Pardon Our Dust

public data infrastructure under construction

- 👷 Rahul Bhargava, Data Culture Group rahulbot@vis.social rahulbot
- 👷 Hayden DelCiello, Lazer Lab
- 👷 Zhen Guo, Lazer Lab VeraGuo5
- 👷 Alyssa Smith, Lazer Lab cetacean_needed cetacean_needed@mas.to

Background

the need for public data infrastructure

Why do we need public data infrastructure?

- Online media data mostly serves private interests, with very high barriers to access
- This data is critical for understanding changes in online media patterns and impacts
- There is a *public good* here here that is poorly understood and lacks tooling

Why are we “under construction”?

The landscape, builders, and tools all keep changing.

We’re trying to make the process of constructing this infrastructure transparent so it seems more achievable.

We need adaptable, functional infrastructure - but that doesn’t come out of thin air!

Existing Tools

Big Picture

An evolving ecosystem of online social/media archives

Building Infrastructure Together

The growing need to understand the societal impacts of online media mean we need to work together across sectors to build new shared digital public infrastructure.

Technically: APIs are how we can interoperate and together create more than our individual sum of parts.

Socially: sharing methods and approaches, governance structures, and convenings can help us build on each other's learnings.

There are a large set of partners we already know about, build on, and work with.

Media Cloud

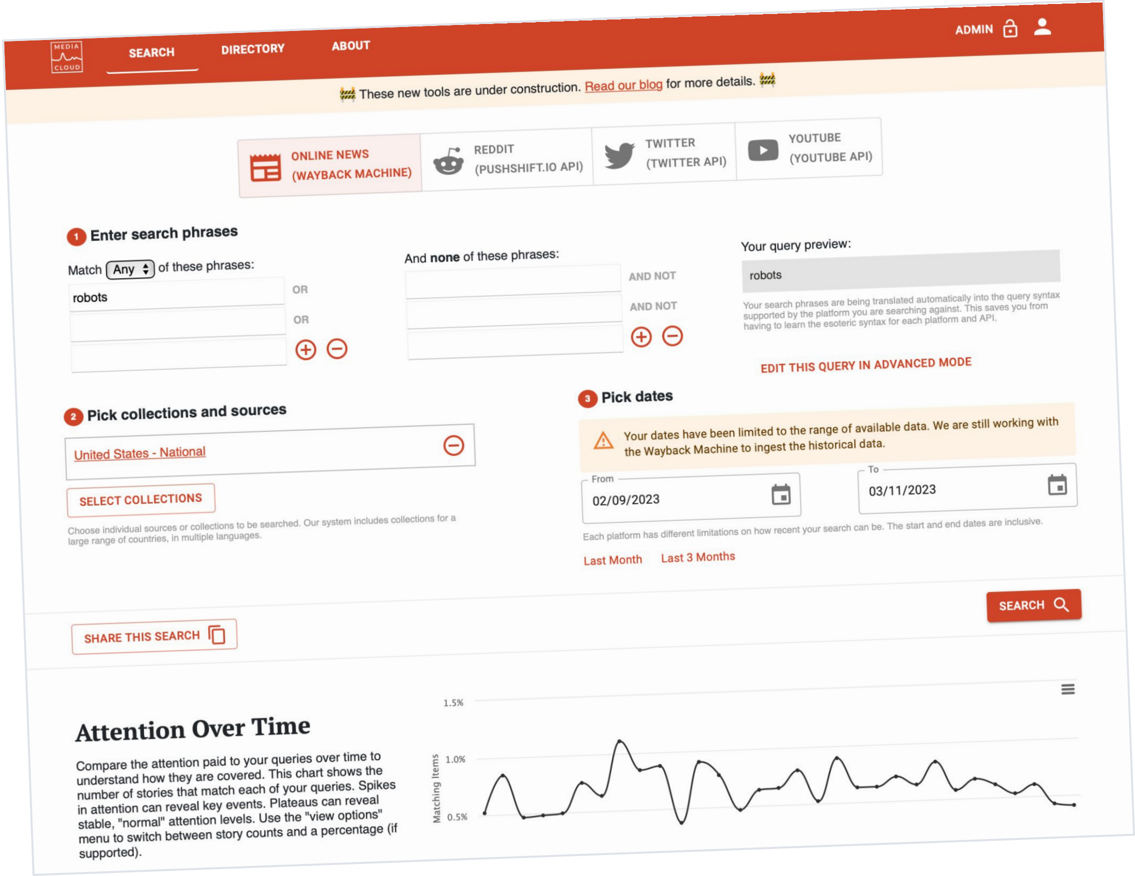
Access: free

Data Source: custom open news scraping

Interface: web-based search UI and API

Metadata: basic

Concerns: current data instability



🚧 These new tools are under construction. Read our blog for more details. 🚧

1 Enter search phrases

Match Any of these phrases:

robots OR

OR

And **none** of these phrases:

AND NOT

AND NOT

Your query preview:

robots

Your search phrases are being translated automatically into the query syntax supported by the platform you are searching against. This saves you from having to learn the esoteric syntax for each platform and API.

2 Pick collections and sources

United States - National

Choose individual sources or collections to be searched. Our system includes collections for a large range of countries, in multiple languages.

3 Pick dates

Your dates have been limited to the range of available data. We are still working with the Wayback Machine to ingest the historical data.

From 02/09/2023

To 03/11/2023

Each platform has different limitations on how recent your search can be. The start and end dates are inclusive.

Attention Over Time

Compare the attention paid to your queries over time to understand how they are covered. This chart shows the number of stories that match each of your queries. Spikes in attention can reveal key events. Plateaus can reveal stable, "normal" attention levels. Use the "view options" menu to switch between story counts and a percentage (if supported).

https://search.mediacloud.org

Stanford TV News Analyzer

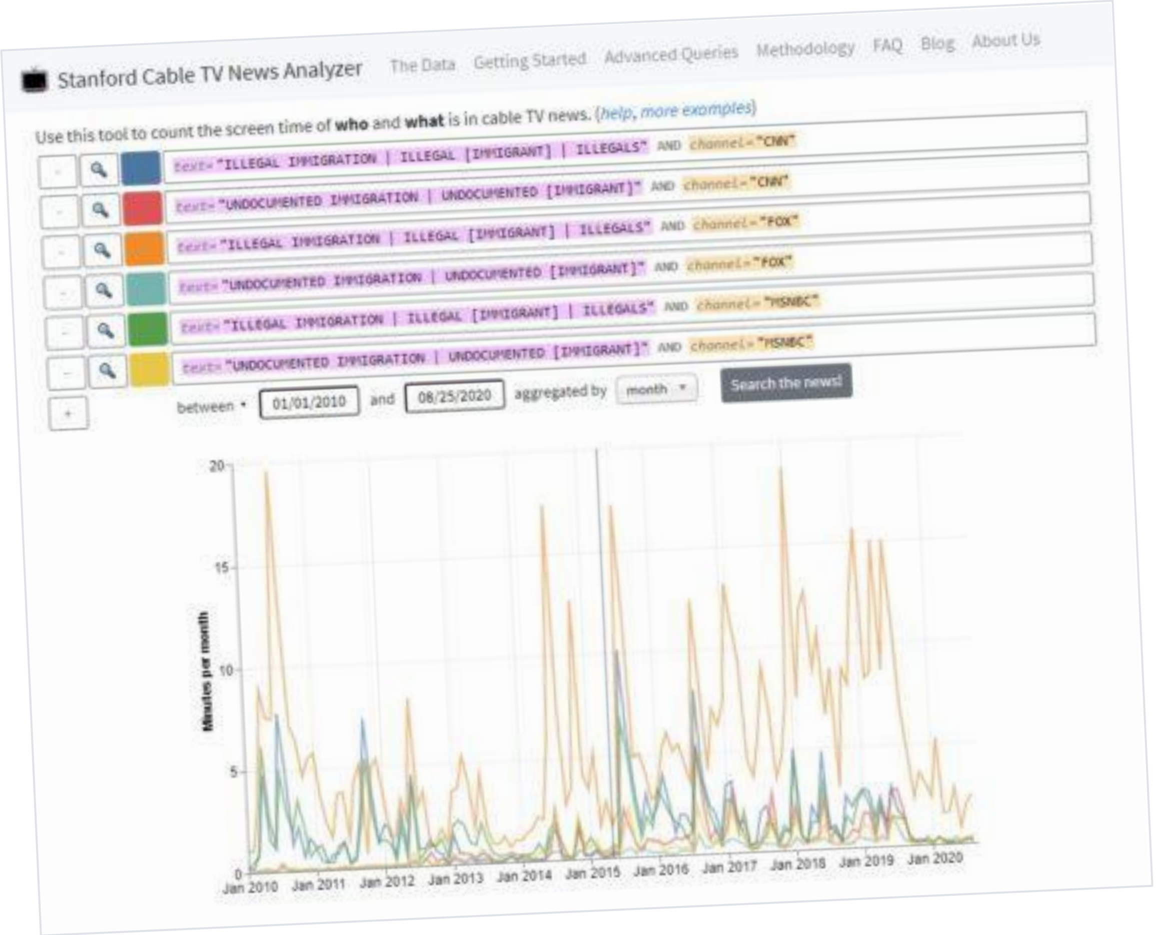
Access: free

Data Source: large-scale ingestion of Reddit content recordings of US-based cable news networks; indexed on closed captions

