Acquiring Social Media Data

Laura Wrubel and Dan Kerchner
January 24, 2018

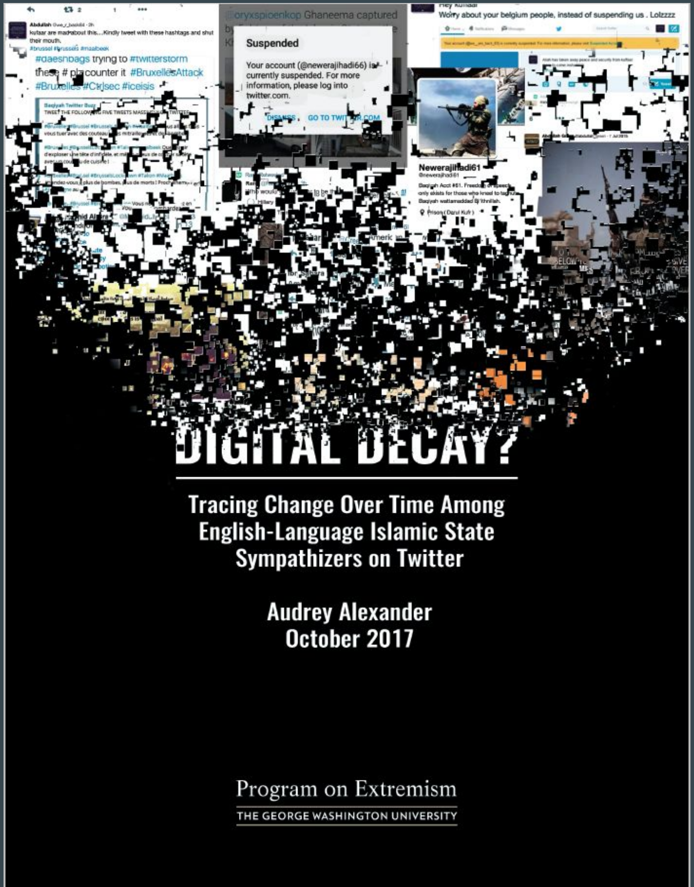
Slides: bit.ly/social-media-data-workshop-2019

The hands-on part of the workshop requires logging into Twitter. Either go to 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
- Using existing datasets
    - Hands-on: TweetSets
- Collecting new datasets
    - Hands-on: Social Feed Manager
- Ethics of social media collecting

Social media research

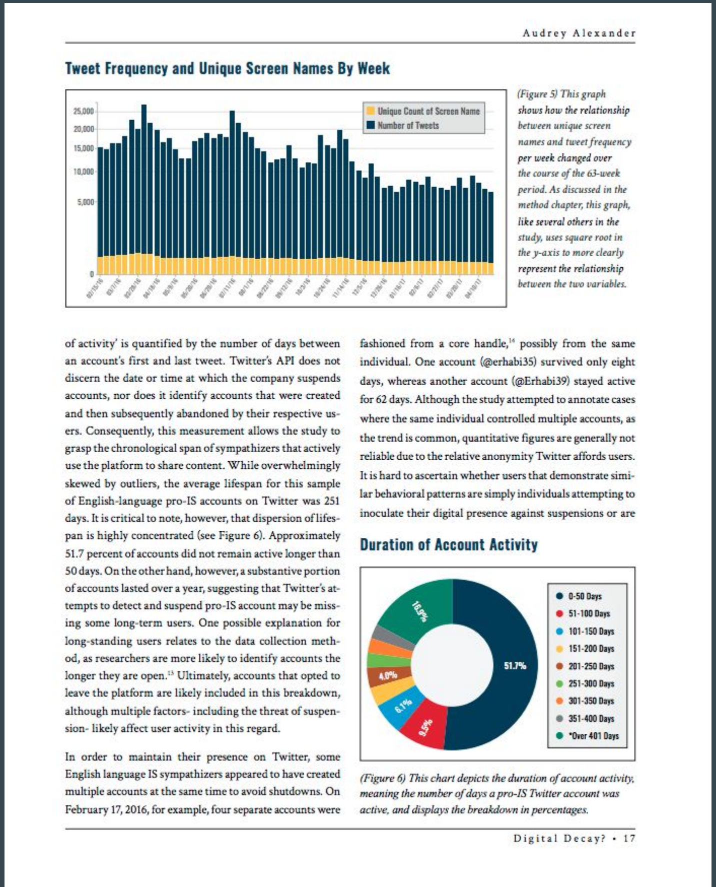

Suspended
Your account (@newerajihadi66) is currently suspended. For more information, please log into twitter.com.
DISMISS
GO TO TWITTER.COM
#daeshbags trying to #twitterstorm these # plz counter it #BruxellesAttack #Bruxelles #Ctrlsec #iceisis
Worry about your belgium people, instead of suspending us . Lolzzzz
Newerajihadi61

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 63-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.

