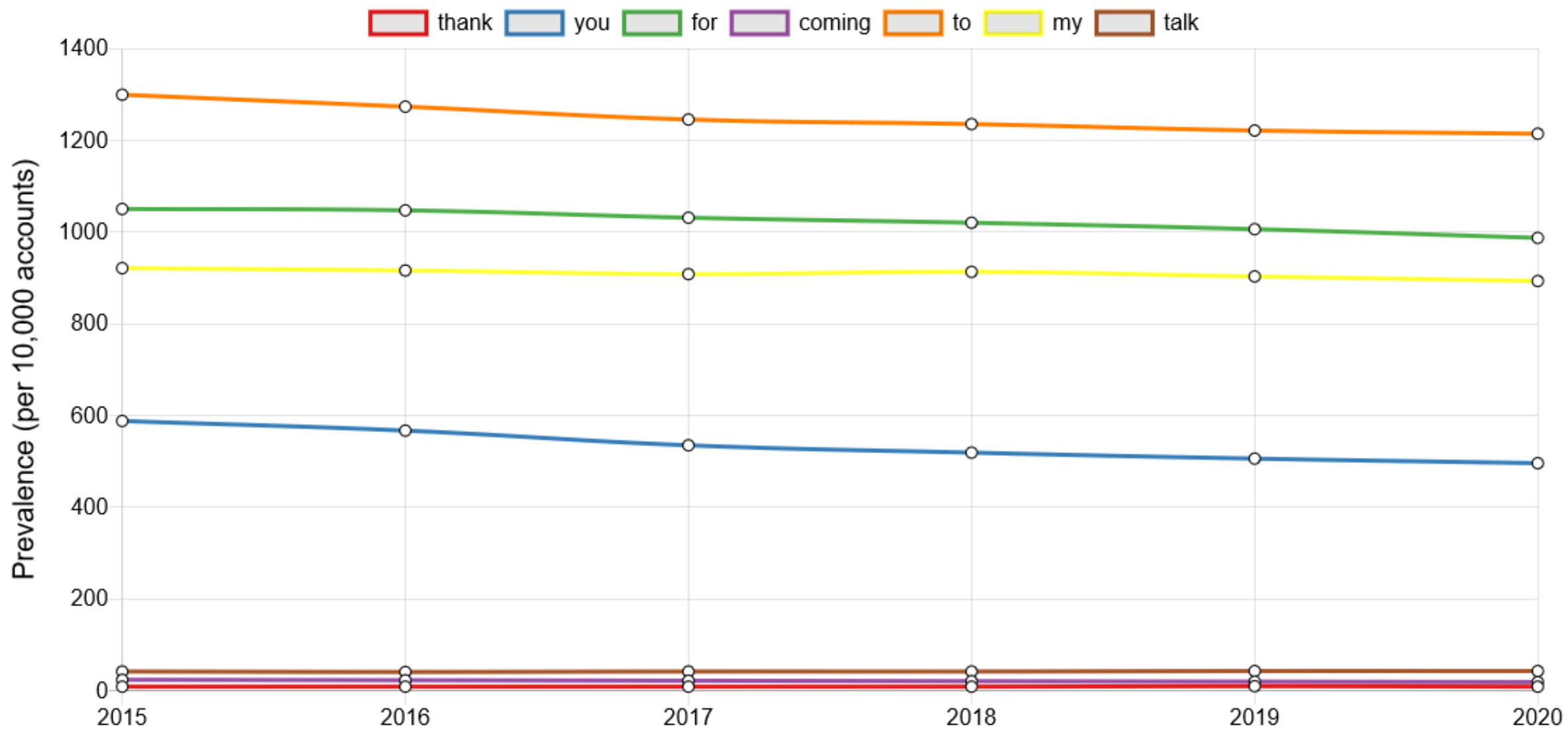
- Explore how Americans are describing themselves online.
- Search for a word, and you will see the prevalence of American Twitter users who chose to include that word in their profile "bio."
 - Trend from 2015-2020.
- <https://jasonjones.ninja/jason-j-jones-identity-trends-v1/>

Data and Methods

- Jones, Jason Jeffrey. *A dataset for the study of identity at scale: Annual Prevalence of American Twitter Users with specified Token in their Profile Bio 2015–2020*. *PloS One* 16.11 (2021): e0260185.
<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0260185>

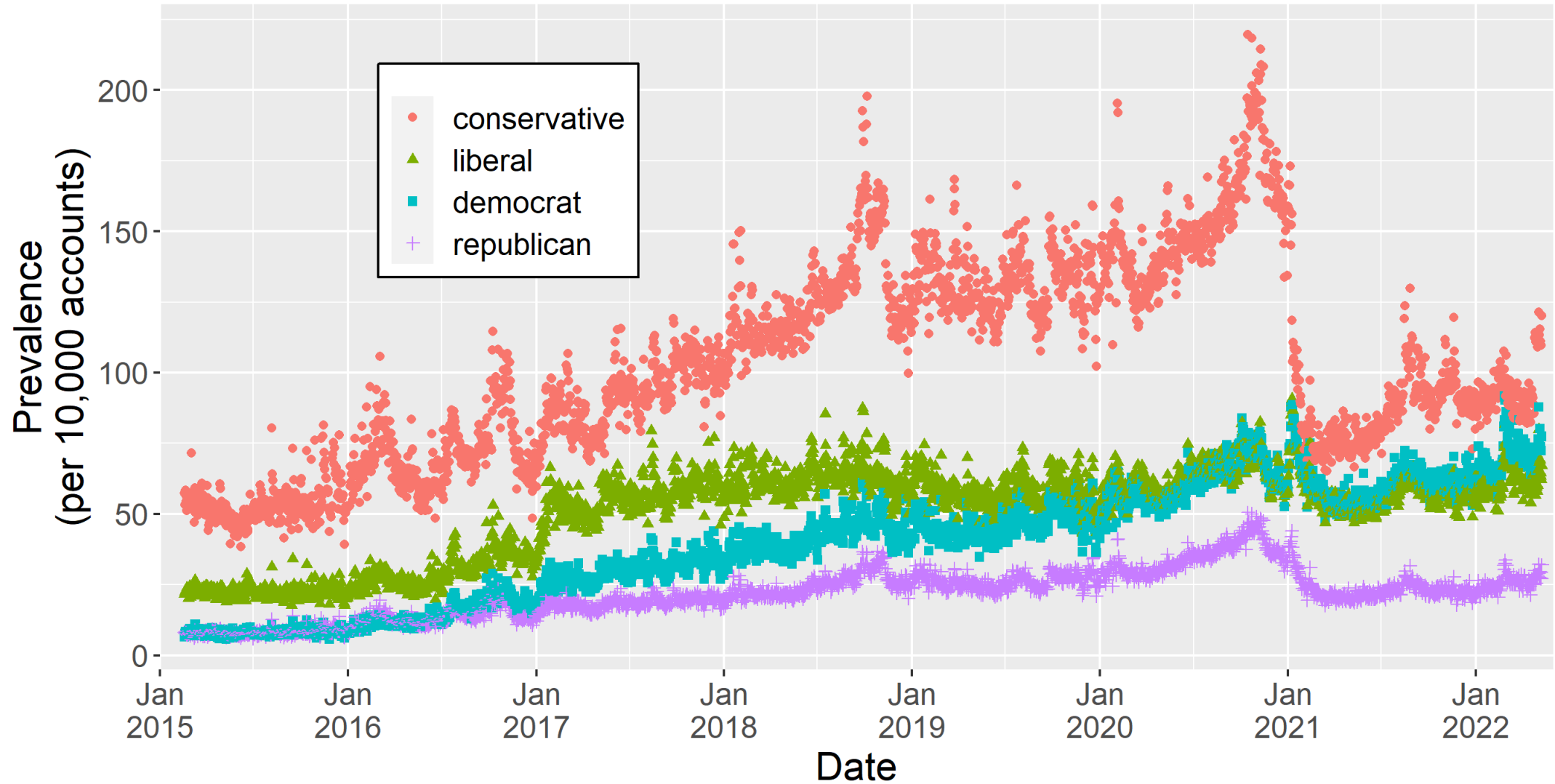
Summary

- We construct our personal identity on media platforms.
- With ***Longitudinal Online Profile Sampling***, one can observe shifting trends in how people describe themselves in near-real-time.
- US Twitter users increasingly use political words to describe themselves.