Interface: web-based search and analysis tool with CSV downloads

Metadata: time, channel, person entities, duration

Concerns: depends on cloud service support grants



Use this tool to count the screen time of **who** and **what** is in cable TV news. (*help*, *more examples*)

text="ILLEGAL IMMIGRATION | ILLEGAL [IMMIGRANT] | ILLEGALS" AND channel="CNN"

text="UNDOCUMENTED IMMIGRATION | UNDOCUMENTED [IMMIGRANT]" AND channel="CNN"

text="ILLEGAL IMMIGRATION | ILLEGAL [IMMIGRANT] | ILLEGALS" AND channel="FOX"

text="UNDOCUMENTED IMMIGRATION | UNDOCUMENTED [IMMIGRANT]" AND channel="FOX"

text="ILLEGAL IMMIGRATION | ILLEGAL [IMMIGRANT] | ILLEGALS" AND channel="MSNBC"

text="UNDOCUMENTED IMMIGRATION | UNDOCUMENTED [IMMIGRANT]" AND channel="MSNBC"

between 01/01/2010 and 08/25/2020 aggregated by month

https://tvnews.stanford.edu/

Pushshift.io

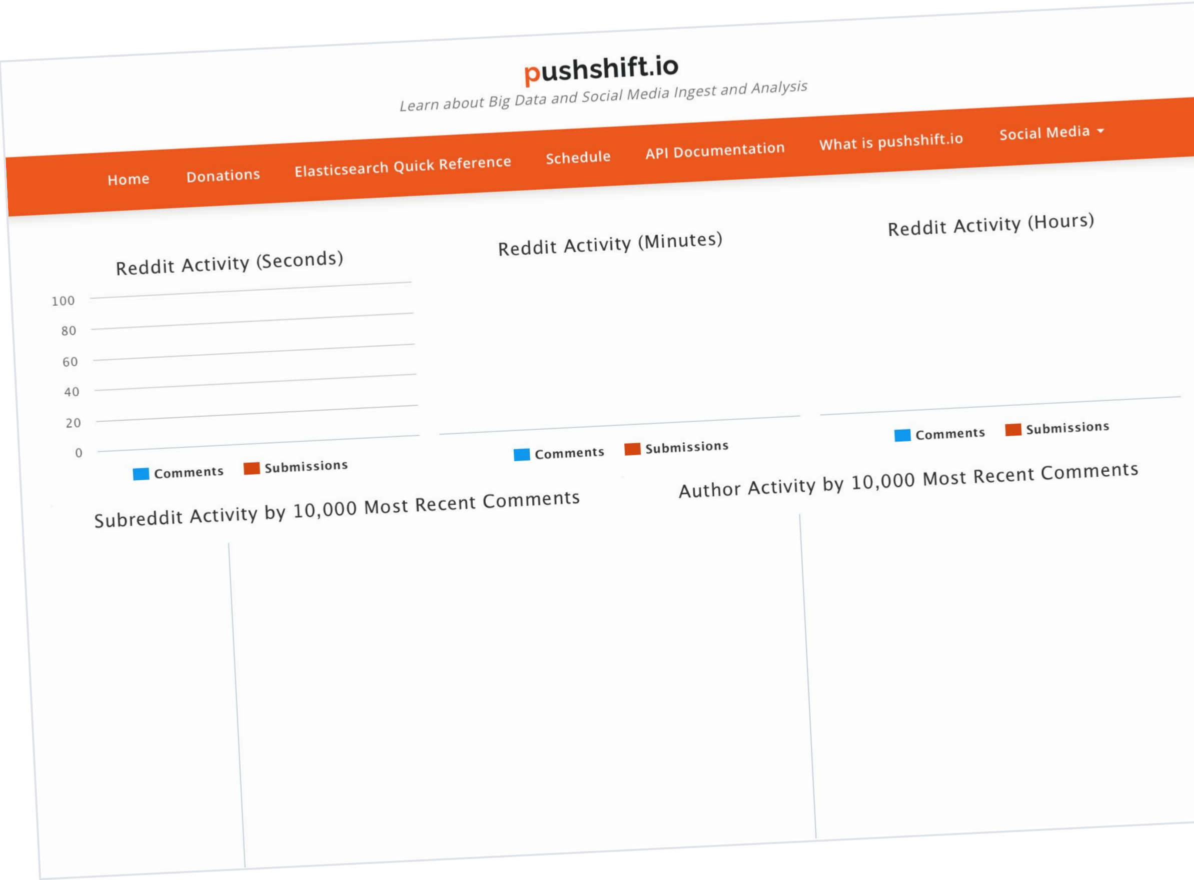
Access: free

Data Source: large-scale custom ingestion of Reddit submissions and comments

Interface: API

Metadata: time, subreddit, score, etc.

Concerns: handoff to NCRI leaves dataset in flux, and aligned with defense interests; take-down requests


pushshift.io
Learn about Big Data and Social Media Ingest and Analysis
Home
Donations
Elasticsearch Quick Reference
Schedule
API Documentation
What is pushshift.io
Social Media
Reddit Activity (Seconds)
100
80
60
40
20
0
Comments
Submissions
Reddit Activity (Minutes)
Comments
Submissions
Reddit Activity (Hours)
Comments
Submissions
Subreddit Activity by 10,000 Most Recent Comments
Author Activity by 10,000 Most Recent Comments

https://pushshift.io

CrowdTangle

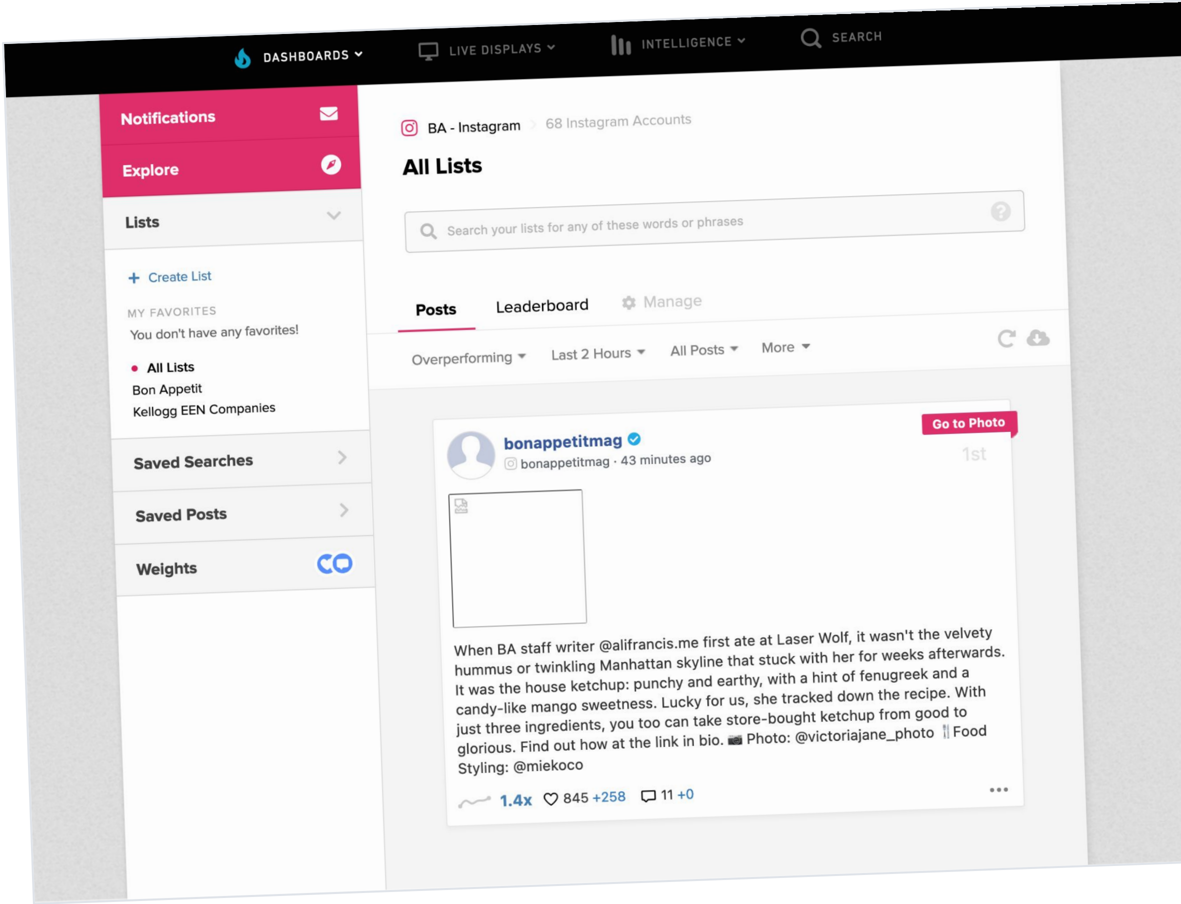
Access: free-ish

Data Source: Facebook, Instagram, Reddit

Interface: web-based monitors, API

Metadata: various

Concerns: owned by Meta; most of staff laid off; unreliable statements about content included and data; data only includes subsets of platform



DASHBOARDS LIVE DISPLAYS INTELLIGENCE SEARCH

Notifications

Explore

Lists

+ Create List

MY FAVORITES

You don't have any favorites!