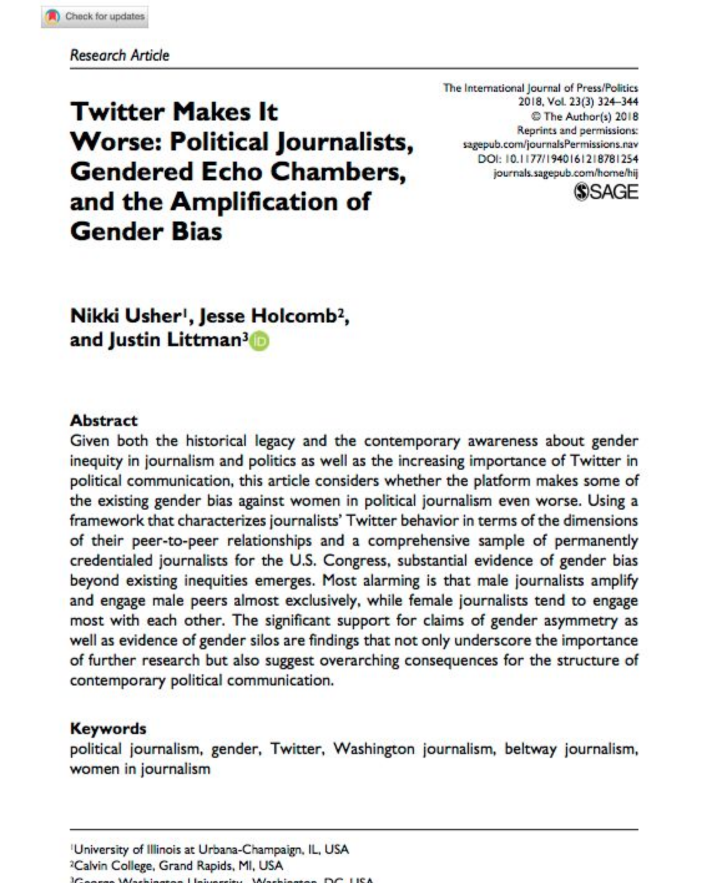
of activity' is quantified by the number of days between an account's first and last tweet. Twitter's API does not discern the date or time at which the company suspends accounts, nor does it identify accounts that were created and then subsequently abandoned by their respective users. Consequently, 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.[13] 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,[14] 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



Check for updates

Research Article

The International Journal of Press/Politics
2018, Vol. 23(3) 324–344
© The Author(s) 2018
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1940161218781254
journals.sagepub.com/home/hij
SAGE

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

Nikki Usher[1], Jesse Holcomb[2], and Justin Littman[3]

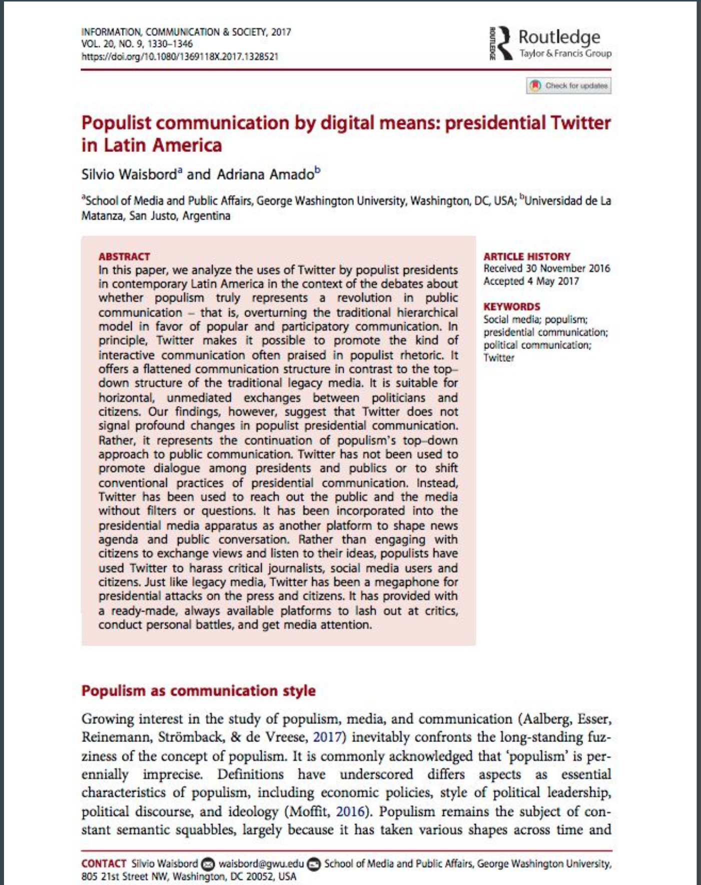
Abstract

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

[1]University of Illinois at Urbana-Champaign, IL, USA
[2]Calvin College, Grand Rapids, MI, USA
[3]George Washington 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

Check for updates

Populist communication by digital means: presidential Twitter in Latin America

Silvio Waisbord[a] and Adriana Amado[b]

[a]School of Media and Public Affairs, George Washington University, Washington, DC, USA; [b]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.