American Users of Twitter with Political Affiliation Signifier in the Bio

Cross-sectional, daily resolution

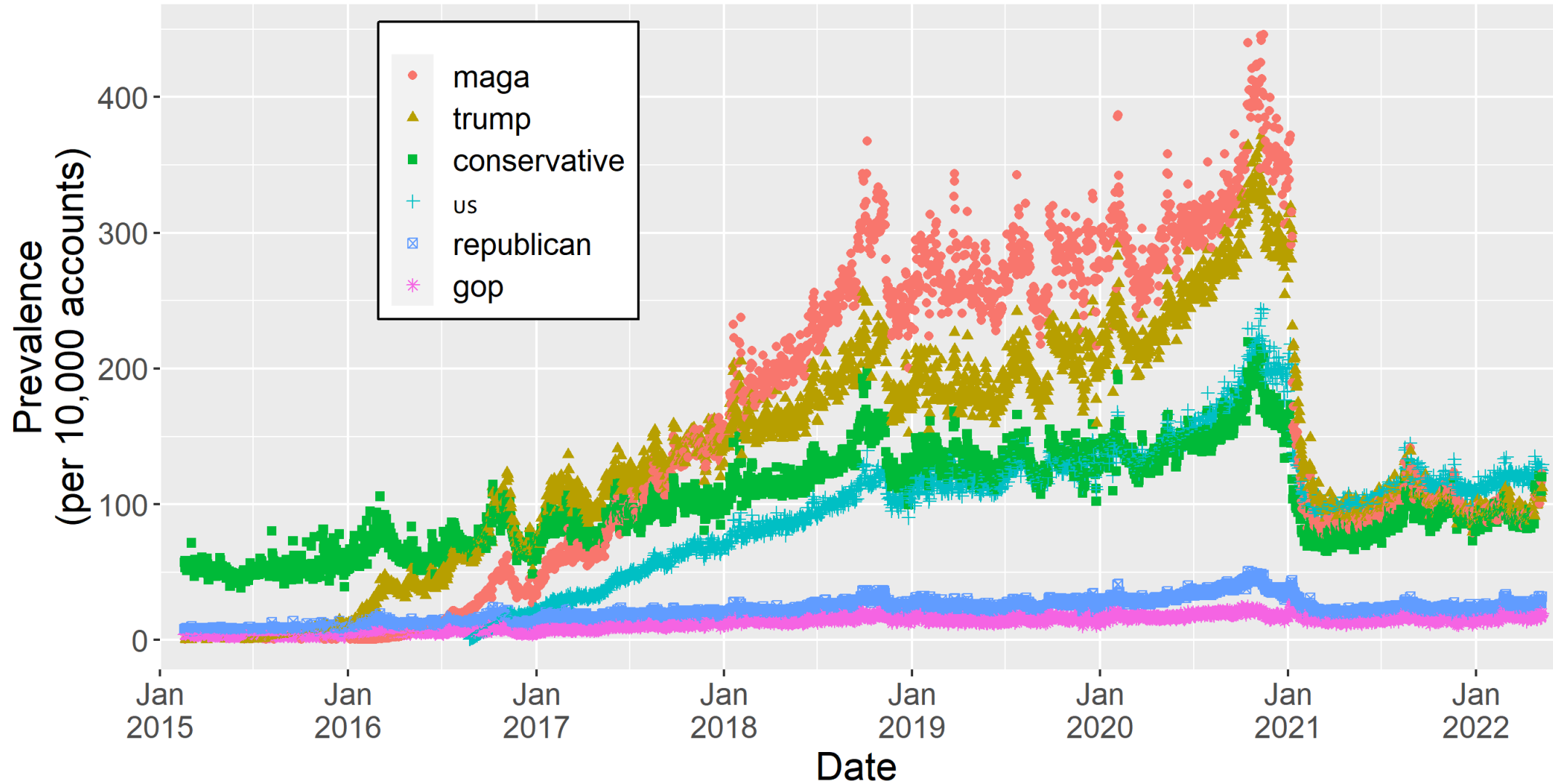


N ~ 200,000 unique US Twitter accounts per day.

© Jason Jeffrey Jones

You may share and adapt this work under terms of the CC BY 4.0 License.

American Users of Twitter with Trump/GOP Affiliation Signifier in the Cross-sectional, daily resolution