- All Lists
- Bon Appetit
- Kellogg EEN Companies

Saved Searches

Saved Posts

Weights

BA - Instagram 68 Instagram Accounts

All Lists

Search your lists for any of these words or phrases

Posts Leaderboard Manage

Overperforming Last 2 Hours All Posts More

bonappetitmag

bonappetitmag · 43 minutes ago

Go to Photo

1st

When BA staff writer @alifrancis.me first ate at Laser Wolf, it wasn't the velvety hummus or twinkling Manhattan skyline that stuck with her for weeks afterwards. It was the house ketchup: punchy and earthy, with a hint of fenugreek and a candy-like mango sweetness. Lucky for us, she tracked down the recipe. With just three ingredients, you too can take store-bought ketchup from good to glorious. Find out how at the link in bio. 📷 Photo: @victoriajane_photo 🍴Food Styling: @miekoco

1.4x ♡ 845 +258 🗨 11 +0

https://crowdtangle.com/

newscatcher

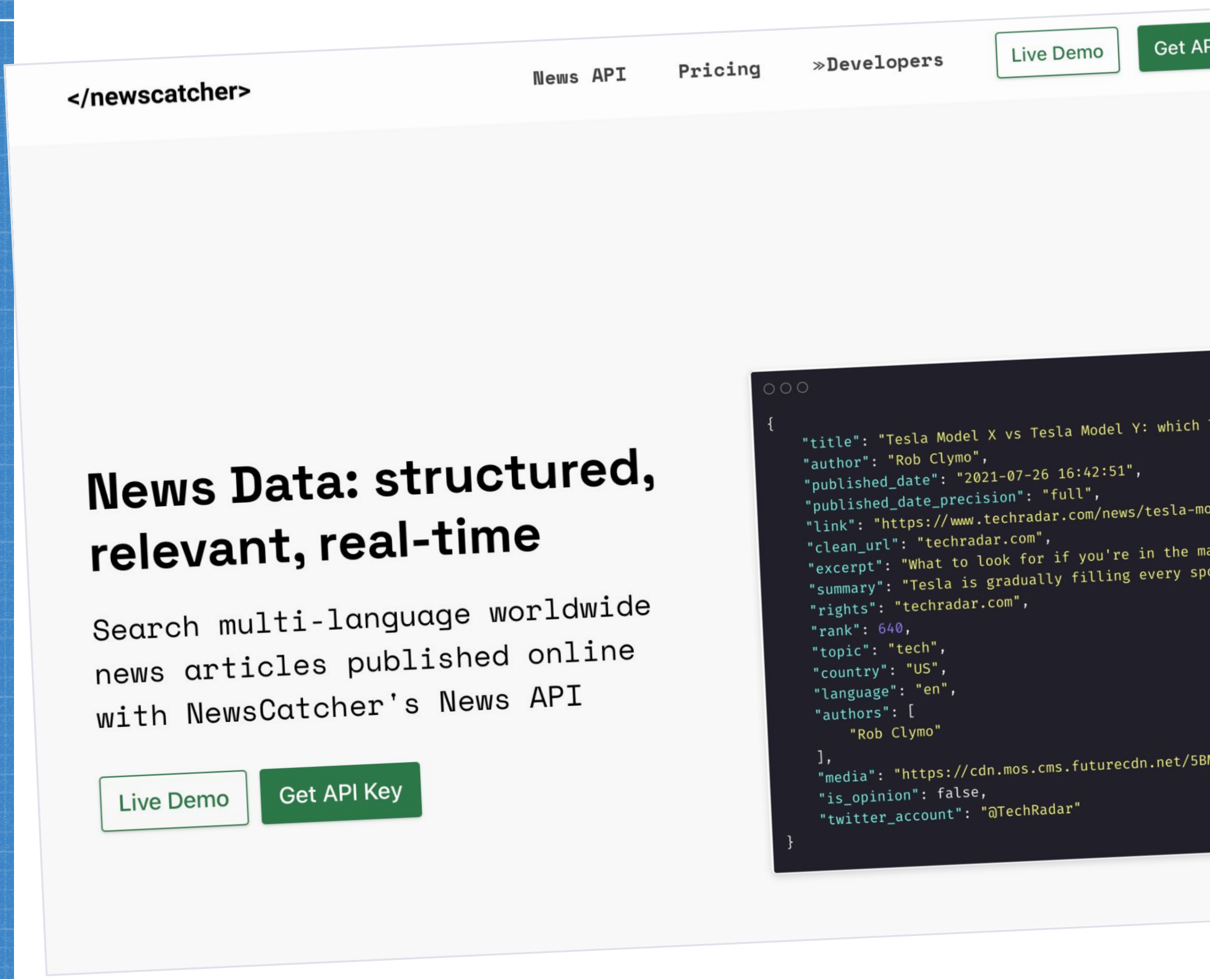
Access: freemium

Data Source: open news on the web

Interface: API

Metadata: date, title, authors, country, and others

Concerns: startup that might go out of business



News Data: structured, relevant, real-time

Search multi-language worldwide news articles published online with NewsCatcher's News API

```
{
    "title": "Tesla Model X vs Tesla Model Y: which
    "author": "Rob Clymo",
    "published_date": "2021-07-26 16:42:51",
    "published_date_precision": "full",
    "link": "https://www.techradar.com/news/tesla-mo
    "clean_url": "techradar.com",
    "excerpt": "What to look for if you're in the ma
    "summary": "Tesla is gradually filling every sp
    "rights": "techradar.com",
    "rank": 640,
    "topic": "tech",
    "country": "US",
    "language": "en",
    "authors": [
        "Rob Clymo"
    ],
    "media": "https://cdn.mos.cms.futurecdn.net/5BM
    "is_opinion": false,
    "twitter_account": "@TechRadar"
}
```

ttps://newscatcherapi.com

SMAT

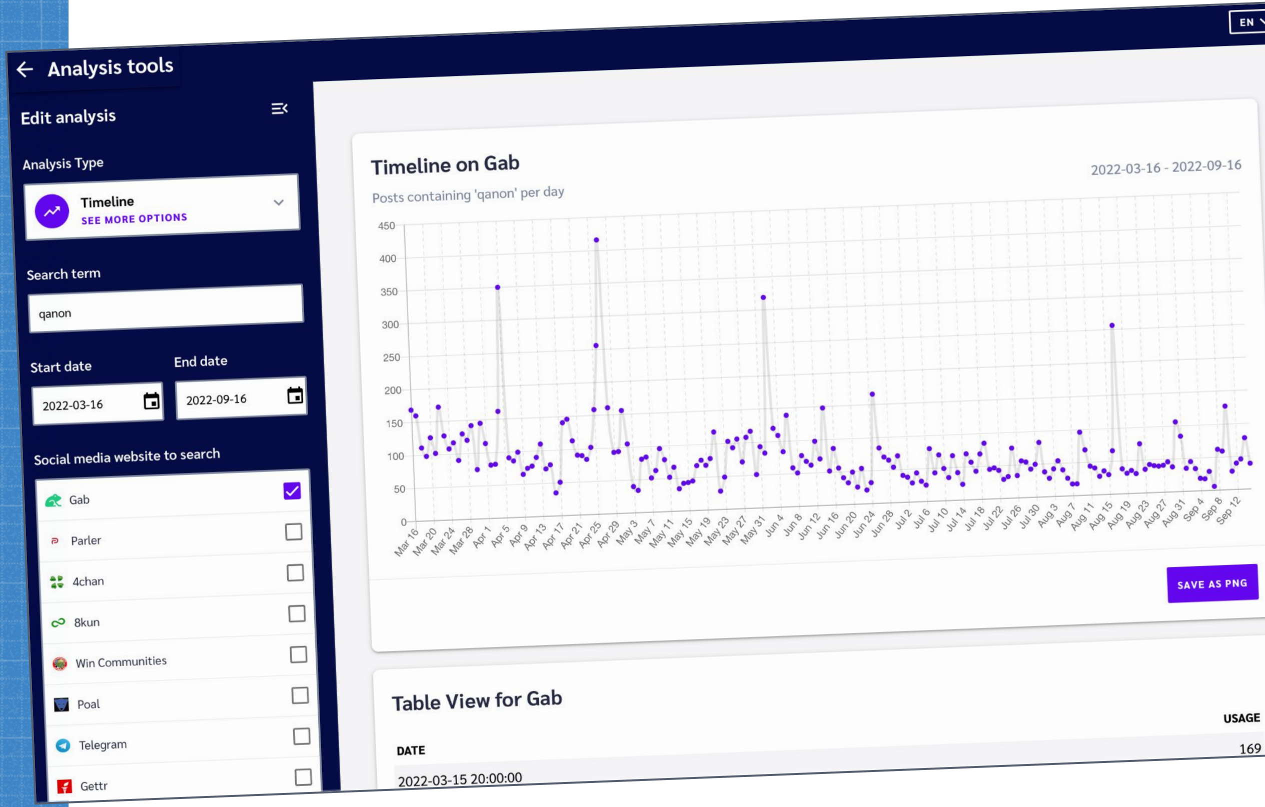
Access: freemium

Data Source: Parler, Telegram, 8kun, 4chan, Gab, and other fringe sites

Interface: web-based search UI and API

Metadata: basic

Concerns: longer term sustainability; sustainability of antagonistic data collection



Analysis Type: Timeline

Search term: qanon

Start date: 2022-03-16

End date: 2022-09-16

Social media website to search

- Gab ☑
- Parler □
- 4chan □
- 8kun □
- Win Communities □
- Poal □
- Telegram □
- Gettr □