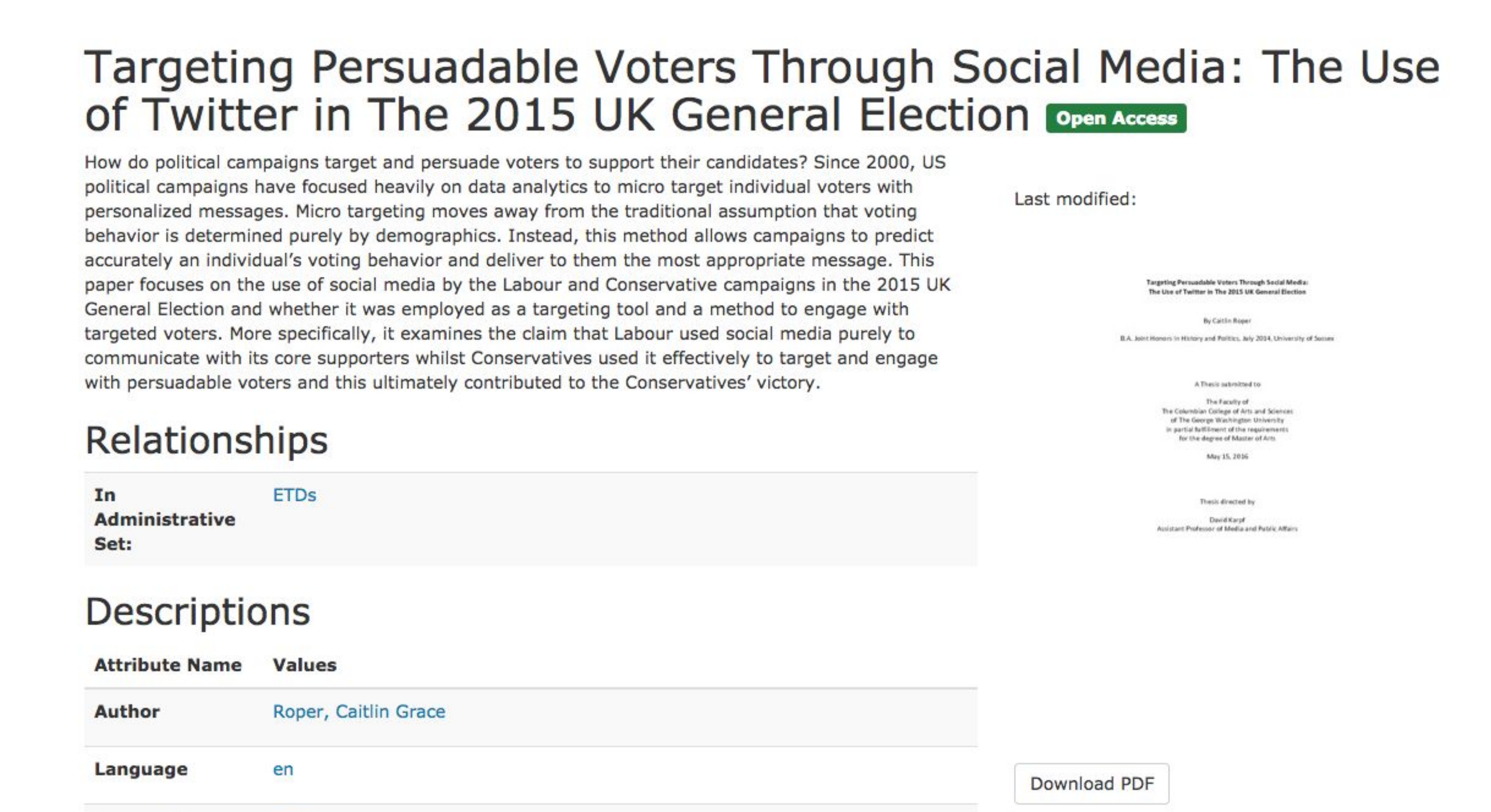
ARTICLE HISTORY
Received 30 November 2016
Accepted 4 May 2017

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 differs aspects as essential characteristics of populism, including economic policies, style of political leadership, political discourse, and ideology (Moffit, 2016). Populism remains the subject of constant semantic squabbles, largely because it has taken various shapes across time and

CONTACT Silvio Waisbord ✉ waisbord@gwu.edu School of Media and Public Affairs, George Washington University, 805 21st Street NW, Washington, DC 20052, USA