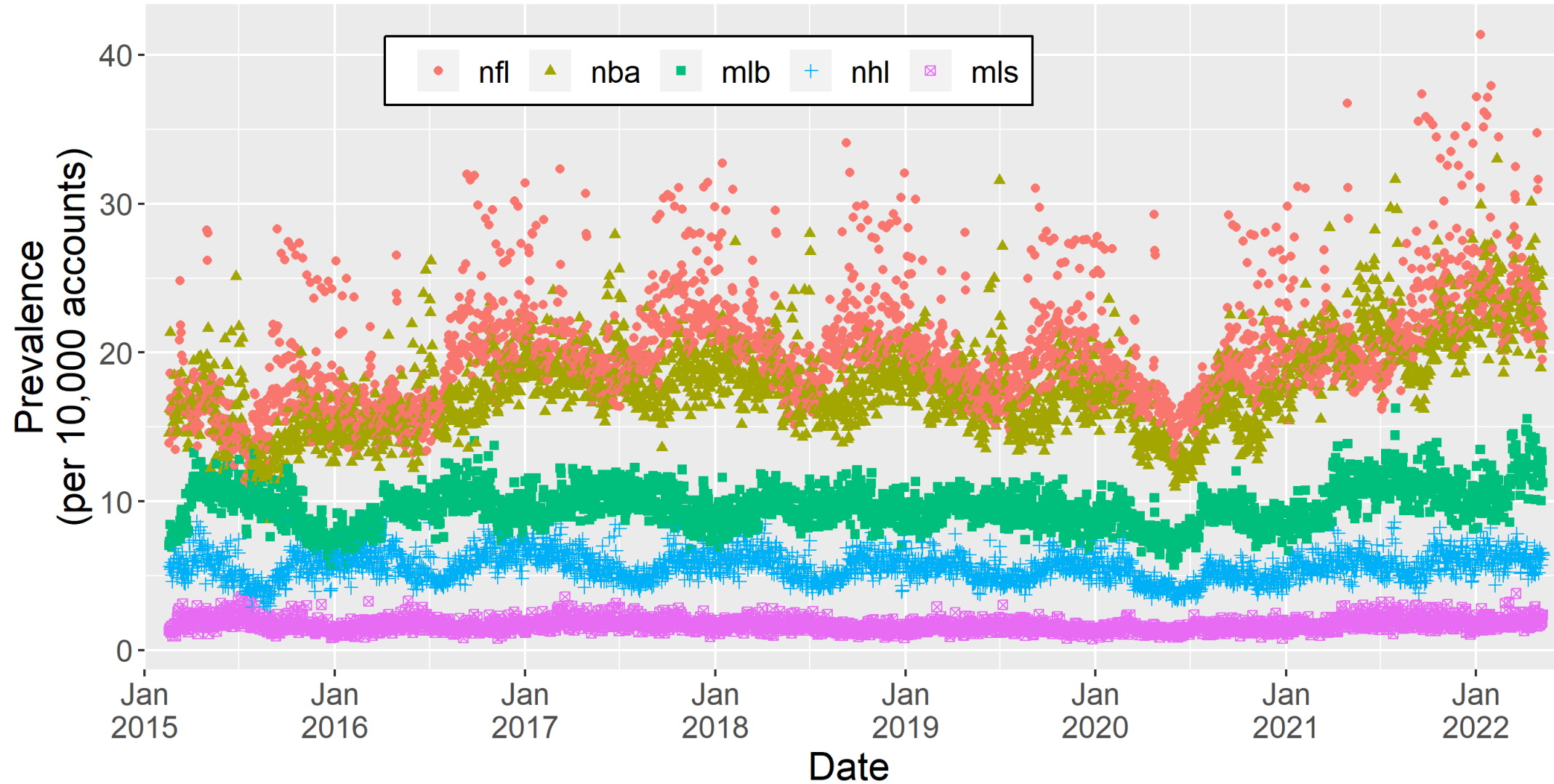
N ~ 200,000 unique US Twitter accounts per day.

© Jason Jeffrey Jones

You may share and adapt this work under terms of the CC BY 4.0 License.

American Users of Twitter with Sport League Signifier in the Bio

Cross-sectional, daily resolution



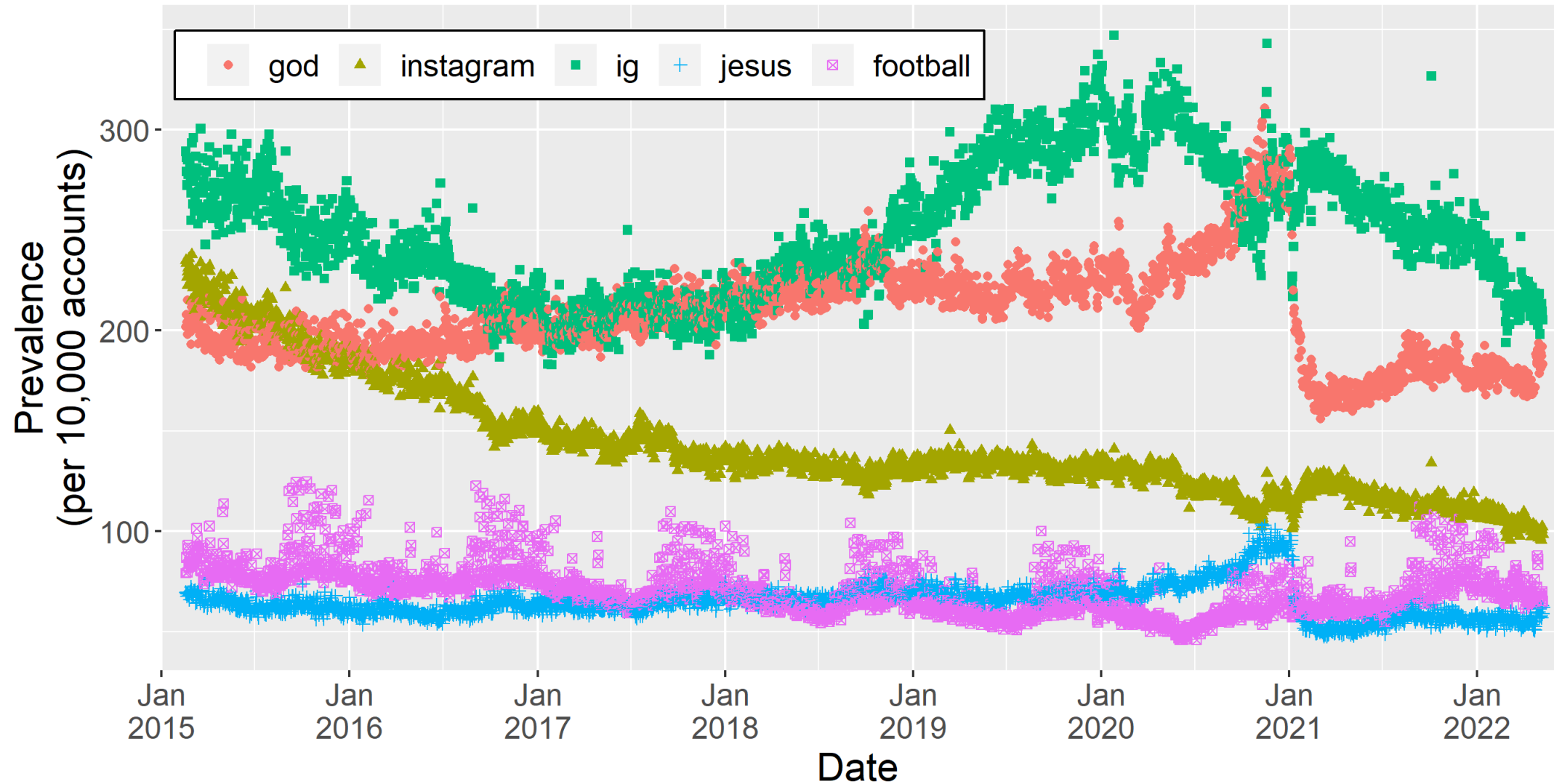
N ~ 200,000 unique US Twitter accounts per day.

© Jason Jeffrey Jones

You may share and adapt this work under terms of the CC BY 4.0 License.

American Users of Twitter with Signifier in the Bio

Cross-sectional, daily resolution



N ~ 200,000 unique US Twitter accounts per day.

© Jason Jeffrey Jones

You may share and adapt this work under terms of the CC BY 4.0 License.

Personally Expressed Identity



Barack Obama ✓

@BarackObama

Dad, husband, President, citizen.