Timeline on Gab

2022-03-16 - 2022-09-16

Posts containing 'qanon' per day

Table View for Gab

DATE	USAGE
2022-03-15 20:00:00	169

https://www.smat-app.com

Platform APIs

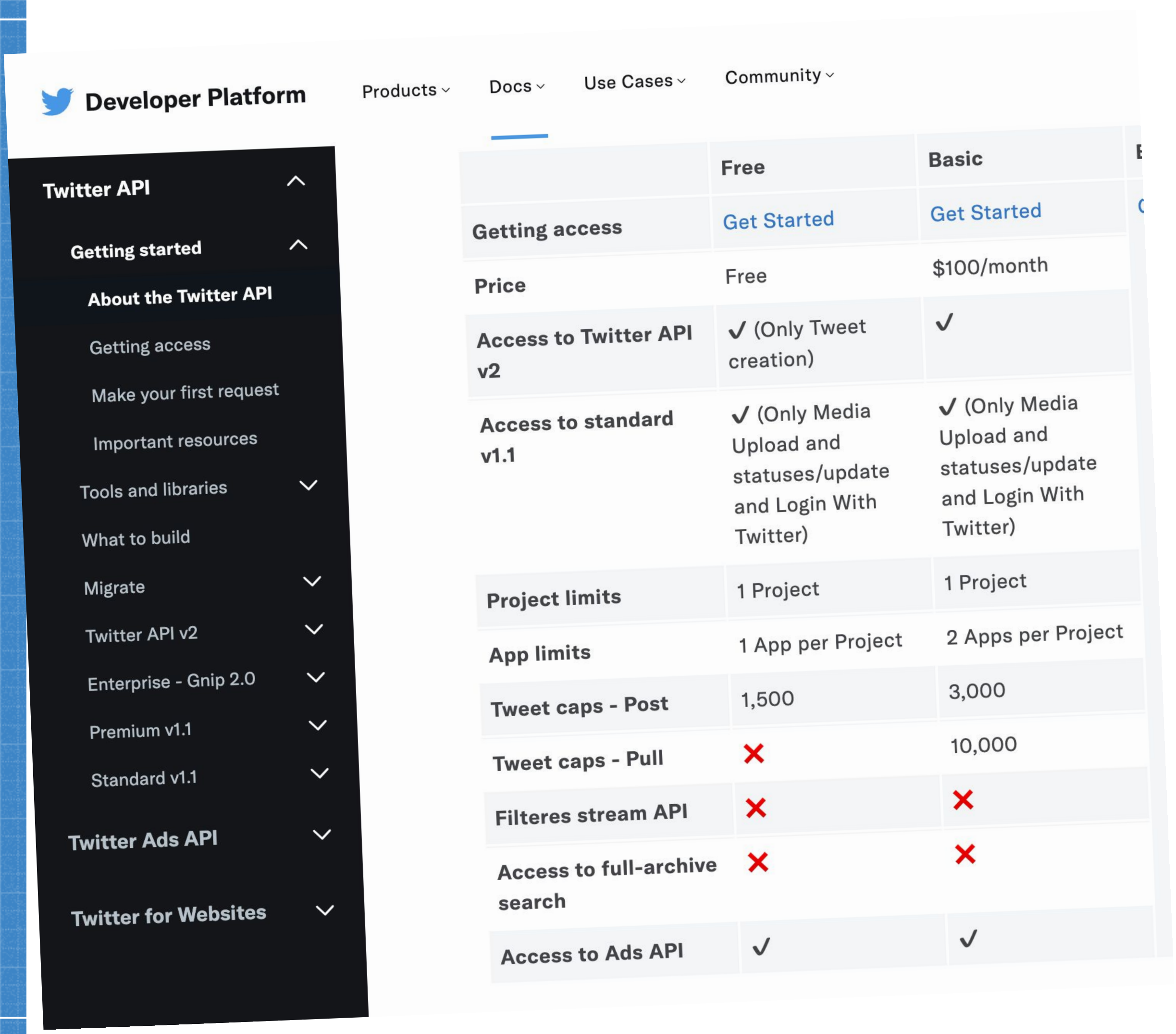
Access: freemium

Data Source: internal

Interface: API calls

Metadata: most of it

Concerns: cost; lack of transparency; real usage limits; changing policies



	Free	Basic
Getting access	Get Started	Get Started
Price	Free	$100/month
Access to Twitter API v2	✓ (Only Tweet creation)	✓
Access to standard v1.1	✓ (Only Media Upload and statuses/update and Login With Twitter)	✓ (Only Media Upload and statuses/update and Login With Twitter)
Project limits	1 Project	1 Project
App limits	1 App per Project	2 Apps per Project
Tweet caps - Post	1,500	3,000
Tweet caps - Pull	✗	10,000
Filteres stream API	✗	✗
Access to full-archive search	✗	✗
Access to Ads API	✓	✓

Social Analytics Companies

We've used Brandwatch (used to be called Crimson Hexagon).

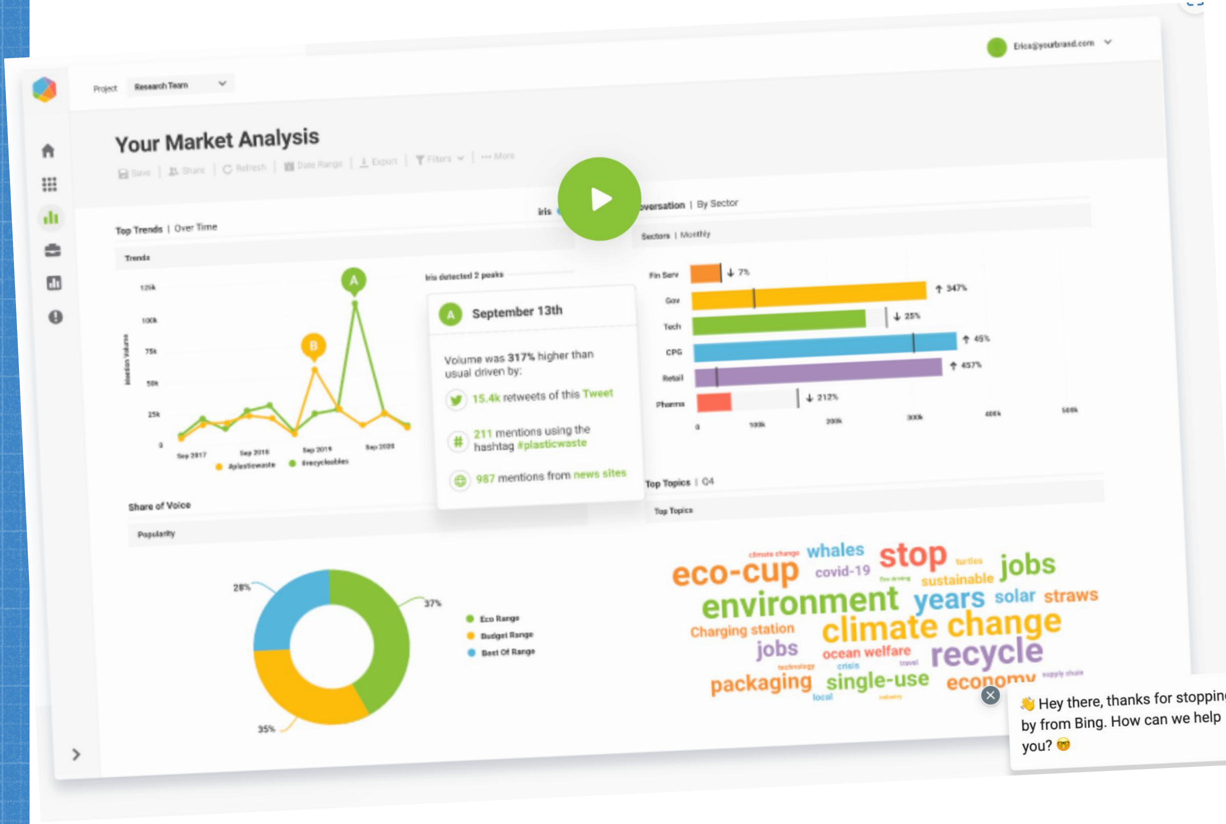
Access: costs real money

Data Source: paying social media platforms for access

Interface: web-based dashboards and API calls

Metadata: loads

Concerns: little transparency; cost; unclear methods for things like sampling



A September 13th

Volume was 317% higher than usual driven by:

15.4k retweets of this Tweet

211 mentions using the hashtag #plasticwaste

987 mentions from news sites

👋 Hey there, thanks for stopping by from Bing. How can we help you? 🤓

Media Cloud

an searchable public archive of
global online news stories

Media Cloud is...

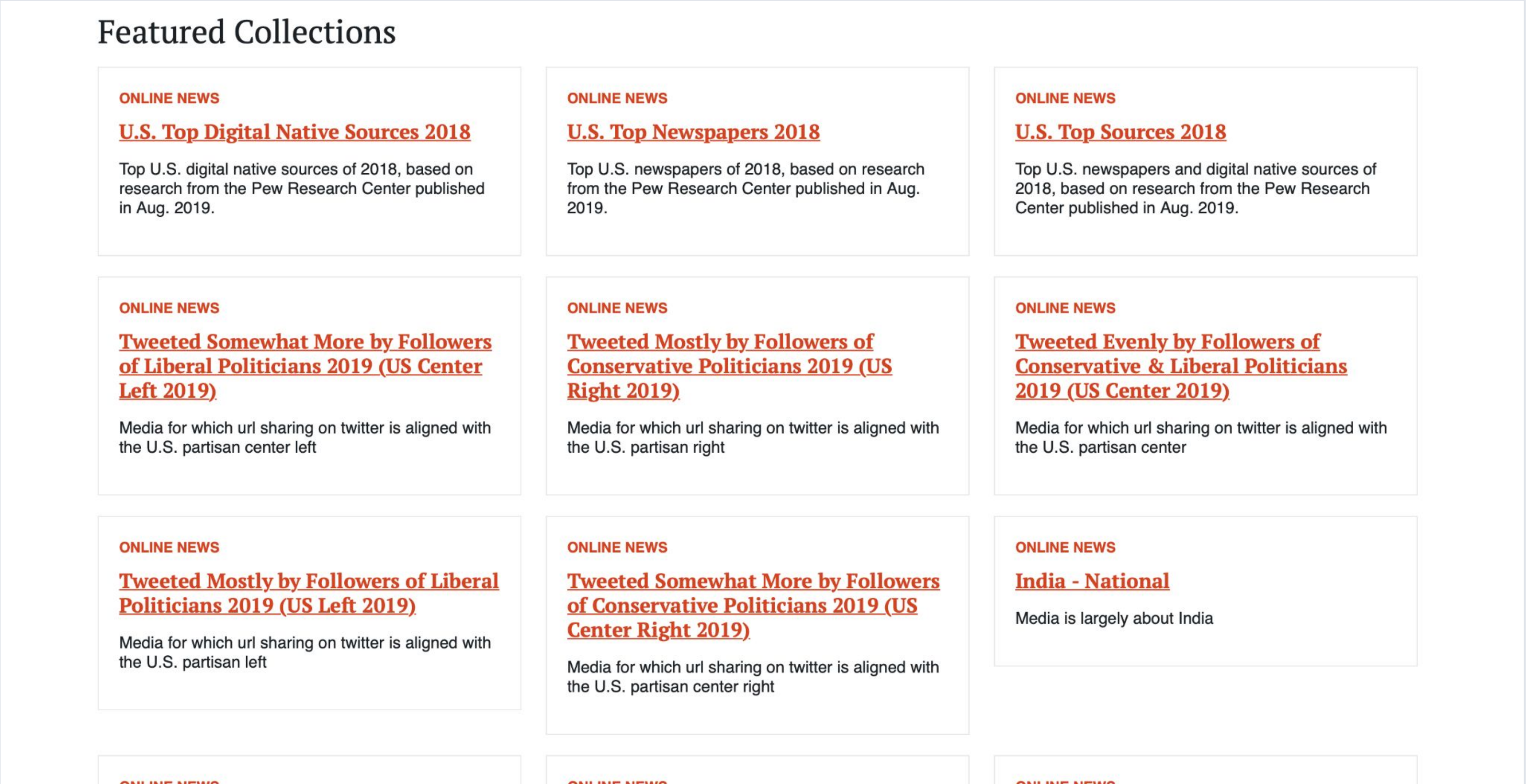
The most comprehensive **database of digital news** in the world available to researchers.

