

# Collecting Social Media Data

Slides: [bit.ly/social-media-data-workshop-2019](https://bit.ly/social-media-data-workshop-2019)

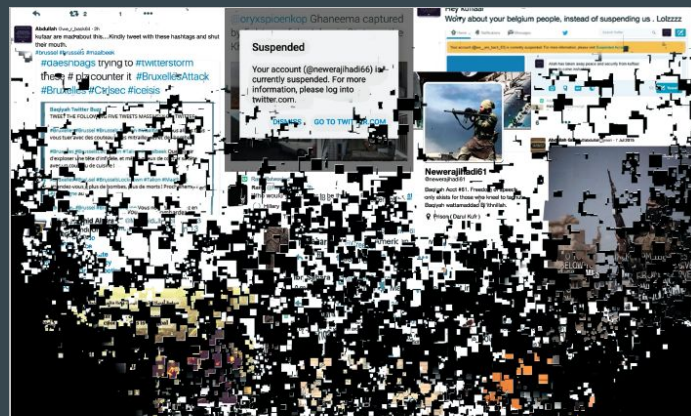
Laura Wrubel  
lwrubel@gwu.edu

Optional hands-on part of the workshop requires logging into Twitter. Either go to [twitter.com](https://twitter.com) to create a Twitter account (if you don't have one), or look on with someone else.

# Agenda

- Overview of social media APIs and data formats
- Twitter's API in depth
- Collecting new datasets
  - Hands-on: Social Feed Manager
- Using existing datasets
  - Hands-on: TweetSets
- Ethics of social media collecting

# Social media research



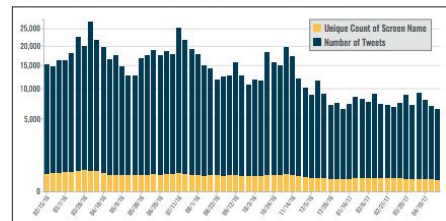
**DIGITAL DECAY?**

Tracing Change Over Time Among English-Language Islamic State Sympathizers on Twitter

Audrey Alexander  
October 2017

Program on Extremism  
THE GEORGE WASHINGTON UNIVERSITY

## Tweet Frequency and Unique Screen Names By Week



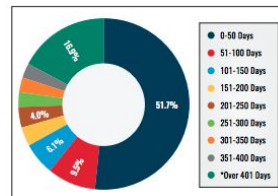
(Figure 5) This graph shows how the relationship between unique screen names and tweet frequency per week changed over the course of the 65-week period. As discussed in the method chapter, this graph, like several others in the study, uses square root in the y-axis to more clearly represent the relationship between the two variables.

of activity' is quantified by the number of days between an account's first and last tweet. Twitter's API does not discern the date or time at which the company suspends accounts, nor does it identify accounts that were created and then subsequently abandoned by their respective users. Consequently, this measurement allows the study to grasp the chronological span of sympathizers that actively use the platform to share content. While overwhelmingly skewed by outliers, the average lifespan for this sample of English-language pro-IS accounts on Twitter was 251 days. It is critical to note, however, that dispersion of lifespan is highly concentrated (see Figure 6). Approximately 51.7 percent of accounts did not remain active longer than 50 days. On the other hand, however, a substantive portion of accounts lasted over a year, suggesting that Twitter's attempts to detect and suspend pro-IS account may be missing some long-term users. One possible explanation for long-standing users relates to the data collection method, as researchers are more likely to identify accounts the longer they are open.<sup>13</sup> Ultimately, accounts that opted to leave the platform are likely included in this breakdown, although multiple factors—including the threat of suspension—likely affect user activity in this regard.

In order to maintain their presence on Twitter, some English language IS sympathizers appeared to have created multiple accounts at the same time to avoid shutdowns. On February 17, 2016, for example, four separate accounts were

fashioned from a core handle,<sup>14</sup> possibly from the same individual. One account (@erhabi35) survived only eight days, whereas another account (@Erhabi39) stayed active for 62 days. Although the study attempted to annotate cases where the same individual controlled multiple accounts, as the trend is common, quantitative figures are generally not reliable due to the relative anonymity Twitter affords users. It is hard to ascertain whether users that demonstrate similar behavioral patterns are simply individuals attempting to inoculate their digital presence against suspensions or are

## Duration of Account Activity



(Figure 6) This chart depicts the duration of account activity, meaning the number of days a pro-IS Twitter account was active, and displays the breakdown in percentages.