Bio

📍 Washington, DC

🔗 obama.org

📅 Joined March 2007

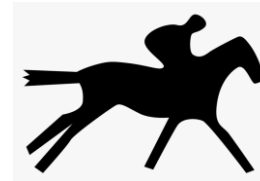
🎂 Born on August 4, 1961

- Twitter bios are the best-ever source of data for *personally expressed identity*.
- *Describe yourself in 160 characters or less.*
- A **computational social science** approach:
 - **Scales** – can request millions of data points each day.
 - Is **longitudinal** – profiles are persistent.
 - Has a data-driven approach to “coding.”
 - Just count users per token

Identity Threshold Theory



Hitman



Actor



Student

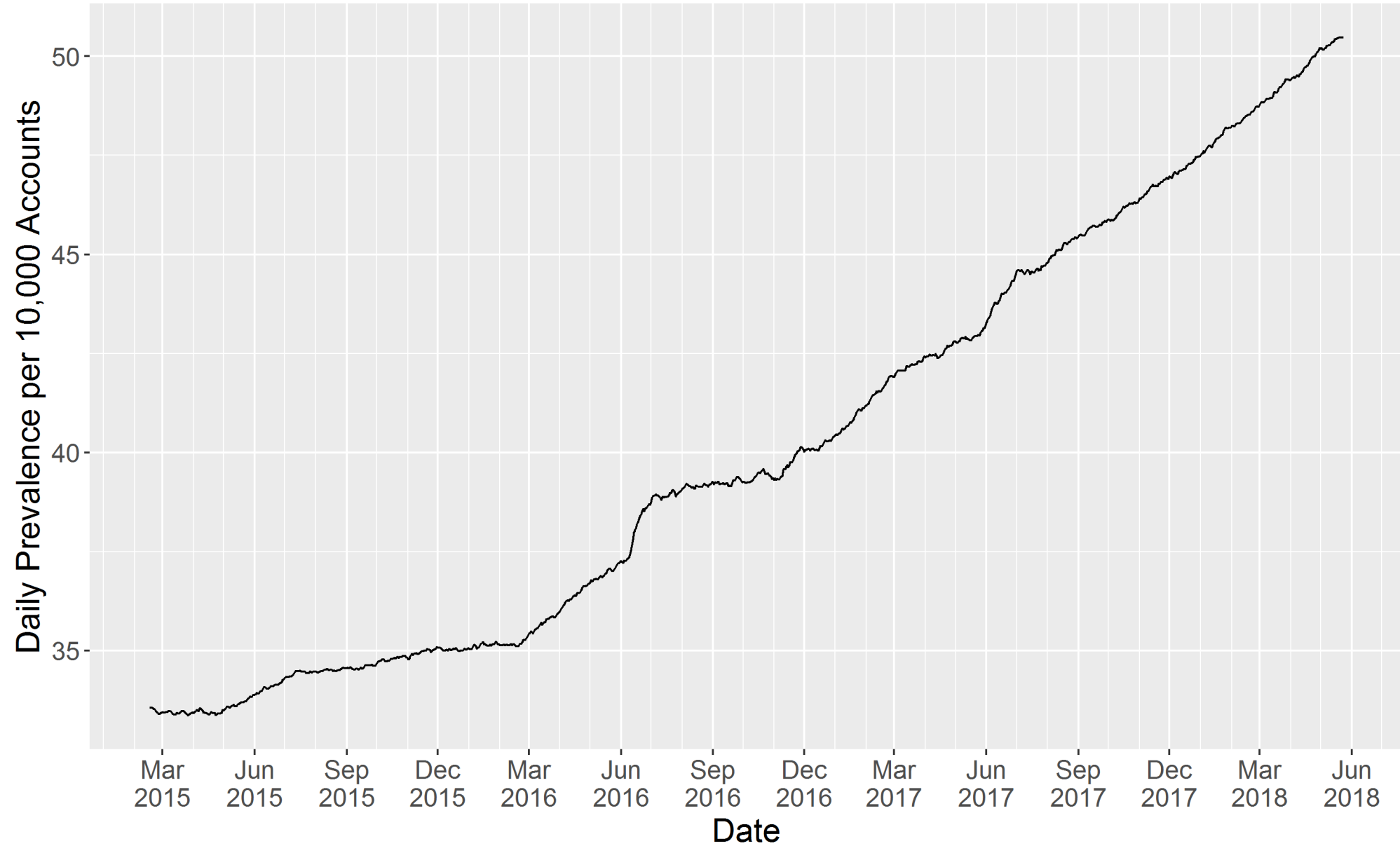


Boyfriend

Identity Threshold

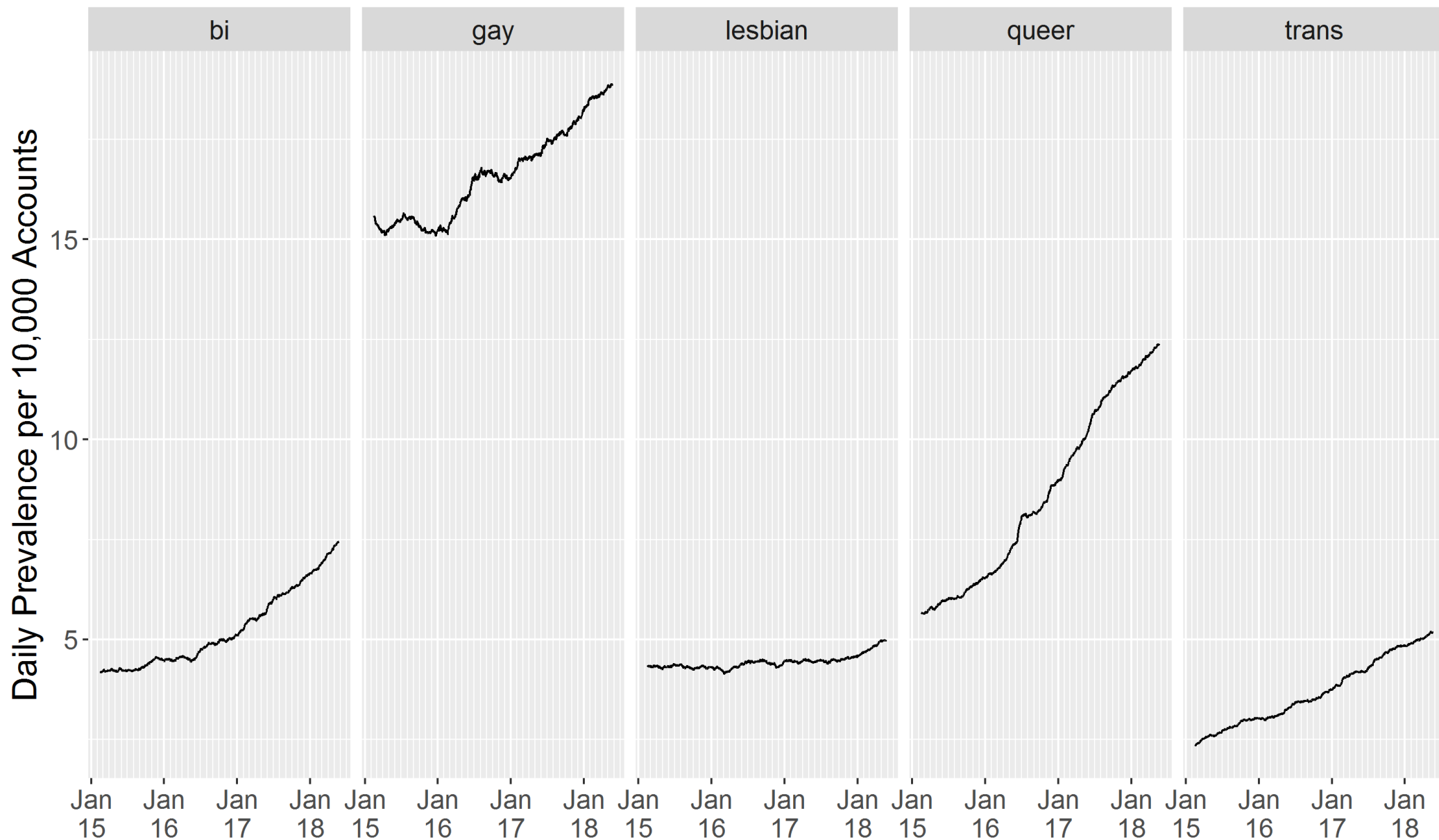


Prevalence of Accounts with Bios Containing at least one LGBTQ Keyword



Note: Longitudinal sample

Prevalence of Accounts with Bios Containing a Specific LGBTQ Keyword

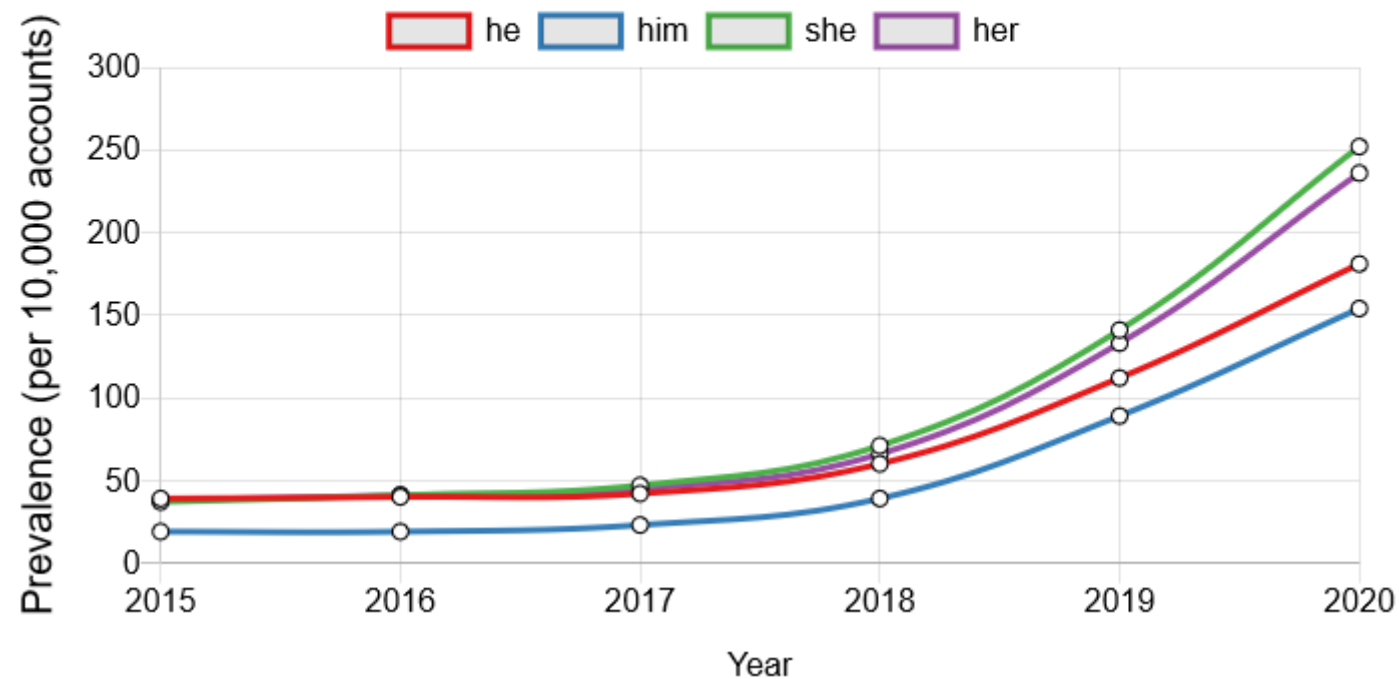


Note: Longitudinal sample

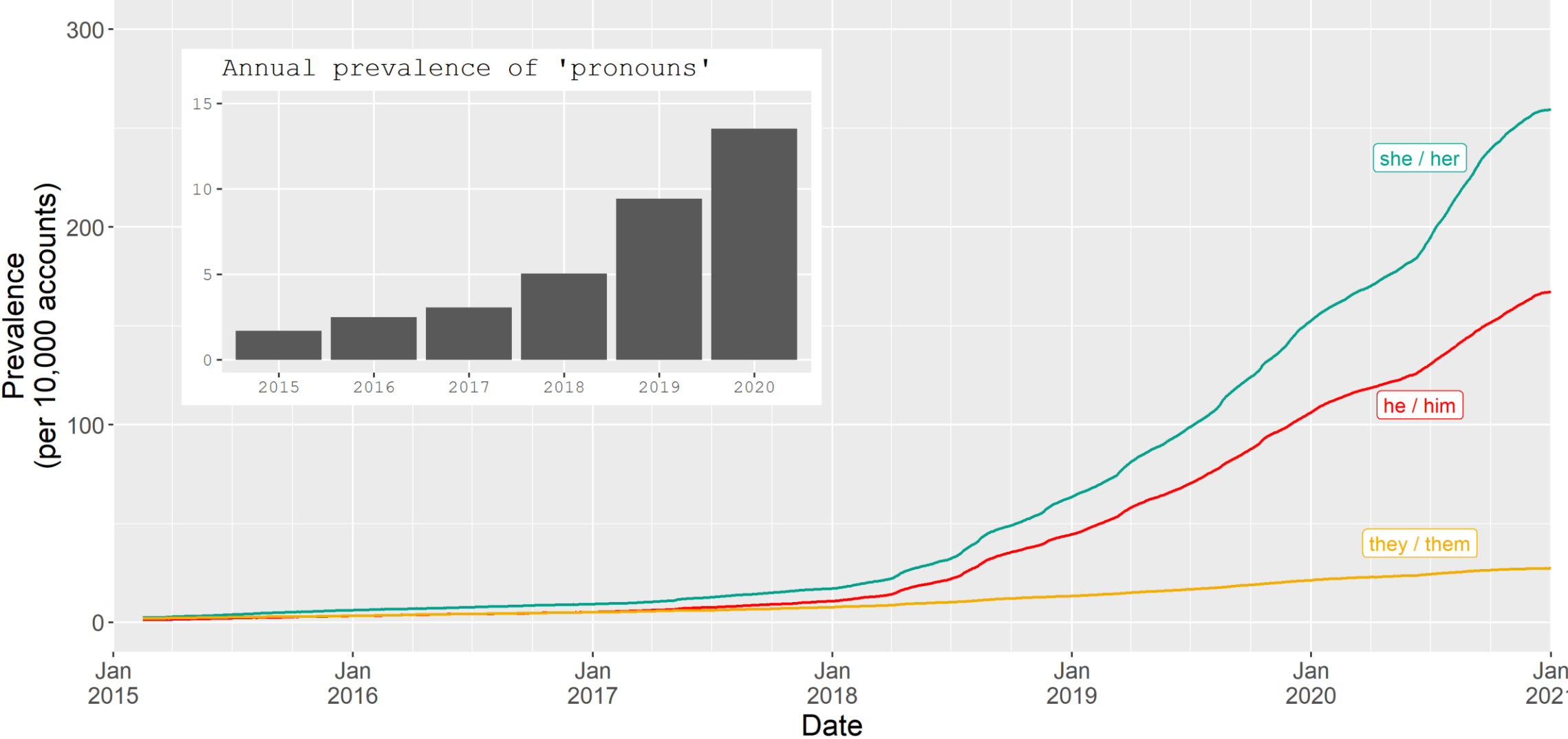
Pronoun-slash-lists in the Bio: A descriptive, exploratory analysis