A set of online analysis and visualization **tools and methods.**

A team of **technologists and researchers**.

A **research service** for a community of academics, journalists, foundations, and nonprofits that want to understand the online media ecosystem.

The open news archive: what is/isn’t it?



Featured Collections

ONLINE NEWS

U.S. Top Digital Native Sources 2018

Top U.S. digital native sources of 2018, based on research from the Pew Research Center published in Aug. 2019.

ONLINE NEWS

U.S. Top Newspapers 2018

Top U.S. newspapers of 2018, based on research from the Pew Research Center published in Aug. 2019.

ONLINE NEWS

U.S. Top Sources 2018

Top U.S. newspapers and digital native sources of 2018, based on research from the Pew Research Center published in Aug. 2019.

ONLINE NEWS

Tweeted Somewhat More by Followers of Liberal Politicians 2019 (US Center Left 2019)

Media for which url sharing on twitter is aligned with the U.S. partisan center left

ONLINE NEWS

Tweeted Mostly by Followers of Conservative Politicians 2019 (US Right 2019)

Media for which url sharing on twitter is aligned with the U.S. partisan right

ONLINE NEWS

Tweeted Evenly by Followers of Conservative & Liberal Politicians 2019 (US Center 2019)

Media for which url sharing on twitter is aligned with the U.S. partisan center

ONLINE NEWS

Tweeted Mostly by Followers of Liberal Politicians 2019 (US Left 2019)

Media for which url sharing on twitter is aligned with the U.S. partisan left

ONLINE NEWS

Tweeted Somewhat More by Followers of Conservative Politicians 2019 (US Center Right 2019)

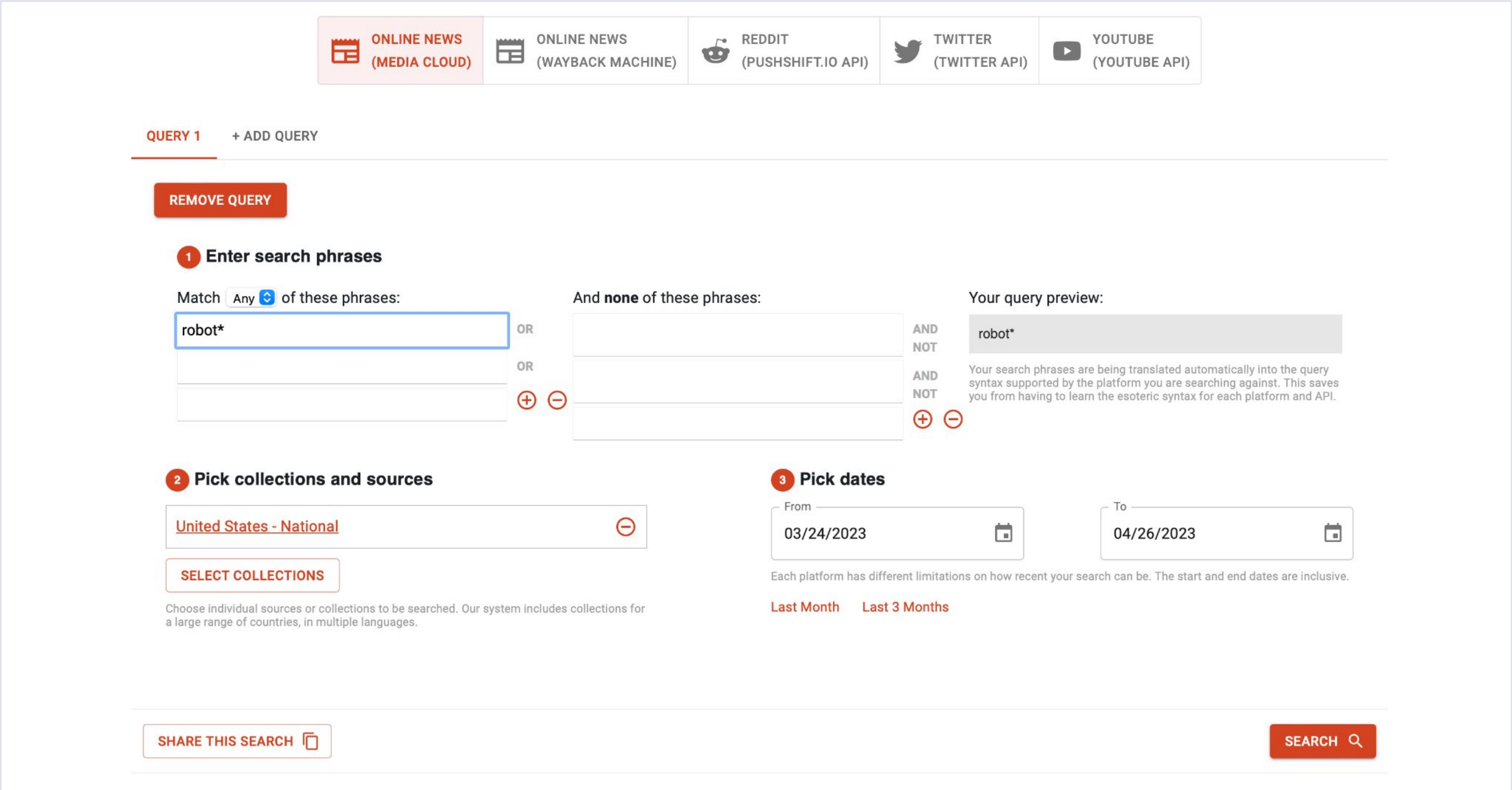
Media for which url sharing on twitter is aligned with the U.S. partisan center right

ONLINE NEWS

India - National

Media is largely about India

Search interface demonstration



1 Enter search phrases

Match Any of these phrases: robot*

And **none** of these phrases:

Your query preview: robot*

Your search phrases are being translated automatically into the query syntax supported by the platform you are searching against. This saves you from having to learn the esoteric syntax for each platform and API.

2 Pick collections and sources

United States - National

Choose individual sources or collections to be searched. Our system includes collections for a large range of countries, in multiple languages.

3 Pick dates

From: 03/24/2023

To: 04/26/2023

Each platform has different limitations on how recent your search can be. The start and end dates are inclusive.

Existing Applications: What are people doing now?

- Researching misinformation
- Monitoring media
- Using news as data

Technical concerns

Stability - large software and data infrastructures are hard to maintain without a team of technical PhD students or staff

Grants - few private funders want to support infrastructure and maintenance, but public funding is competitive too

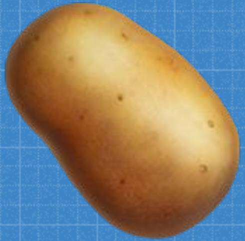
Techno-social concerns

Copyright - 2 billions documents is a *lot* of copyright liability

Compute Power - processor cycles add up to become expensive very quickly

Governance - shared resources require shared governance structures, which need trust, shared interests, etc.

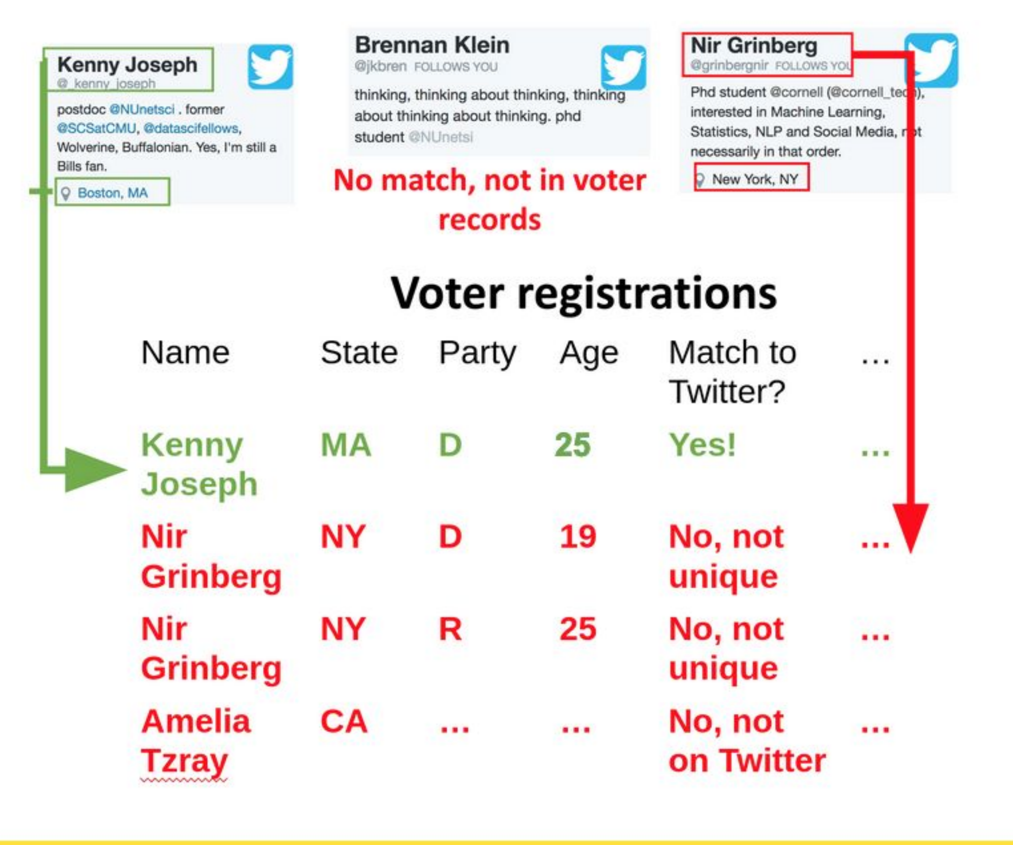
POTATO



Panel-based Open Term-level
Aggregate Twitter Observatory

What we’re starting with: the Twitter panel dataset

- Matches state-level voter records + demographic info to social media accounts.
- We have ~1.6 million users’ tweets from the inception of Twitter to around October 2022.