# Social media research



## Research Article

### Twitter Makes It Worse: Political Journalists, Gendered Echo Chambers, and the Amplification of Gender Bias

Nikki Usher<sup>1</sup>, Jesse Holcomb<sup>2</sup>, and Justin Littman<sup>3</sup>

#### Abstract

Given both the historical legacy and the contemporary awareness about gender inequity in journalism and politics as well as the increasing importance of Twitter in political communication, this article considers whether the platform makes some of the existing gender bias against women in political journalism even worse. Using a framework that characterizes journalists' Twitter behavior in terms of the dimensions of their peer-to-peer relationships and a comprehensive sample of permanently credentialed journalists for the U.S. Congress, substantial evidence of gender bias beyond existing inequities emerges. Most alarming is that male journalists amplify and engage male peers almost exclusively, while female journalists tend to engage most with each other. The significant support for claims of gender asymmetry as well as evidence of gender silos are findings that not only underscore the importance of further research but also suggest overarching consequences for the structure of contemporary political communication.

#### Keywords

political journalism, gender, Twitter, Washington journalism, beltway journalism, women in journalism

The International Journal of Press/Politics  
2018, Vol. 23(3) 324–344  
© The Author(s) 2018  
Reprints and permissions:  
sagepub.com/journalsPermissions.nav  
DOI: 10.1177/1040161218781254  
journals.sagepub.com/home/ijp

<sup>1</sup>University of Illinois at Urbana-Champaign, IL, USA

<sup>2</sup>Calvin College, Grand Rapids, MI, USA

<sup>3</sup>Georgetown University, Washington, DC, USA

INFORMATION, COMMUNICATION & SOCIETY, 2017  
Vol. 20, No. 9, 1330–1346  
<https://doi.org/10.1080/1369118X.2017.1328521>

Routledge  
Taylor & Francis Group



### Populist communication by digital means: presidential Twitter in Latin America

Silvio Waisbord<sup>a</sup> and Adriana Amado<sup>b</sup>

<sup>a</sup>School of Media and Public Affairs, George Washington University, Washington, DC, USA; <sup>b</sup>Universidad de La Matanza, San Justo, Argentina

#### ABSTRACT

In this paper, we analyze the uses of Twitter by populist presidents in contemporary Latin America in the context of the debates about whether populism truly represents a revolution in public communication – that is, overturning the traditional hierarchical model in favor of popular and participatory communication. In principle, Twitter makes it possible to promote the kind of interactive communication often praised in populist rhetoric. It offers a flattened communication structure in contrast to the top-down structure of the traditional legacy media. It is suitable for horizontal, unmediated exchanges between politicians and citizens. Our findings, however, suggest that Twitter does not signal profound changes in populist presidential communication. Rather, it represents the continuation of populism's top-down approach to public communication. Twitter has not been used to promote dialogue among presidents and publics or to shift conventional practices of presidential communication. Instead, Twitter has been used to reach out the public and the media without filters or questions. It has been incorporated into the presidential media apparatus as another platform to shape news agenda and public conversation. Rather than engaging with citizens to exchange views and listen to their ideas, populists have used Twitter to harass critical journalists, social media users and citizens. Just like legacy media, Twitter has been a megaphone for presidential attacks on the press and citizens. It has provided with a ready-made, always available platforms to lash out at critics, conduct personal battles, and get media attention.