Query: he, him, she, her

Sample: Longitudinal sample



Daily prevalence of bios containing pronoun-slash-lists

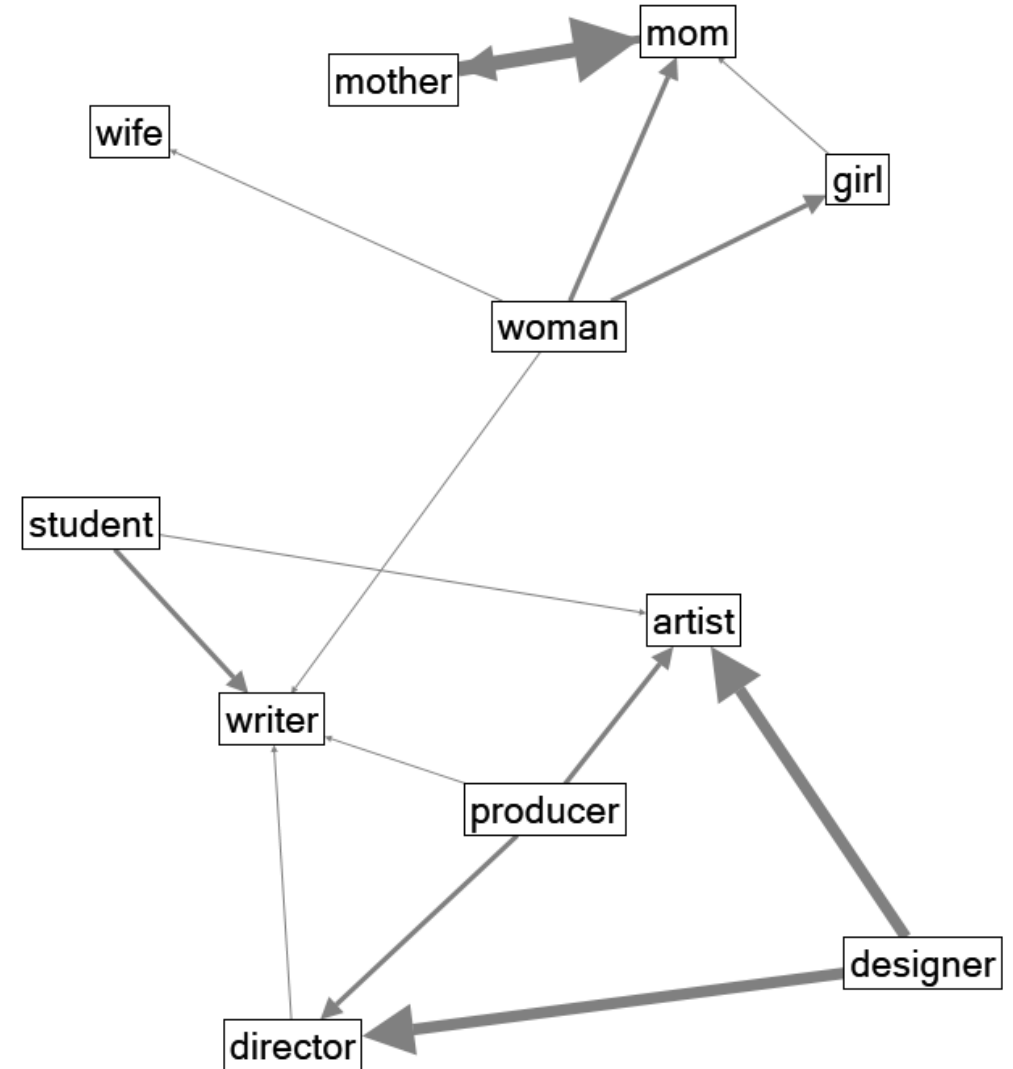
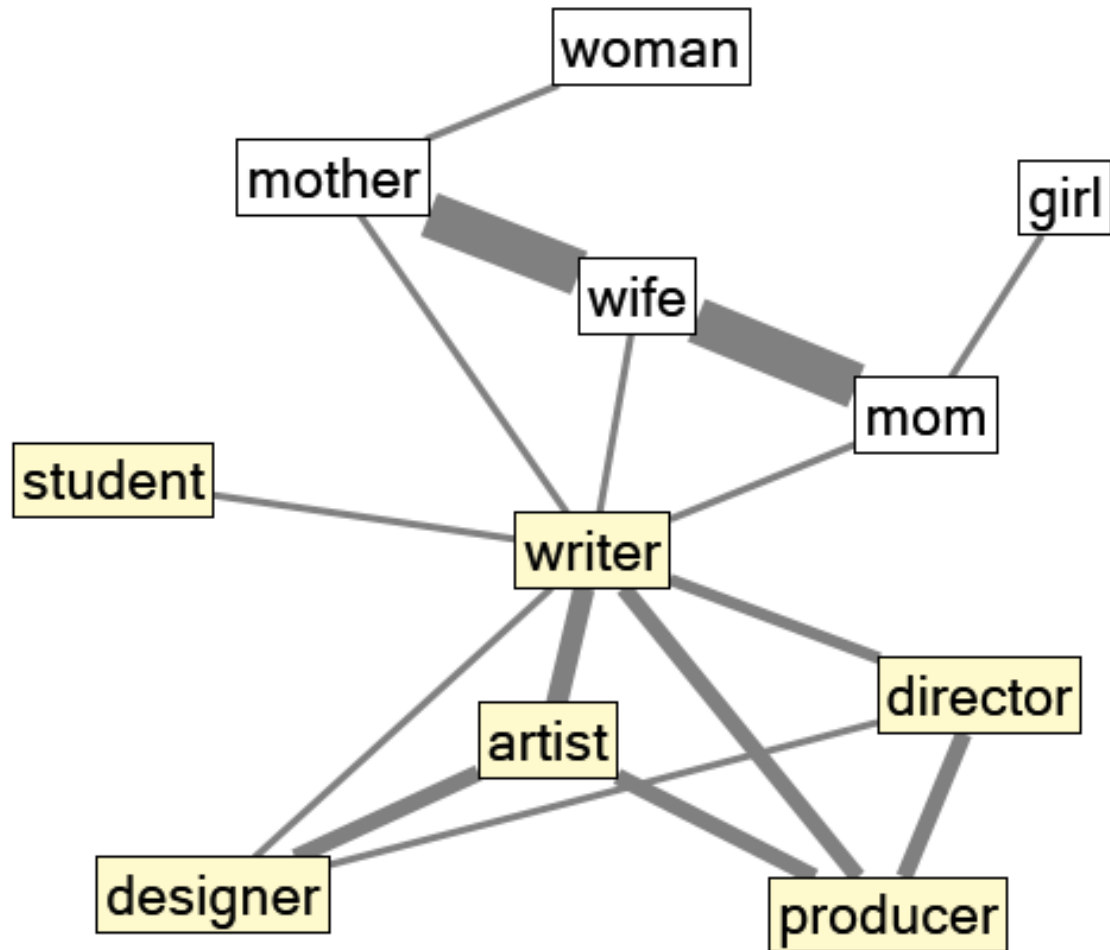