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

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

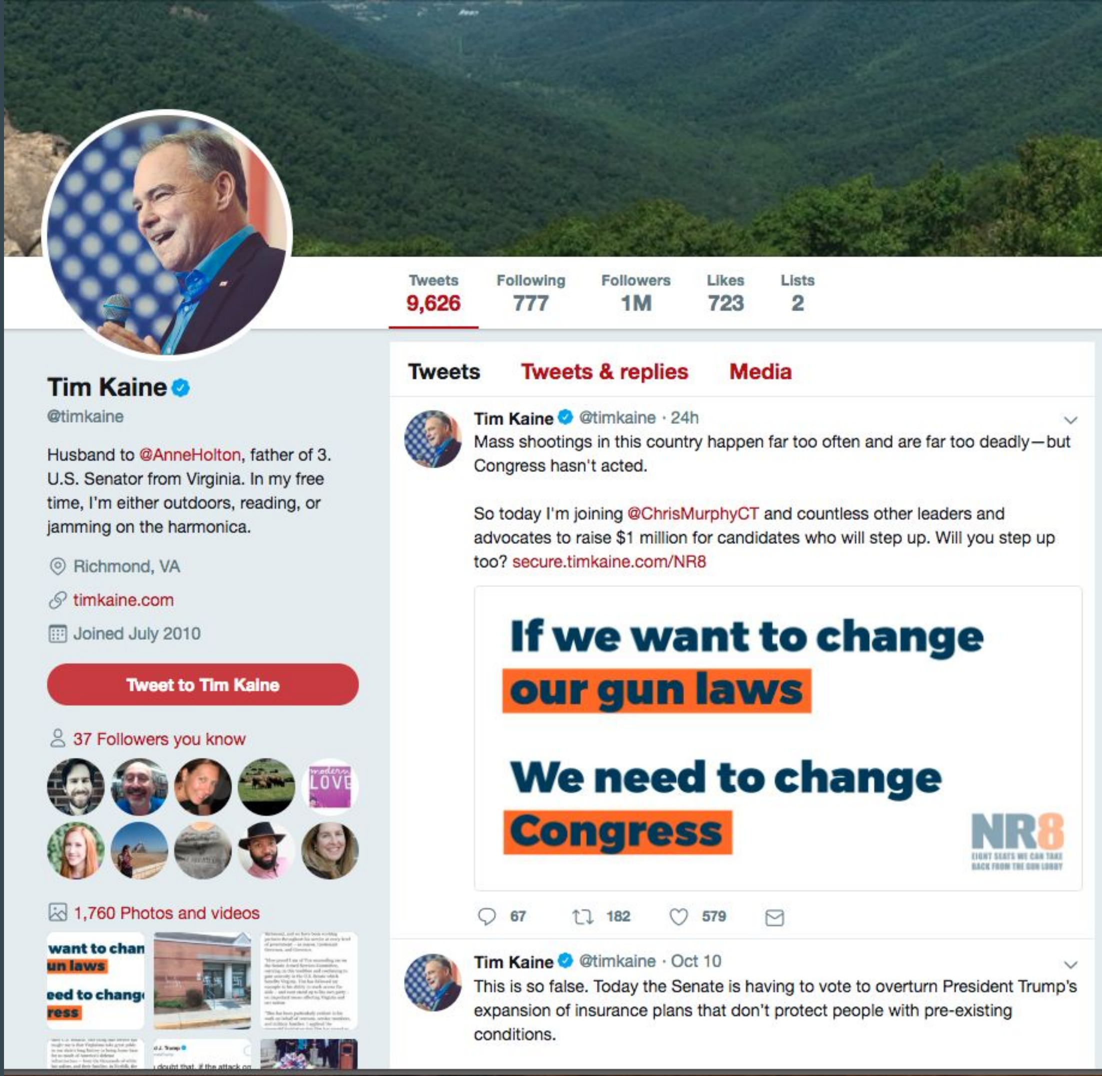
Relationships

In Administrative Set:	ETDs

Descriptions

Attribute Name	Values
Author	Roper, Caitlin Grace
Language	en

Social media on the web



Tweets 9,626 | Following 777 | Followers 1M | Likes 723 | Lists 2

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

If we want to change our gun laws

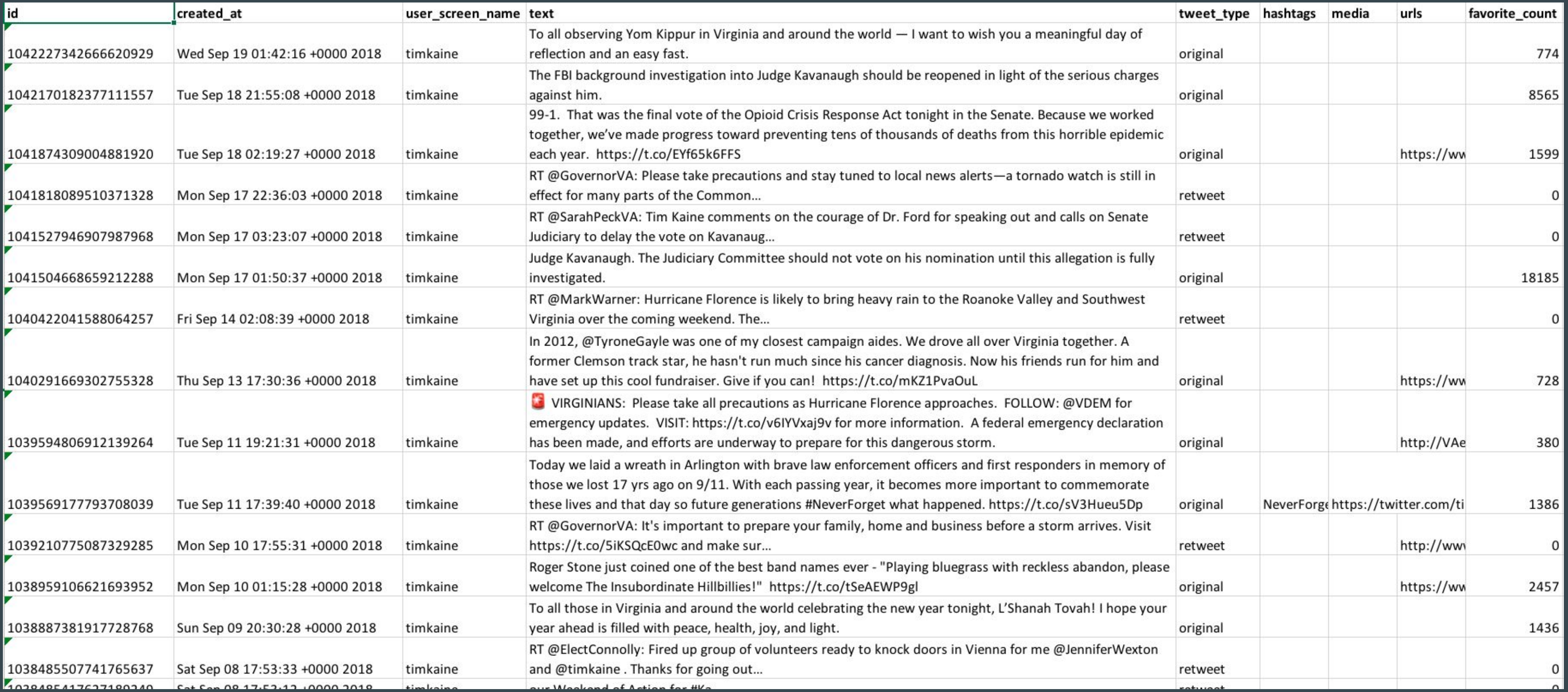
We need to change Congress

NR8
EIGHT SEATS 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
1042227342666620929	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. https://t.co/EYf65k6FFS	original			https://ww	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 Kavanaug…	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! https://t.co/mKZ1PvaOuL	original			https://ww	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: https://t.co/v6IYVxaj9v for more information. A federal emergency declaration has been made, and efforts are underway to prepare for this dangerous storm.	original			http://VAe	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. https://t.co/sV3Hueu5Dp	original	NeverForge	https://twitter.com/ti		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 https://t.co/5iKSQcE0wc and make sur…	retweet			http://ww	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!" https://t.co/tSeAEWP9gl	original			https://ww	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
1038485417627189240	Sat Sep 08 17:53:12 +0000 2018	timkaine	our Weekend of Action for #Ka	retweet				0

Tweets are data, too.

- Example tweet: go.gwu.edu/emse4197sampletweet
- Twitter's guide to the structure of a tweet: developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object