#### ARTICLE HISTORY

Received 30 November 2016  
Accepted 4 May 2017

#### KEYWORDS

Social media; populism;  
presidential communication;  
political communication;  
Twitter

### Populism as communication style

Growing interest in the study of populism, media, and communication (Aalberg, Esser, Reinemann, Strömback, & de Vreese, 2017) inevitably confronts the long-standing fuzziness of the concept of populism. It is commonly acknowledged that 'populism' is perennially imprecise. Definitions have underscored different aspects as essential characteristics of populism, including economic policies, style of political leadership, political discourse, and ideology (Moffitt, 2016). Populism remains the subject of constant semantic squabbles, largely because it has taken various shapes across time and

# Targeting Persuadable Voters Through Social Media: The Use of Twitter in The 2015 UK General Election

[Open Access](#)

How do political campaigns target and persuade voters to support their candidates? Since 2000, US political campaigns have focused heavily on data analytics to micro target individual voters with personalized messages. Micro targeting moves away from the traditional assumption that voting behavior is determined purely by demographics. Instead, this method allows campaigns to predict accurately an individual's voting behavior and deliver to them the most appropriate message. This paper focuses on the use of social media by the Labour and Conservative campaigns in the 2015 UK General Election and whether it was employed as a targeting tool and a method to engage with targeted voters. More specifically, it examines the claim that Labour used social media purely to communicate with its core supporters whilst Conservatives used it effectively to target and engage with persuadable voters and this ultimately contributed to the Conservatives' victory.

Last modified:

Targeting Persuadable Voters Through Social Media:  
The Use of Twitter in The 2015 UK General Election

By Caitlin Roper

B.A. Joint Honors in History and Politics, July 2014, University of Sussex

A Thesis submitted to

The Faculty of  
The Columbian College of Arts and Sciences  
of The George Washington University  
in partial fulfillment of the requirements  
for the degree of Master of Arts

May 15, 2016

Thesis directed by

David Karpf  
Assistant Professor of Media and Public Affairs

## Relationships

In [ETDs](#)  
**Administrative  
Set:**

## Descriptions

Attribute Name	Values
Author	<a href="#">Roper, Caitlin Grace</a>
Language	<a href="#">en</a>

[Download PDF](#)



## Elites and foreign actors among the alt-right: The Gab social media platform by Yuchen Zhou, Mark Dredze, David A. Broniatowski, and William D. Adler

### Abstract

Content regulation and censorship of social media platforms is increasingly discussed by governments and the platforms themselves. To date, there has been little data-driven analysis of the effects of regulated content deemed inappropriate on online user behavior. We therefore compared Twitter — a popular social media platform that occasionally removes content in violation of its Terms of Service — to Gab — a platform that markets itself as completely unregulated. Launched in mid-2016, Gab is, in practice, dominated by individuals who associate with the “alt-right” political movement in the United States. Despite its billing as “The Free Speech Social Network,” Gab users display more extreme social hierarchy and elitism when compared to Twitter. Although the framing of the site welcomes all people, Gab users’ content is more homogeneous, preferentially sharing material from sites traditionally associated with the extremes of American political discourse, especially the far right. Furthermore, many of these sites are associated with state-sponsored propaganda from foreign governments. Finally, we discovered a significant presence of German language posts on Gab, with several topics focusing on German domestic politics, yet sharing significant amounts of content from U.S. and Russian sources. These results indicate possible emergent linkages between domestic politics in European and American far right political movements. Implications for regulation of social media platforms are discussed.

### Contents

# Social media on the web

The image shows a screenshot of a Twitter profile for Tim Kaine (@timkaine). The profile picture is a circular headshot of Tim Kaine speaking into a microphone. The background of the header is a scenic view of a forested mountain. The profile statistics are: 9,626 Tweets, 777 Following, 1M Followers, 723 Likes, and 2 Lists. The bio states: "Husband to @AnneHolton, father of 3. U.S. Senator from Virginia. In my free time, I'm either outdoors, reading, or jamming on the harmonica." The location is "Richmond, VA" and the website is "timkaine.com". He joined in July 2010. Below the bio is a red button that says "Tweet to Tim Kaine". A section titled "37 Followers you know" shows a grid of 12 circular profile pictures. Below that is a section titled "1,760 Photos and videos" showing a grid of image thumbnails. The first tweet is from Tim Kaine, posted 24 hours ago, with the text: "Mass shootings in this country happen far too often and are far too deadly—but Congress hasn't acted." Below the tweet is a large graphic with the text "If we want to change our gun laws" and "We need to change Congress", with the NR8 logo in the bottom right corner. The second tweet is from Tim Kaine, posted on Oct 10, with the text: "This is so false. Today the Senate is having to vote to overturn President Trump's expansion of insurance plans that don't protect people with pre-existing conditions."