Kenny Joseph
@_kenny_joseph
postdoc @NUnetsci . former @SCSatCMU, @datascifellows, Wolverine, Buffalonian. Yes, I'm still a Bills fan.
Boston, MA

Brennan Klein
@jkbren FOLLOWS YOU
thinking, thinking about thinking, thinking about thinking about thinking. phd student @NUnetsi

No match, not in voter records

Nir Grinberg
@grinbergnir FOLLOWS YOU
Phd student @cornell (@cornell_tech), interested in Machine Learning, Statistics, NLP and Social Media, not necessarily in that order.
New York, NY

Voter registrations

Name	State	Party	Age	Match to Twitter?	…
Kenny Joseph	MA	D	25	Yes!	…
Nir Grinberg	NY	D	19	No, not unique	…
Nir Grinberg	NY	R	25	No, not unique	…
Amelia Tzray	CA	…	…	No, not on Twitter	…

Privacy Concerns

- **This is sensitive data!** Our data usage agreement doesn’t allow us to share the voter files or the linkages from voter file to Twitter accounts.
- We are able to report aggregate statistics about the panel (e.g. “What is the gender breakdown of people tweeting about lizards?”)
- Even when we report aggregate statistics, if the buckets become small, we worry about identifiability.

Our Solution (a very brief overview)

- POTATO (Panel-Based Open Term-Level Aggregate Twitter Observatory) is a searchable version of the Twitter panel.
- It returns search results in aggregate along with data visualizations of the demographic/geographic information.
- Users can also download a JSON blob of the aggregate data for further analysis.

Searchability

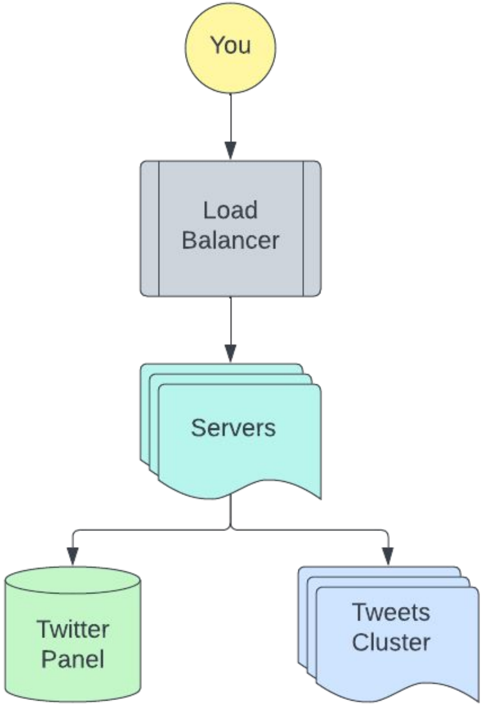
- You search Tweet text -> We tell you who Tweeted it

- We store Tweets from the Twitter Panel

- Searches not hindered by Twitter API limits

Scalability

- On-demand aggregation is expensive
- Vertically scalable in hardware
- Horizontally scalable in software



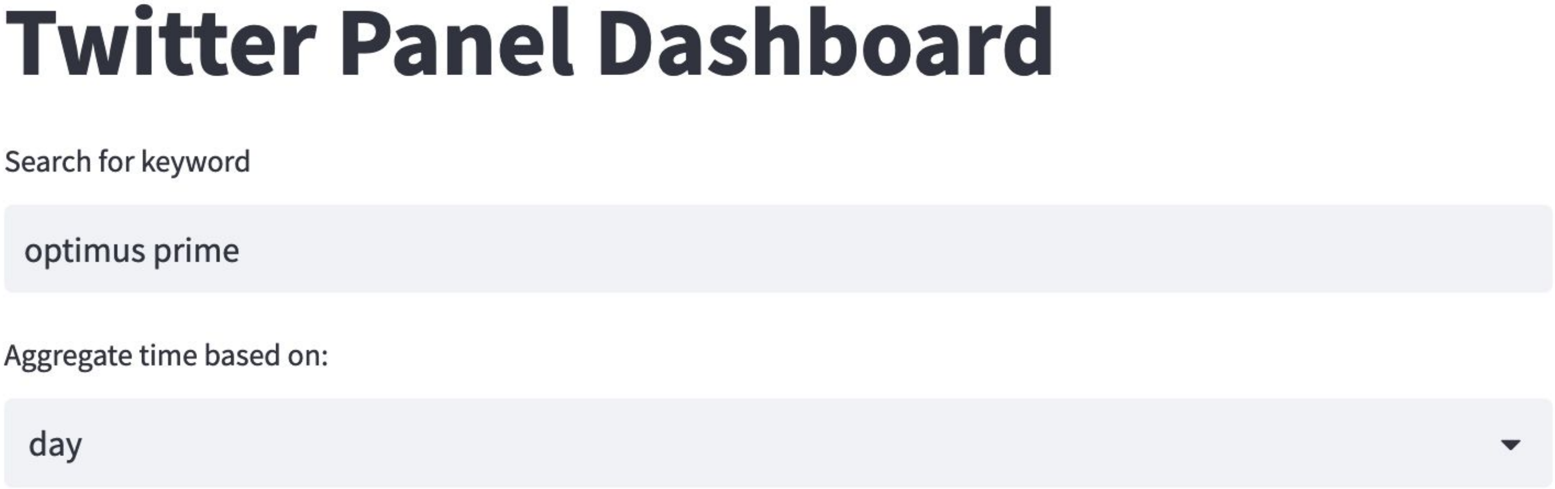
Visualization Tool & Website UI

- Streamlit
- Open-source
- Pure python
- Beautiful visualization
- Easy and quick to learn
- Support all the features we need

Streamlit

A Quick Video of POTATO In Action

Visualization Tool & Website UI



Twitter Panel Dashboard

Search for keyword

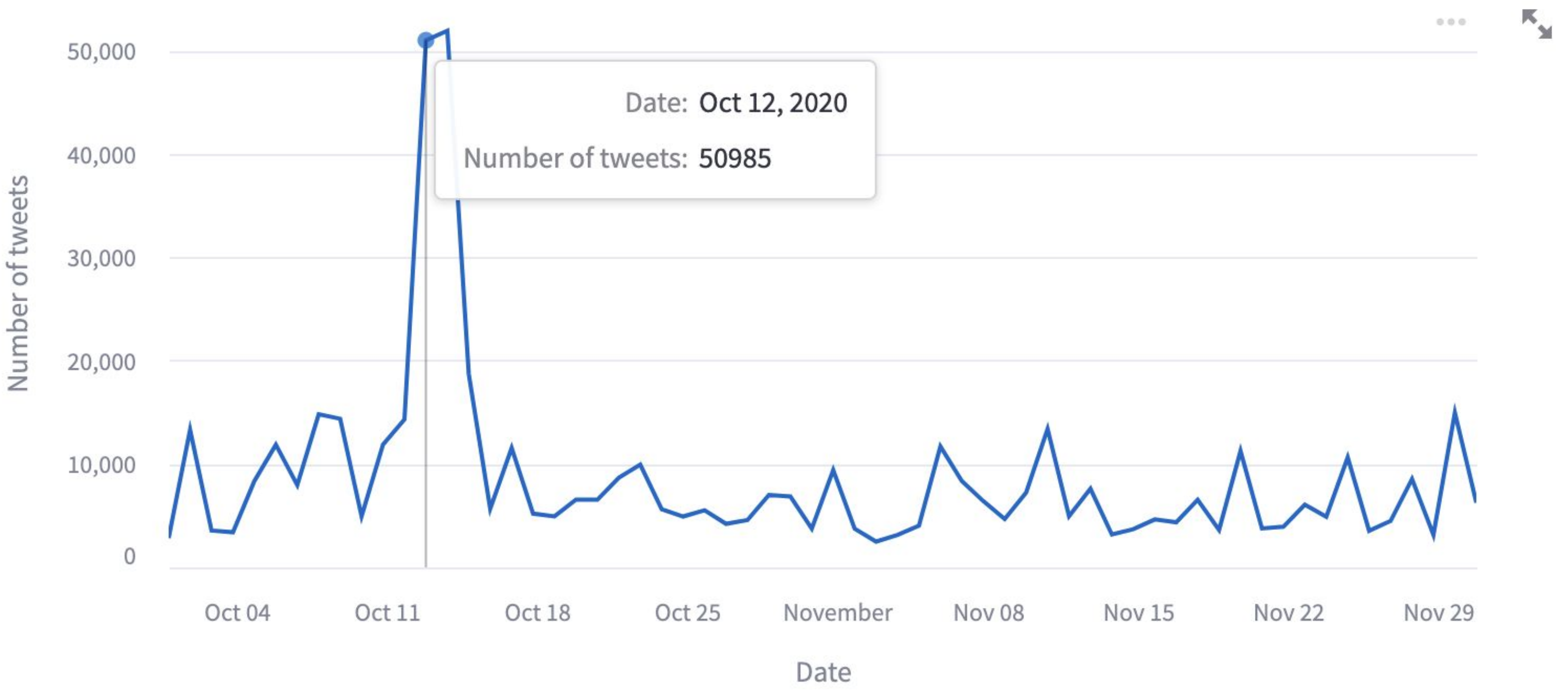
optimus prime

Aggregate time based on:

day

Visualization Tool & Website UI

Number of tweets per day

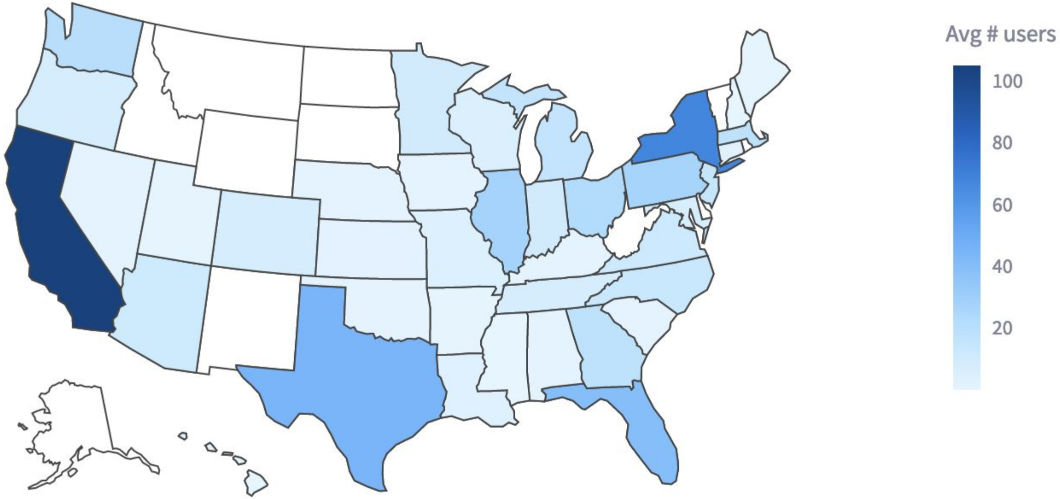


Date: Oct 12, 2020

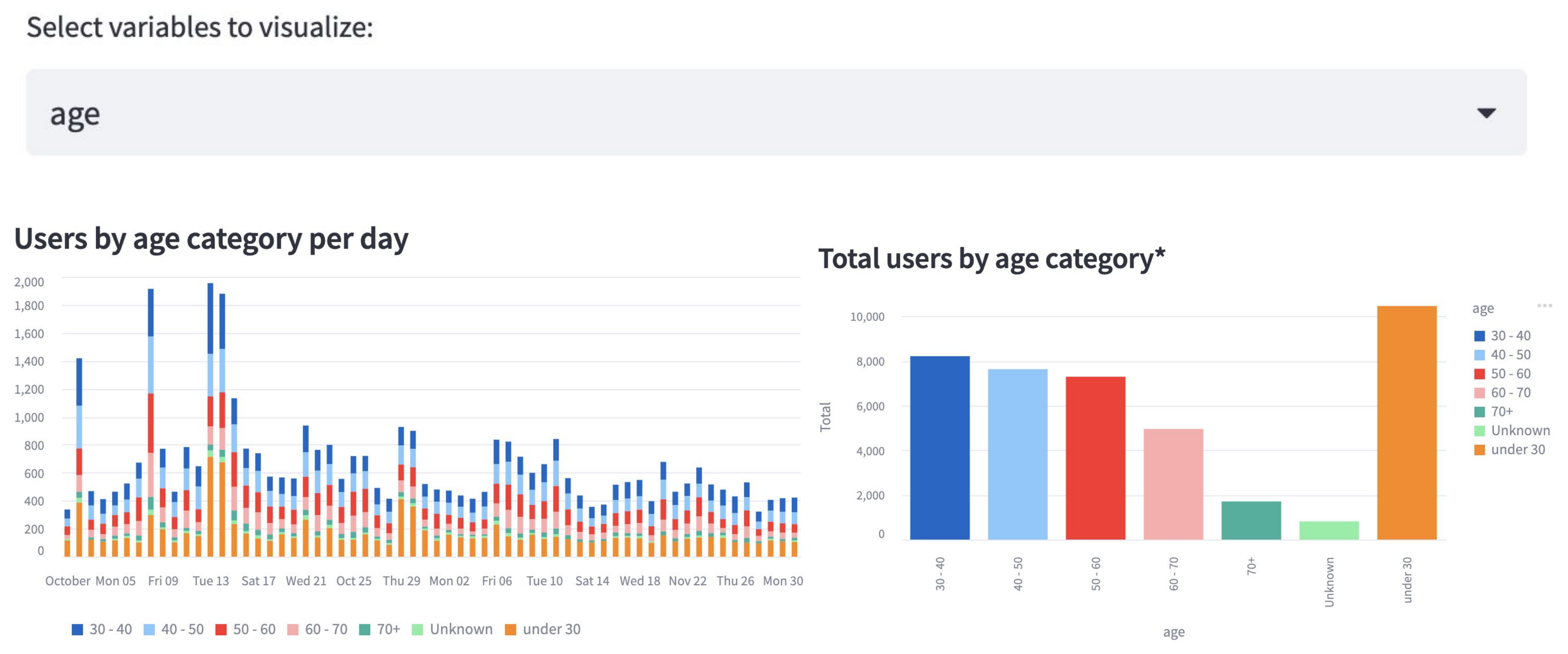
Number of tweets: 50985

Visualization Tool & Website UI

Average number of unique users per day



Visualization Tool & Website UI



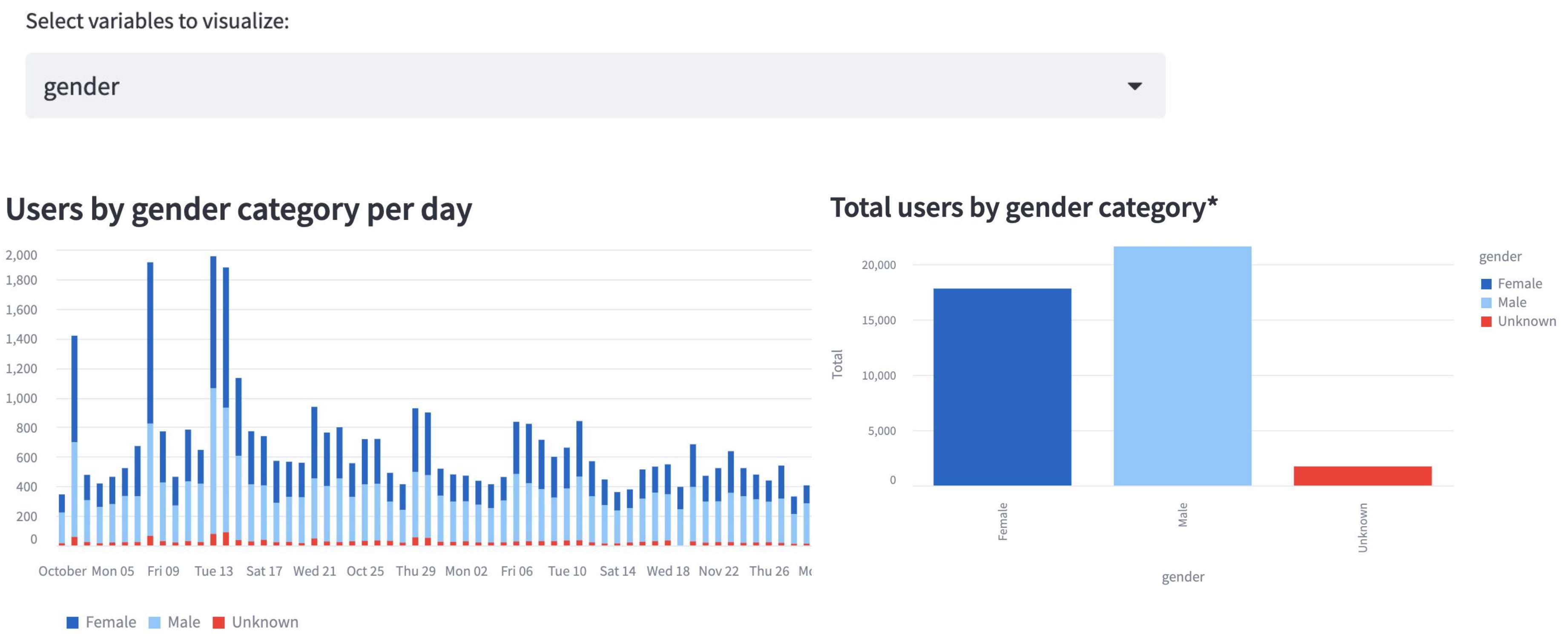
Select variables to visualize:

age

Users by age category per day

Total users by age category*

Visualization Tool & Website UI



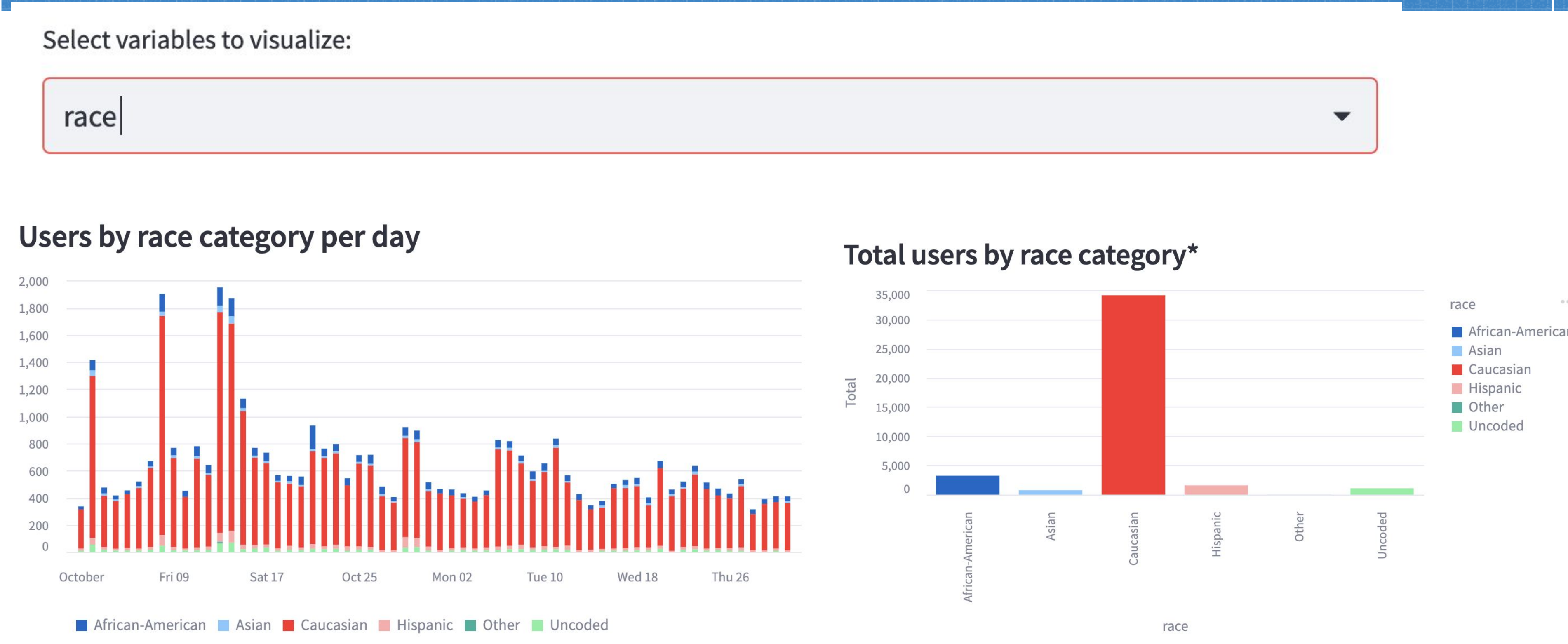
Select variables to visualize:

gender ▾

Users by gender category per day

Total users by gender category*

Visualization Tool & Website UI



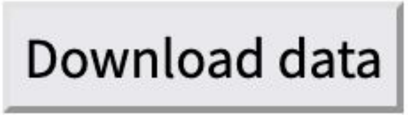
Select variables to visualize:

race

Users by race category per day

Total users by race category*

Visualization Tool & Website UI



Download data

age.csv

df.csv

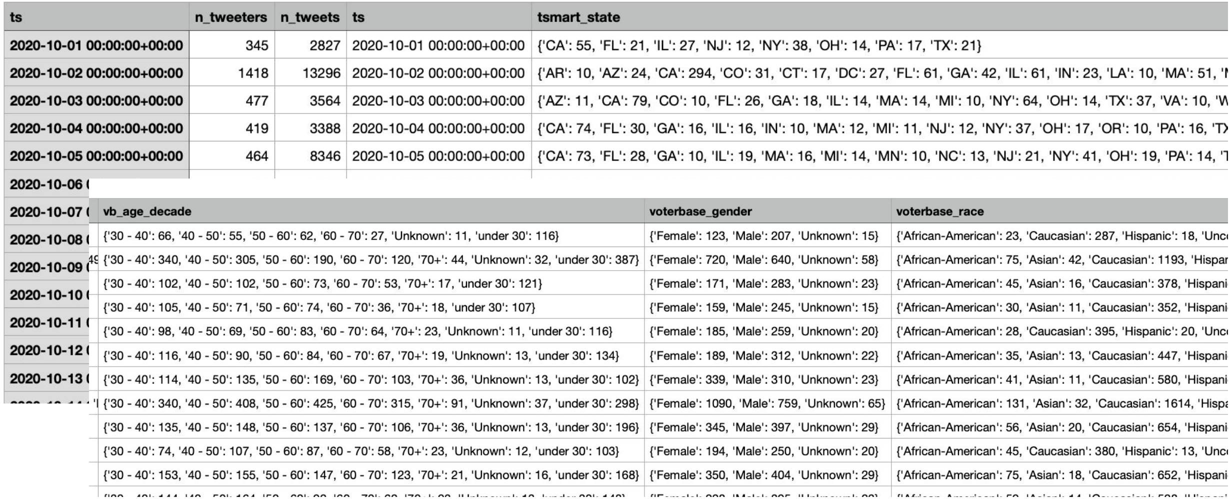
gender.csv

race.csv

states.csv

data.zip.zip

44 KB



ts	n_tweeters	n_tweets	ts	tsmart_state
2020-10-01 00:00:00+00:00	345	2827	2020-10-01 00:00:00+00:00	{'CA': 55, 'FL': 21, 'IL': 27, 'NJ': 12, 'NY': 38, 'OH': 14, 'PA': 17, 'TX': 21}
2020-10-02 00:00:00+00:00	1418	13296	2020-10-02 00:00:00+00:00	{'AR': 10, 'AZ': 24, 'CA': 294, 'CO': 31, 'CT': 17, 'DC': 27, 'FL': 61, 'GA': 42, 'IL': 61, 'IN': 23, 'LA': 10, 'MA': 51, 'N
2020-10-03 00:00:00+00:00	477	3564	2020-10-03 00:00:00+00:00	{'AZ': 11, 'CA': 79, 'CO': 10, 'FL': 26, 'GA': 18, 'IL': 14, 'MA': 14, 'MI': 10, 'NY': 64, 'OH': 14, 'TX': 37, 'VA': 10, 'W
2020-10-04 00:00:00+00:00	419	3388	2020-10-04 00:00:00+00:00	{'CA': 74, 'FL': 30, 'GA': 16, 'IL': 16, 'IN': 10, 'MA': 12, 'MI': 11, 'NJ': 12, 'NY': 37, 'OH': 17, 'OR': 10, 'PA': 16, 'TX
2020-10-05 00:00:00+00:00	464	8346	2020-10-05 00:00:00+00:00	{'CA': 73, 'FL': 28, 'GA': 10, 'IL': 19, 'MA': 16, 'MI': 14, 'MN': 10, 'NC': 13, 'NJ': 21, 'NY': 41, 'OH': 19, 'PA': 14, 'T
2020-10-06 (				
2020-10-07 (				
2020-10-08 (				
2020-10-09 (				
2020-10-10 (				
2020-10-11 (				
2020-10-12 (				
2020-10-13 (				

vb_age_decade	voterbase_gender	voterbase_race
{'30 - 40': 66, '40 - 50': 55, '50 - 60': 62, '60 - 70': 27, 'Unknown': 11, 'under 30': 116}	{'Female': 123, 'Male': 207, 'Unknown': 15}	{'African-American': 23, 'Caucasian': 287, 'Hispanic': 18, 'Unc
{'30 - 40': 340, '40 - 50': 305, '50 - 60': 190, '60 - 70': 120, '70+': 44, 'Unknown': 32, 'under 30': 387}	{'Female': 720, 'Male': 640, 'Unknown': 58}	{'African-American': 75, 'Asian': 42, 'Caucasian': 1193, 'Hispar
{'30 - 40': 102, '40 - 50': 102, '50 - 60': 73, '60 - 70': 53, '70+': 17, 'under 30': 121}	{'Female': 171, 'Male': 283, 'Unknown': 23}	{'African-American': 45, 'Asian': 16, 'Caucasian': 378, 'Hispani
{'30 - 40': 105, '40 - 50': 71, '50 - 60': 74, '60 - 70': 36, '70+': 18, 'under 30': 107}	{'Female': 159, 'Male': 245, 'Unknown': 15}	{'African-American': 30, 'Asian': 11, 'Caucasian': 352, 'Hispani
{'30 - 40': 98, '40 - 50': 69, '50 - 60': 83, '60 - 70': 64, '70+': 23, 'Unknown': 11, 'under 30': 116}	{'Female': 185, 'Male': 259, 'Unknown': 20}	{'African-American': 28, 'Caucasian': 395, 'Hispanic': 20, 'Unc
{'30 - 40': 116, '40 - 50': 90, '50 - 60': 84, '60 - 70': 67, '70+': 19, 'Unknown': 13, 'under 30': 134}	{'Female': 189, 'Male': 312, 'Unknown': 22}	{'African-American': 35, 'Asian': 13, 'Caucasian': 447, 'Hispani
{'30 - 40': 114, '40 - 50': 135, '50 - 60': 169, '60 - 70': 103, '70+': 36, 'Unknown': 13, 'under 30': 102}	{'Female': 339, 'Male': 310, 'Unknown': 23}	{'African-American': 41, 'Asian': 11, 'Caucasian': 580, 'Hispani
{'30 - 40': 340, '40 - 50': 408, '50 - 60': 425, '60 - 70': 315, '70+': 91, 'Unknown': 37, 'under 30': 298}	{'Female': 1090, 'Male': 759, 'Unknown': 65}	{'African-American': 131, 'Asian': 32, 'Caucasian': 1614, 'Hispa
{'30 - 40': 135, '40 - 50': 148, '50 - 60': 137, '60 - 70': 106, '70+': 36, 'Unknown': 13, 'under 30': 196}	{'Female': 345, 'Male': 397, 'Unknown': 29}	{'African-American': 56, 'Asian': 20, 'Caucasian': 654, 'Hispani
{'30 - 40': 74, '40 - 50': 107, '50 - 60': 87, '60 - 70': 58, '70+': 23, 'Unknown': 12, 'under 30': 103}	{'Female': 194, 'Male': 250, 'Unknown': 20}	{'African-American': 45, 'Caucasian': 380, 'Hispanic': 13, 'Unc
{'30 - 40': 153, '40 - 50': 155, '50 - 60': 147, '60 - 70': 123, '70+': 21, 'Unknown': 16, 'under 30': 168}	{'Female': 350, 'Male': 404, 'Unknown': 29}	{'African-American': 75, 'Asian': 18, 'Caucasian': 652, 'Hispani

What Now?

a call to action

Calls to Action

- How would you, or are you, using these types of tools now?
- How can we leverage collective infrastructure to open access to more archives?
- What approaches to creating sustainable technologies like this are working?

General Discussion

General Discussion (with audience)

- What else exists in this space that we can support & work with?
- What are people working on at the moment?
- What needs do you, in particular, have?
- What are you hoping to build? How can we help you build it?

Thank You