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"
     }
   }
}
```

APIs, social media APIs, and their data

What's an API?

- Short for “Application Programming Interface”
- Allows code to request or send data to a website
- API calls consist of:
  - **requests**: http://an.api.com/request?foo=15
  - **response:** structured data, e.g., XML or 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 similar to what the platform stores.
  - Slices of data you can't get by scraping.
  - Data in structured format, which makes it easy to analyze as data, with analysis tools.

The Twitter API

Understanding the Twitter API

- There are many Twitter APIs, only some 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=829886156334571 5`

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`

Some other Twitter API methods

- Get a specific tweet: GET users/lookup
- Post a tweet: POST statuses/update
- Follow a user: POST friendships/create
- Get user info: GET users/lookup
- Get trends near a location: GET trends/place

More: 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.

More:
gwu-libraries.github.io/sfm-ui/posts/2017-09-14-twitter-data

Collecting a new dataset - using coding

These require some coding skills:

- Command line:
  - Twarc: github.com/docnow/twarc
  - Twurl: github.com/twitter/twurl
- Libraries:
  - Python
    - twarc github.com/DocNow/twarc
    - tweepy: www.tweepy.org
  - R - rtweet: github.com/mkearney/rtweet

Collecting a new dataset - no coding required

Social Feed Manager: go.gwu.edu/sfmgw

TAGS (Twitter Archiving Google Sheet): 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

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

Exporting datasets

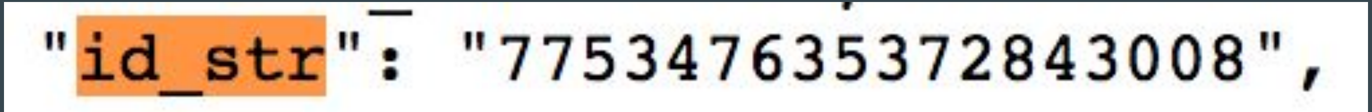
- Formats: Excel, CSV, JSON