**Tim Kaine** ✓  
@timkaine

Husband to @AnneHolton, father of 3. U.S. Senator from Virginia. In my free time, I'm either outdoors, reading, or jamming on the harmonica.

Richmond, VA

timkaine.com

Joined July 2010

**Tweet to Tim Kaine**

37 Followers you know

1,760 Photos and videos

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

**Tim Kaine** ✓ @timkaine · 24h  
Mass shootings in this country happen far too often and are far too deadly—but Congress hasn't acted.

So today I'm joining @ChrisMurphyCT and countless other leaders and advocates to raise \$1 million for candidates who will step up. Will you step up too? [secure.timkaine.com/NR8](https://secure.timkaine.com/NR8)

**If we want to change our gun laws**

**We need to change Congress**

**NR8**  
EIGHT DOLLARS WE CAN TAKE BACK FROM THE GUN LOBBY

67   182   579

**Tim Kaine** ✓ @timkaine · Oct 10  
This is so false. Today the Senate is having to vote to overturn President Trump's expansion of insurance plans that don't protect people with pre-existing conditions.

# Social media as data

id	created_at	user_screen_name	text	tweet_type	hashtags	media	urls	favorite_count
104222734266620929	Wed Sep 19 01:42:16 +0000 2018	timkaine	To all observing Yom Kippur in Virginia and around the world — I want to wish you a meaningful day of reflection and an easy fast.	original				774
1042170182377111557	Tue Sep 18 21:55:08 +0000 2018	timkaine	The FBI background investigation into Judge Kavanaugh should be reopened in light of the serious charges against him.	original				8565
1041874309004881920	Tue Sep 18 02:19:27 +0000 2018	timkaine	99-1. That was the final vote of the Opioid Crisis Response Act tonight in the Senate. Because we worked together, we've made progress toward preventing tens of thousands of deaths from this horrible epidemic each year. <a href="https://t.co/EYf65k6FFS">https://t.co/EYf65k6FFS</a>	original			<a href="https://www.foxnews.com/politics/kavanaugh-act">https://ww</a>	1599
1041818089510371328	Mon Sep 17 22:36:03 +0000 2018	timkaine	RT @GovernorVA: Please take precautions and stay tuned to local news alerts—a tornado watch is still in effect for many parts of the Common...	retweet				0
1041527946907987968	Mon Sep 17 03:23:07 +0000 2018	timkaine	RT @SarahPeckVA: Tim Kaine comments on the courage of Dr. Ford for speaking out and calls on Senate Judiciary to delay the vote on Kavanaugh...	retweet				0
1041504668659212288	Mon Sep 17 01:50:37 +0000 2018	timkaine	Judge Kavanaugh. The Judiciary Committee should not vote on his nomination until this allegation is fully investigated.	original				18185
1040422041588064257	Fri Sep 14 02:08:39 +0000 2018	timkaine	RT @MarkWarner: Hurricane Florence is likely to bring heavy rain to the Roanoke Valley and Southwest Virginia over the coming weekend. The...	retweet				0
1040291669302755328	Thu Sep 13 17:30:36 +0000 2018	timkaine	In 2012, @TyroneGayle was one of my closest campaign aides. We drove all over Virginia together. A former Clemson track star, he hasn't run much since his cancer diagnosis. Now his friends run for him and have set up this cool fundraiser. Give if you can! <a href="https://t.co/mKZ1PvaOuL">https://t.co/mKZ1PvaOuL</a>	original			<a href="https://www.facebook.com/tyronegayle">https://ww</a>	728
1039594806912139264	Tue Sep 11 19:21:31 +0000 2018	timkaine	🔴 VIRGINIANS: Please take all precautions as Hurricane Florence approaches. FOLLOW: @VDEM for emergency updates. VISIT: <a href="https://t.co/v6lYVxaj9v">https://t.co/v6lYVxaj9v</a> for more information. A federal emergency declaration has been made, and efforts are underway to prepare for this dangerous storm.	original			<a href="http://VAemergency.com">http://VAe</a>	380
1039569177793708039	Tue Sep 11 17:39:40 +0000 2018	timkaine	Today we laid a wreath in Arlington with brave law enforcement officers and first responders in memory of those we lost 17 yrs ago on 9/11. With each passing year, it becomes more important to commemorate these lives and that day so future generations #NeverForget what happened. <a href="https://t.co/sV3Hueu5Dp">https://t.co/sV3Hueu5Dp</a>	original	NeverForget		<a href="https://twitter.com/timkaine">https://twitter.com/ti</a>	1386
1039210775087329285	Mon Sep 10 17:55:31 +0000 2018	timkaine	RT @GovernorVA: It's important to prepare your family, home and business before a storm arrives. Visit <a href="https://t.co/5iKSQcE0wc">https://t.co/5iKSQcE0wc</a> and make sur...	retweet			<a href="http://www.gva.com">http://www</a>	0
1038959106621693952	Mon Sep 10 01:15:28 +0000 2018	timkaine	Roger Stone just coined one of the best band names ever - "Playing bluegrass with reckless abandon, please welcome The Insubordinate Hillbillies!" <a href="https://t.co/tSeAEWP9gl">https://t.co/tSeAEWP9gl</a>	original			<a href="https://www.facebook.com/rogerstone">https://ww</a>	2457
1038887381917728768	Sun Sep 09 20:30:28 +0000 2018	timkaine	To all those in Virginia and around the world celebrating the new year tonight, L'Shanah Tovah! I hope your year ahead is filled with peace, health, joy, and light.	original				1436
1038485507741765637	Sat Sep 08 17:53:33 +0000 2018	timkaine	RT @ElectConnolly: Fired up group of volunteers ready to knock doors in Vienna for me @JenniferWexton and @timkaine. Thanks for going out...	retweet				0
1038485417637180340	Sat Sep 08 17:53:12 +0000 2018	timkaine	our Weekend of Action for #K...	retweet				0