N = 1.35 million US Twitter accounts.
Source: <https://osf.io/4h8jk/>
© Jason Jeffrey Jones

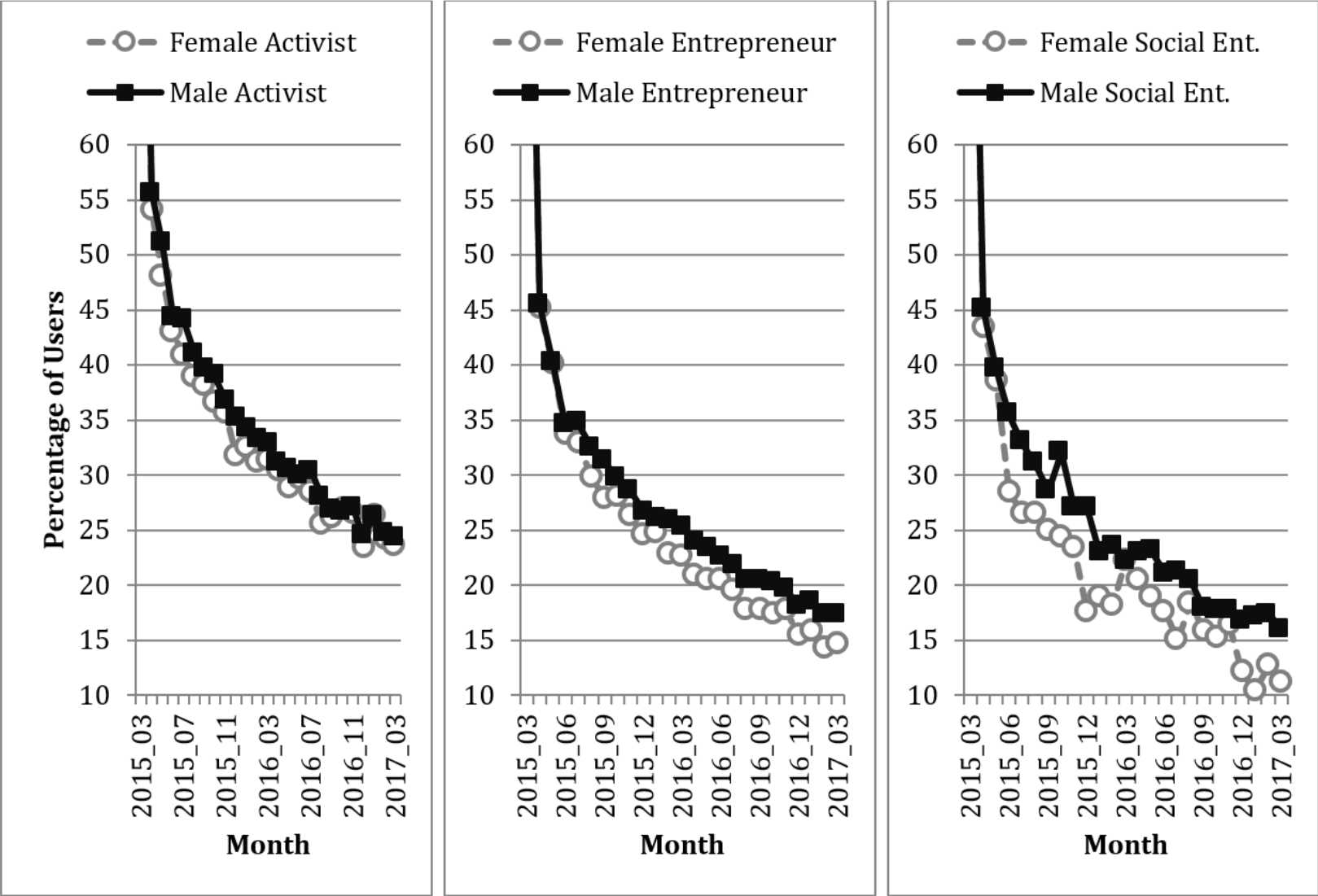
```
1 value,total_users
2 "日本",28049710
3 "United States",24045859
4 "Rio de Janeiro, Brasil",17858770
5 "14003018,"المملكة العربية السعودية
6 Brasil,13803252
7 "Los Angeles, CA",11827516
8 Indonesia,11333132
9 India,10582763
10 "London, England",10490193
11 she/her,10302513
12 "California, USA",10279992
13 "ประเทศไทย",10217045
14 Argentina,10177693
15 "São Paulo, Brasil",10103606
16 Thailand,9948245
17 "Buenos Aires, Argentina",9226902
18 "México",9127749
19 France,8592975
20 "İstanbul, Türkiye",8153317
```



Joint and Transition Probability Networks among Self-Descriptive Tokens in Personally Expressed Identity