- Limit by date ranges
- Splits into separate files

Using existing Twitter data

Datasets from other researchers

Twitter’s terms generally do not allow datasets of full tweet data to be shared.

OK to share: text file of tweet identifiers:



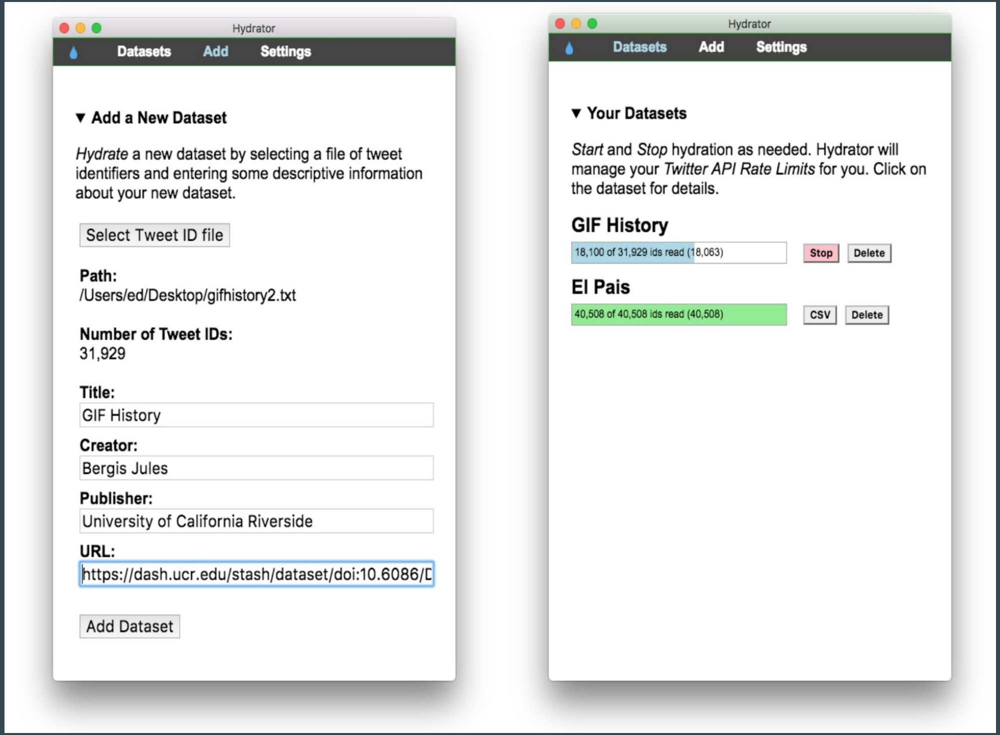
```
"id_str": "775347635372843008",
```

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



▼ Add a New Dataset

Hydrate a new dataset by selecting a file of tweet identifiers and entering some descriptive information about your new dataset.

Path:
/Users/ed/Desktop/gifhistory2.txt

Number of Tweet IDs:
31,929

Title:
GIF History

Creator:
Bergis Jules

Publisher:
University of California Riverside

URL:
https://dash.ucr.edu/stash/dataset/doi:10.6086/D

▼ Your Datasets

Start and *Stop* hydration as needed. Hydrator will manage your *Twitter API Rate Limits* for you. Click on the dataset for details.

GIF History

18,100 of 31,929 ids read (18,063)

El Pais

40,508 of 40,508 ids read (40,508)

Using an existing dataset

- DocNow Catalog: www.docnow.io/catalog/
  - Tweet ids only. Will need to hydrate.
- Data repositories such as Dataverse: dataverse.harvard.edu
- TweetSets: tweetsets.library.gwu.edu/
  - Filter datasets collected by GW Libraries.
  - Full tweets available as JSON or CSV (when on campus network, GW users only).

Datasets collected by GW Libraries

- 2016 U.S. election (280 million tweets)
- 2018 U.S. midterm election
- Congress (all senators and representatives)
- Federal govt (3000 U.S. government accounts)
- News outlets (4500 media organization accounts)

- Hurricane Florence/Harvey / Irma
- Trump Admin officials
- Make America Great Again
- Tax reform
- Immigration & travel ban
- Charlottesville
- Climate change

More ...

TweetSets

Steps we'll demo:

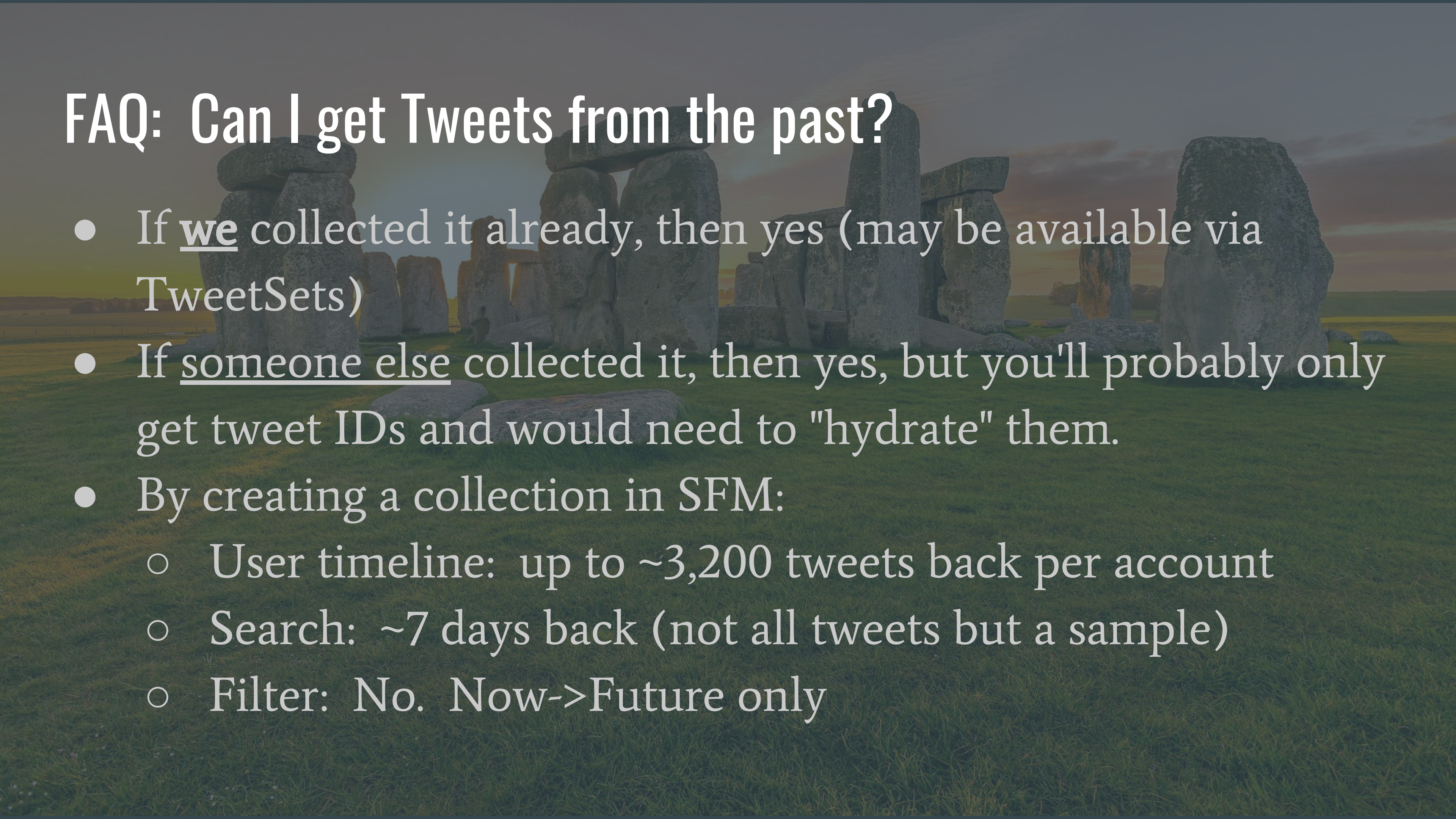
1. Select a source dataset.
2. Filter the source dataset.
3. Create a new dataset.
4. Generate and download dataset derivatives.

tweetsets.library.gwu.edu/

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.



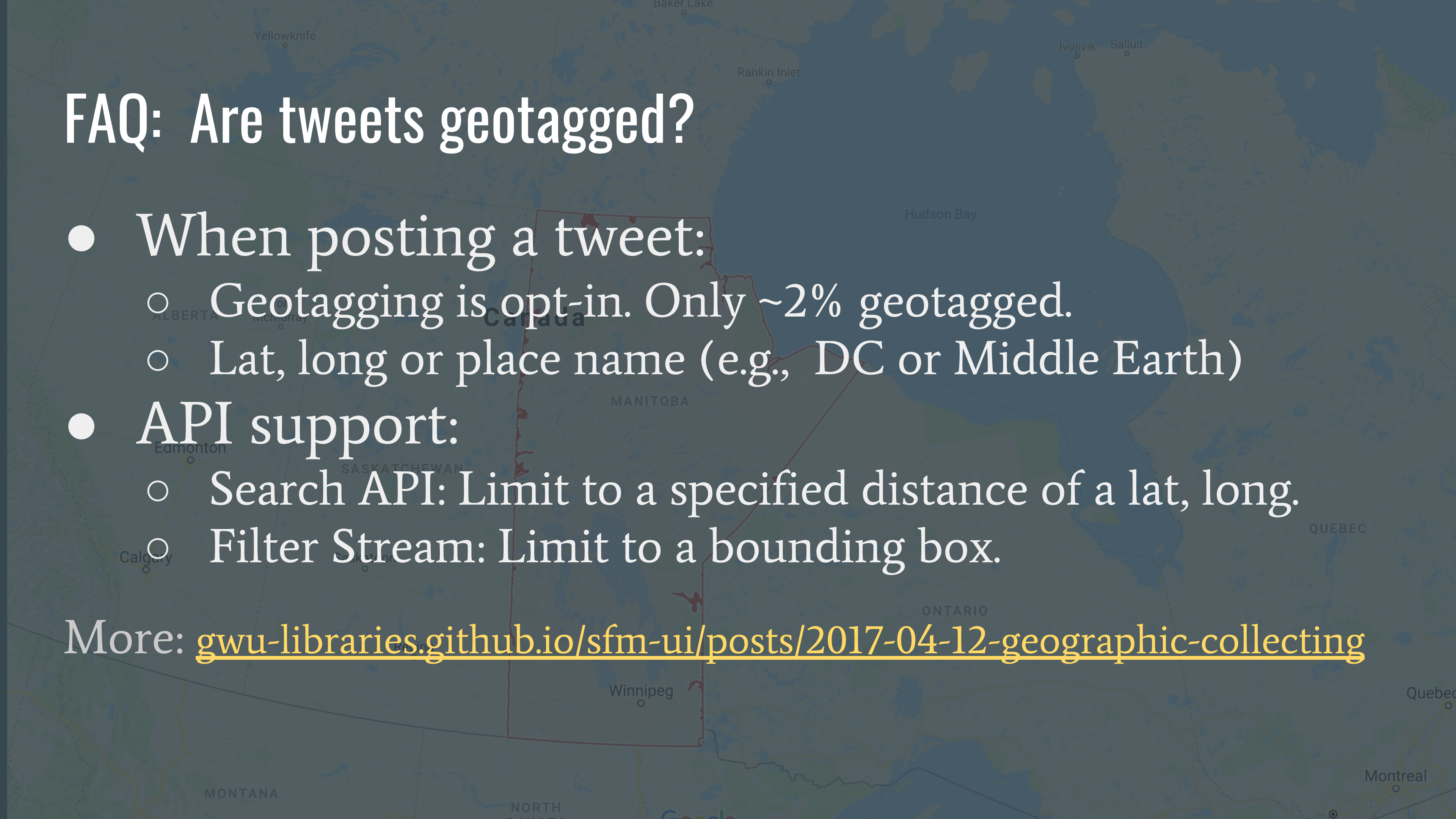
FAQ: Can I get Tweets from the past?

- If **we** collected it already, then yes (may be available via TweetSets)
- If someone else collected it, then yes, but you'll probably only get tweet IDs and would need to "hydrate" them.
- By creating a collection in SFM:
  - User timeline: up to ~3,200 tweets back per account
  - Search: ~7 days back (not all tweets but a sample)
  - Filter: No. Now->Future only

FAQ: Are tweets geotagged?

- When posting a tweet:
  - Geotagging is opt-in. Only ~2% geotagged.
  - Lat, long or place name (e.g., DC or Middle Earth)
- API support:
  - 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



Exploring and analyzing Twitter data

Before analysis

- Clean and validate your data
- Examples of why this is necessary:
    - Our 2016 U.S. election collection includes tweets from the Indian election.
    - Our U.S. government collection includes accounts that were deleted, claimed by other users, and tweeting in Russian.

Working with datasets

- Jupyter notebooks:
  - Python and pandas: bit.ly/2uhN252 (also see here)
- R
- jq command-line tool
  - Recipes for Twitter data: bit.ly/2t9cStF)
- Excel or Google Sheets

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

"'Participant Perceptions of Twitter Research Ethics." Casey Fiesler, Nicholas Proferes, *Social Media + Society.* First published March 10, 2018. 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

- sfm@gwu.edu
- @liblaura   lwrubel@gwu.edu
- @DanKerchner   kerchner@gwu.edu