# Tweets are data, too.

- Example tweet: [go.gwu.edu/emse4197sampletweet](https://go.gwu.edu/emse4197sampletweet)
- [Twitter's guide to the structure of a tweet](#)

# JSON: JavaScript Object Notation

- `{ key: value, key: value... }`
- keys are strings
- a value may be:
  - string - in quotes: `"GW"`
  - number
  - boolean - `true` or `false`
  - another JSON object
  - array (denoted by square brackets `[ ]`) of JSON objects
  - `null`

# JSON example

```
{
  "text": "Yesterday, #GWU students, faculty,
staff...https://t.co/8Tz29odc11",
  "favorite_count": 56,
  "truncated": false,
  "entities": {
    "user_mentions": [],
    "hashtags": {
      "indices": [11, 15],
      "text": "GWU"
    }
  }
}
```

# Social Media APIs



# What's an API?

Short for “Application Programming Interface”

Allows you to request or send data to another service on the web, using HTTP.

Request: `http://an.api.com/query?term=pizza`

Response: structured data (XML, JSON)

# Why use an API for working with social media?

- You don't want to scrape it from the web page!
- An API gives you:
  - Data you can't get by scraping.
  - Data in a structured format, easier for analyzing.

# The Twitter API

## Scale your Twitter data access



### Standard APIs

Our free, standard APIs are great for getting started, testing an integration, validating a concept, or creating solutions that complement what you can create with premium and enterprise products. Examples include posting content to Twitter and getting data not available in high volumes.

### Premium APIs

Our premium APIs offer scalable access to Twitter data for those looking to grow, experiment, and innovate. When the standard API doesn't offer the amount of data necessary, upgrading to premium allows you to continue building and growing. Test in the free sandbox and then upgrade to month-to-month access.

### Enterprise APIs

Our enterprise APIs offer the highest level of access and reliability to those who depend on Twitter data. Perfect as you scale beyond premium and need more reliable access, custom tailored packages, or annual contracts. Enterprise API access comes with dedicated account managers and technical support.



# Understanding the Twitter API

- There are many Twitter APIs, only some are free.
- Their restrictions and affordances shape what you can collect.
- Understanding the APIs allows you to best choose which research questions can be addressed.