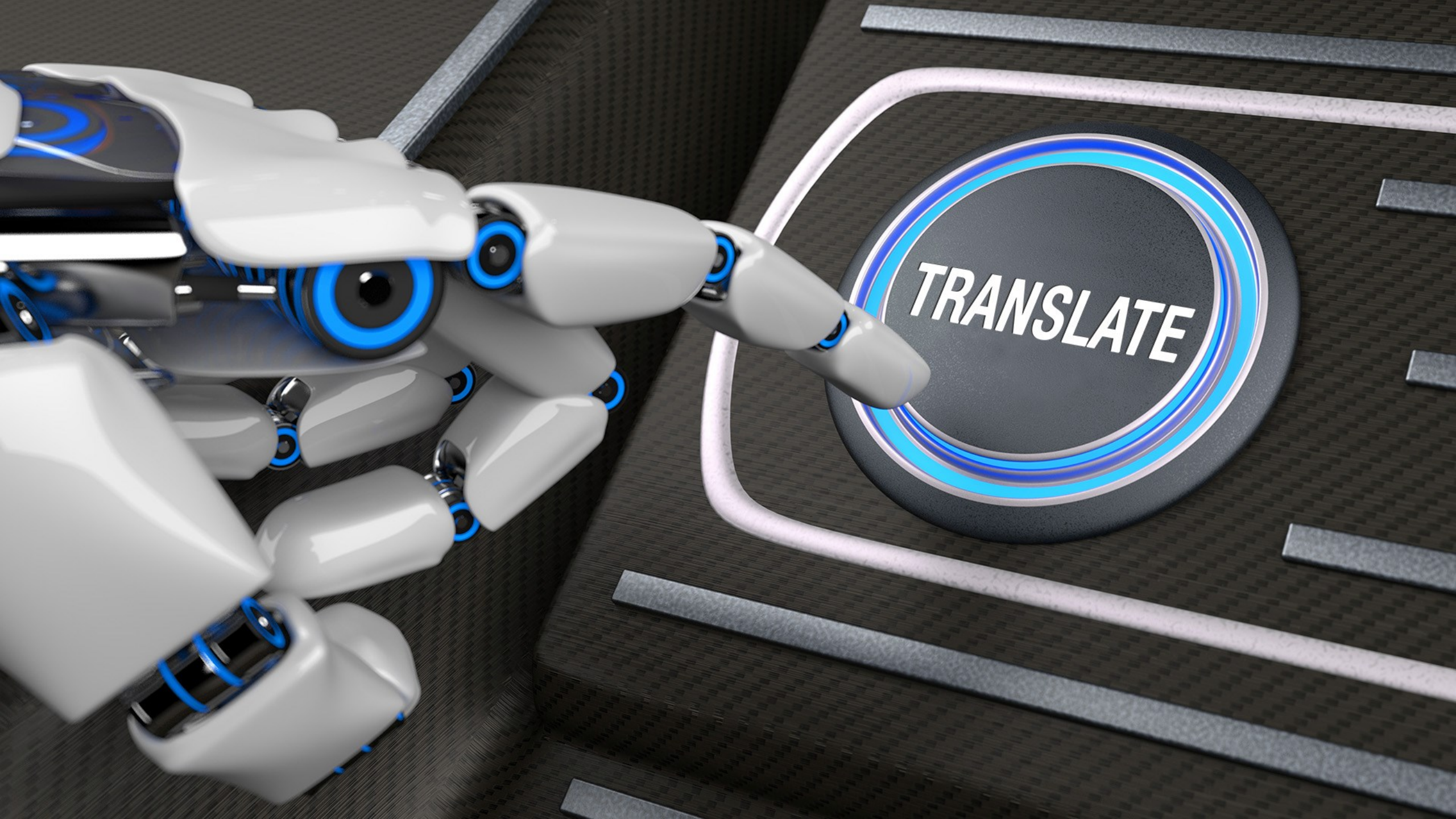


Durability Outliers in Social Role Signifiers



Predicting the Self

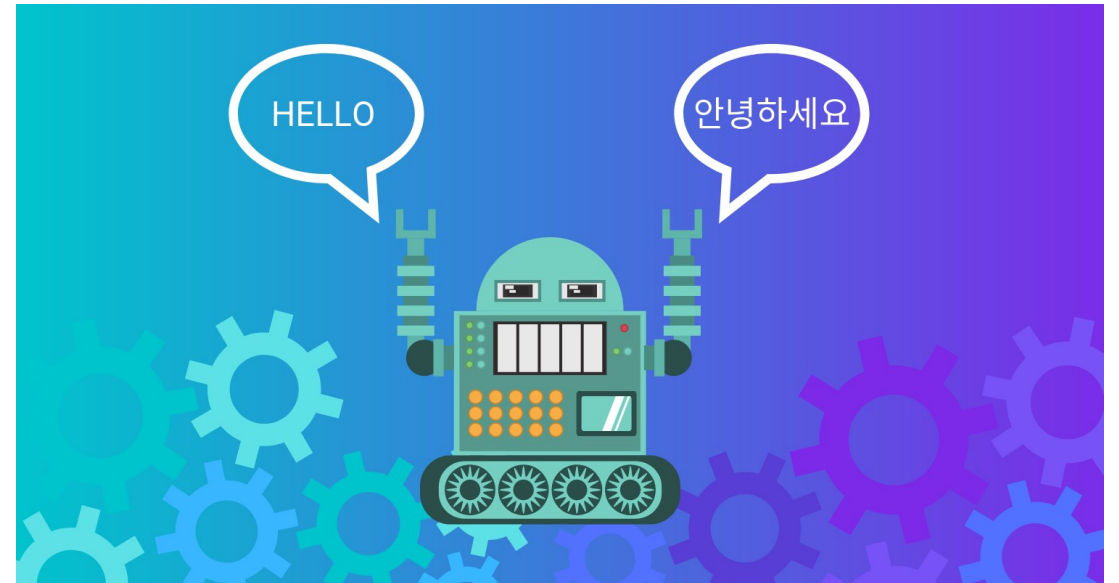
- We have best-ever data for ***personally expressed identity***.
 - Longitudinal, annual data for over 1 million Americans
- Retrospectively, we have examined changes in identity.
 - Personal development, cultural trends and responses to offline events
- What can we do with prediction?



TRANSLATE

Predicting the Self with Machine Translation

1. Use a pre-trained but flexible question-answering system.
 - Specifically, Google's T5.
2. Train to a new task:
 - Learn to predict one's 2020 bio given one's 2015 bio.
3. ???
4. Insight



T5: Text-To-Text Transfer Transformer



- “We propose reframing all NLP tasks into a unified text-to-text-format where the input and output are always text strings”
- <https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html>
- <https://arxiv.org/abs/1910.10683>
- <https://github.com/google-research/text-to-text-transfer-transformer>

T5: Text-To-Text Transfer Transformer

- Language model(s) trained on many online documents.
 - C4: Colossal Clean Crawled Corpus
 - ~400 million unique documents from Common Crawl
 - 1 TB of text
- Fine-tune to any arbitrary task by feeding examples of:
 - <name-of-task>: <training-example-input-string>
 - <training-example-output-string>
- For fun, compete with T5 on trivia:
 - <https://ai.googleblog.com/2020/02/exploring-transfer-learning-with-t5.html>

2015

I am a computational social scientist.

2020

I am a computational social scientist and an NFT artist.

TRAIN

2020

PREDICT

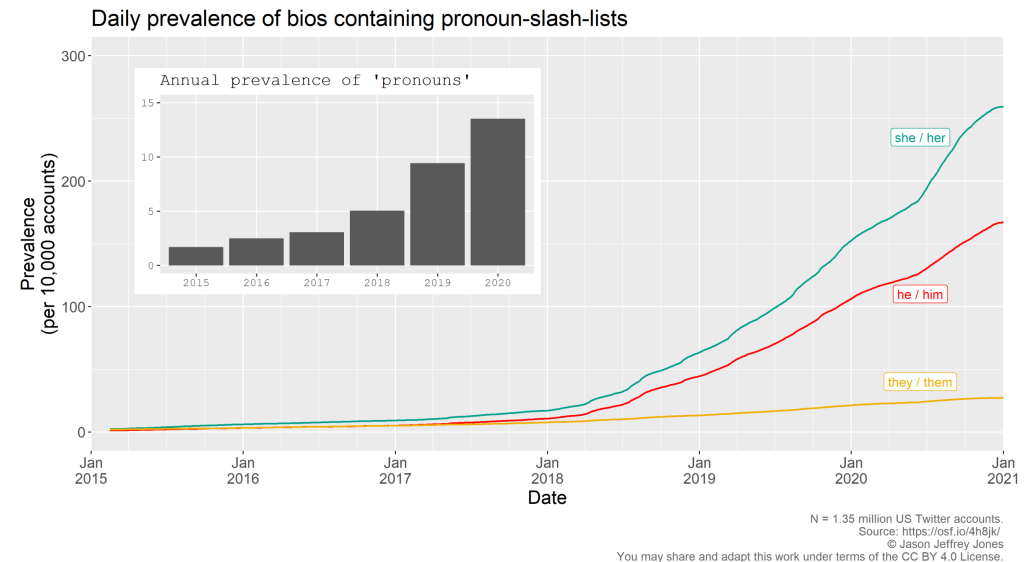
I am a computational social scientist and an NFT artist.

2025

I am a street performer and former scientist.

Predicting the Self with Machine Translation

- Are human lives predictable?
 - (Self-presentations)
 - Some aspects more than others?
 - E.g. parenthood vs. career
- Will the system make interesting mistakes?
 - Who will be mistakenly labelled Trump/MAGA, LGBTQ?
 - Who is “missing” pronouns?
- What will users think of their predicted future selves?
 - Amused, offended or bored?



Query: now, my, talk, is, actually, over

Sample: Cross-sectional sample