# Most useful API methods for collecting tweets

- **User timeline:** GET statuses/user\_timeline
  - Up to the most recent 3,200 tweets
- **Search:** GET search/tweets
  - Sampling of tweets from last 7 days.
  - Query by keyword, phrases, hashtags, author, date, more.
  - Not the same as search via twitter.com
- **Filter stream:** POST statuses/filter
  - Filter by keyword, user, or location

# User timeline: GET statuses/user\_timeline

- Gets most recent tweets posted by a user.
- Limited to last 3,200 tweets.
- Returns 200 at a time, so must page.
- Rate limit: 900 tweets per 15 minutes
- `https://api.twitter.com/1.1/statuses/user_timeline.json?screen_name=gelmanlibrary&max_id=8298861563345715`

# Search: GET search/tweets

- Search recent tweets.
  - Sampling of tweets from last 7 days.
  - Query by keyword, phrases, hashtags, author, date, more.
- Returns up to 100 at a time, so must page.
- Not the same as search on Twitter website.
- Rate limit: 180 tweets per 15 minutes
- `https://api.twitter.com/1.1/search/tweets.json?q=%23onlyatgw`



# Filter Stream: POST statuses/filter

- Real-time filtering of all public tweets.
  - Filter by keyword, user, or location.
- Continue to receive additional tweets over a single call to API. (No paging.)
- Limits:
  - When high volume, will not receive all tweets.
  - One stream at a time per set of credentials.
- `https://stream.twitter.com/1.1/statuses/filter.json?track=gwu`

# More Twitter API methods

Get a specific tweet: GET users/lookup

Get user info: GET users/lookup

Get trends near a location: GET trends/place

More: [developer.twitter.com/en/docs](https://developer.twitter.com/en/docs)

# Acquiring Twitter data sets

# Options for acquiring a Twitter dataset

- Collect a new dataset.
- Use an existing dataset.
- Purchase data or access to a platform.

# Collecting a new dataset - using coding

Command line:

- Twarc: [github.com/docnow/twarc](https://github.com/docnow/twarc)
- Twurl: [github.com/twitter/twurl](https://github.com/twitter/twurl)

Python libraries

- twarc [github.com/DocNow/twarc](https://github.com/DocNow/twarc)
- tweepy: [www.tweepy.org](https://www.tweepy.org)

R package: rtweet: [github.com/mkearney/rtweet](https://github.com/mkearney/rtweet)

# Collecting a new dataset - no coding required

Social Feed Manager:

[go.gwu.edu/sfmgw](https://go.gwu.edu/sfmgw)

TAGS (Twitter Archiving Google Sheet):

[tags.hawksey.info](https://tags.hawksey.info)

# Collecting new Twitter data

# Social Feed Manager software

- Open source software by GW Libraries.
- User interface for collecting, managing, and exporting social media data.
- Collect from Twitter, Tumblr, Flickr, Sina Weibo.
- Libraries run this for their users as a service.  
(Not typically a local install on your laptop.)

More: [go.gwu.edu/sfm](https://go.gwu.edu/sfm)



# Hands-on: Social Feed Manager

Steps we'll perform:

1. Sign up
2. Request credentials (API keys)
3. Create a collection
4. Perform a harvest
5. Export data

Go to: [gwsfm-sandbox.wrlc.org](https://gwsfm-sandbox.wrlc.org)

# Exporting datasets

- Formats: Excel, CSV, JSON
- Limit by date ranges
- Splits into separate files

**Using existing Twitter data**

# Datasets from other researchers

Twitter's terms generally do not allow datasets of full JSON data to be shared.

OK to share: Text file of tweet identifiers

```
id_str: "775347040196894720"
```

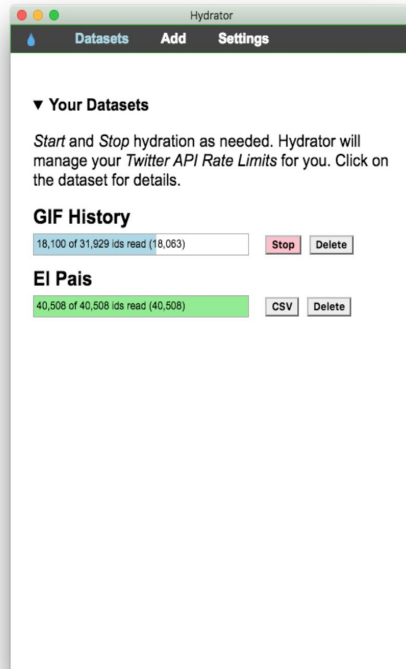
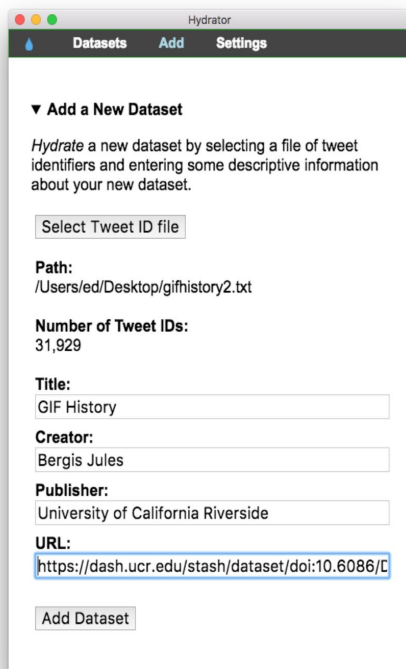
Use Twitter API to request tweets by identifier and get back the full tweet.

Won't include deleted/protected tweets.

# Working with tweet identifiers

## Hydrator desktop app

<https://github.com/DocNow/hydrator>



# Using an existing dataset

- DocNow Catalog: [www.docnow.io/catalog/](http://www.docnow.io/catalog/)
  - Data repositories such as [Dataverse](#)
  - TweetSets: [tweetsets.library.gwu.edu/](http://tweetsets.library.gwu.edu/)
    - Datasets collected by GW Libraries.
    - Full tweets available as JSON or CSV
- Only for GW users, for academic purposes only.

# Datasets collected by GW Libraries

- 2020 U.S. Presidential election (in progress)
  - 2018 U.S. midterm election
  - 2016 U.S. election (280 million tweets)
  - Congress (all senators and representatives)
  - Federal govt (3000 U.S. government accounts)
  - News outlets (4500 media organization accounts)
  - Hurricanes
  - Make America Great Again
  - Tax reform
  - Immigration & travel ban
  - Climate change
- More ...

# TweetSets

Steps we'll demo:

1. Select a source dataset.
2. Filter the source dataset.
3. Create a new dataset.
4. Generate and download dataset derivatives.

[tweetsets.library.gwu.edu/](https://tweetsets.library.gwu.edu/)



**Purchasing data or access to a platform**

# Options

- Subscribe to an analytics platform such as CrimsonHexagon. \*Can only download 50,000 tweets/day
- Subscribe to Twitter Premium or Enterprise APIs.
- Purchase historical batch data from Twitter.
- Subscribe to historical search API access from Twitter.

# Can I get Tweets from the past without cost?

- If GW collected it already: yes (TweetSets or SFM)
- If someone else collected it:
  - Need to hydrate tweet IDs , won't be complete.
- Using Twitter collections in SFM:
  - User timeline: up to ~3,200 tweets per account
  - Search: ~7 days
  - Filter: No.

## FAQ: Are tweets geotagged?

- Geotagging is opt-in. Only ~2% geotagged.  
Lat, long or place name (e.g., DC or Middle Earth)
- Search API: Limit to a specified distance of a lat, long.
- Filter Stream: Limit to a bounding box.

More: [gwu-libraries.github.io/sfm-ui/posts/2017-04-12-geographic-collecting](https://gwu-libraries.github.io/sfm-ui/posts/2017-04-12-geographic-collecting)

# Exploring and analyzing Twitter data

# Before analysis

Clean and validate your data.

- Are the terms you queried used for other meanings and events?
- Are the accounts valid?
- Are there gaps in the data?

# Working with datasets

- Jupyter notebooks for Python and pandas analysis: [bit.ly/2uhN252](https://bit.ly/2uhN252) also see [here](#)

- R

- jq command-line tool

“Recipes for Twitter data” [bit.ly/2t9cStF](https://bit.ly/2t9cStF)

- Excel or Google Sheets

**Other social media platforms**



Other platforms?

Facebook: little/no API available

Instagram: no API available



## The Forum

### Computational Research in the Post-API Age

DEEN FREELON

**Keywords** API, computational, Facebook, Twitter, social media

On April 4, 2018, the post-API age reached a milestone. On that day, Facebook closed access to its Pages API, which had allowed researchers to extract posts, comments, and associated metadata from public Facebook pages (Schroepfer, 2018). This decision followed the company's April 2015 closure of its public search Application Programming Interface (API), which provided searchable access to all public posts within a rolling two-week window (Facebook, n.d.). The closure of the Pages API eliminated all terms of service (TOS)-compliant access to Facebook content. Let me underscore the magnitude of this shift: There is currently no way to independently extract content from Facebook without violating its TOS.

At the flip of a metaphorical switch, Facebook instantly invalidated all methods that depended on the Pages API. For example, I gave a Facebook data collection workshop in January 2018 at the University of Michigan whose lessons are now mostly unusable. A Python module I wrote to extract data from the Pages API is similarly obsolete. The specific implications for Facebook research are immense, but larger still are those for API-based research more generally. When companies can restrict or eliminate API access at any time, for any reason, and without any recourse, computational researchers and students need to seriously consider how to proceed. We find ourselves in a situation where heavy investment in teaching and learning platform-specific methods can be rendered useless overnight: This is what I mean by "the post-API age."

In this brief article I provide two guiding lights for graduate education in computational methods going forward. APIs will continue to be important sources of digital communication data, but the closure of the Pages API demonstrates the dangers of relying on them exclusively. Researchers of social and other online media content should start by doing two things as they brace themselves for the uncertainty ahead. First, they should learn how to scrape the Web; and second, they should understand the potential consequences of violating platforms' TOS by doing so.

Deen Freelon is an associate professor in the School of Media and Journalism at the University of North Carolina at Chapel Hill.

Address correspondence to Deen Freelon, UNC School of Media and Journalism, Carroll Hall, CB 3365, Chapel Hill, NC 27599. E-mail: [freelon@email.unc.edu](mailto:freelon@email.unc.edu)

What do you do when there's no API available?

Web scraping and capture tools are an alternative approach.

Deen Freelon (2018) [Computational Research in the Post-API Age](#), Political Communication, 35:4, 665-668. [Also available as preprint.](#)

# Webrecorder

[webrecorder.io](https://webrecorder.io)

- “Record” your web browsing and capture sites as viewed by a human.
- Provides a complementary view to API data.
- Sign in for 500GB account.

# Ethical considerations

# Social media data comes from people

- Consider impact of your work on the creator of the social media.
- Do not have creator's permission for research.
- Impact on creator is balanced against public good of your research.
- Requires judgment call.

More: [go.gwu.edu/sfmethics](https://go.gwu.edu/sfmethics)

“Participant Perceptions of Twitter Research Ethics.” Casey Fiesler, Nicholas Proferes, *Social Media + Society*.

First published March 10, 2018. [doi.org/10.1177/2056305118763366](https://doi.org/10.1177/2056305118763366)

# Data collecting

Be thoughtful collecting social media of:

- Vulnerable individuals (e.g., minors, social activists)
- Sensitive or harmful topics (e.g., questionable behavior, mental illness)
- Geography-based collecting

# Data sharing

- Get familiar with platform terms of use.
  - Don't republish full datasets
  - Share in accordance with terms (e.g., tweet ids only)
  - Consider copyright
- Sharing summary statistics is usually OK.

# Publishing

- When possible, get permission from creator for quotes.
- Do not rely on anonymizing posts.



# Questions?

Make a consultation appointment:

[calendly.com/social-media-consulting-gw](https://calendly.com/social-media-consulting-gw)

Social Feed Manager team:

[sfm@gwu.edu](mailto:sfm@gwu.edu)

Laura Wrubel

[lwrubel@gwu.edu](mailto:lwrubel@gwu.edu)