



# Building Predictive Applications with Social-Media

Predictive Analytics and Applications (PAA) 2019

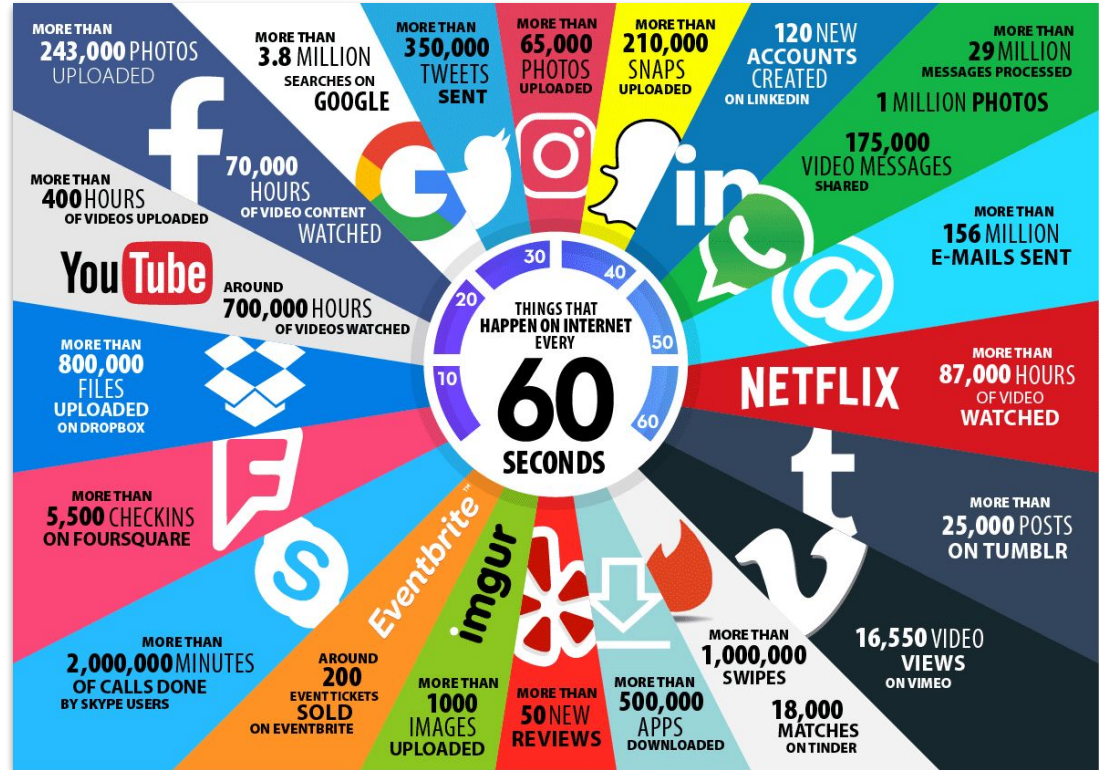


# Big-Data and Anonymity are at odds?

*Digital footprint* of an average office worker is **5GB** per day. As we connect more and more parts of our lives online, this digital footprint is only going to expand. These trails can serve as *primary documents* for our stories.



Most of this data is *invisible*, but how much of it is *impersonal*?



# What's in it for you, today?

1. Social-Media Data?
2. Data Gathering
  - a. APIs
  - b. Archives
  - c. Websites
3. Analysis Models
4. Data Caveats
5. Further Reading!

220 1:40 AM - Mar 28, 2017

106 people are talking about this

 **MotherPlaylist**  
@MotherPlaylist

Parenting is 99% getting roasted by your kids.

160 6:26 AM - Sep 1, 2017

63 people are talking about this

 **James Breakwell, Exploding Unicorn** ✓  
@XplodingUnicorn

My 2-year-old called the vehicle for sick people a "wee woo truck" and now I don't even remember what the right name is anymore.

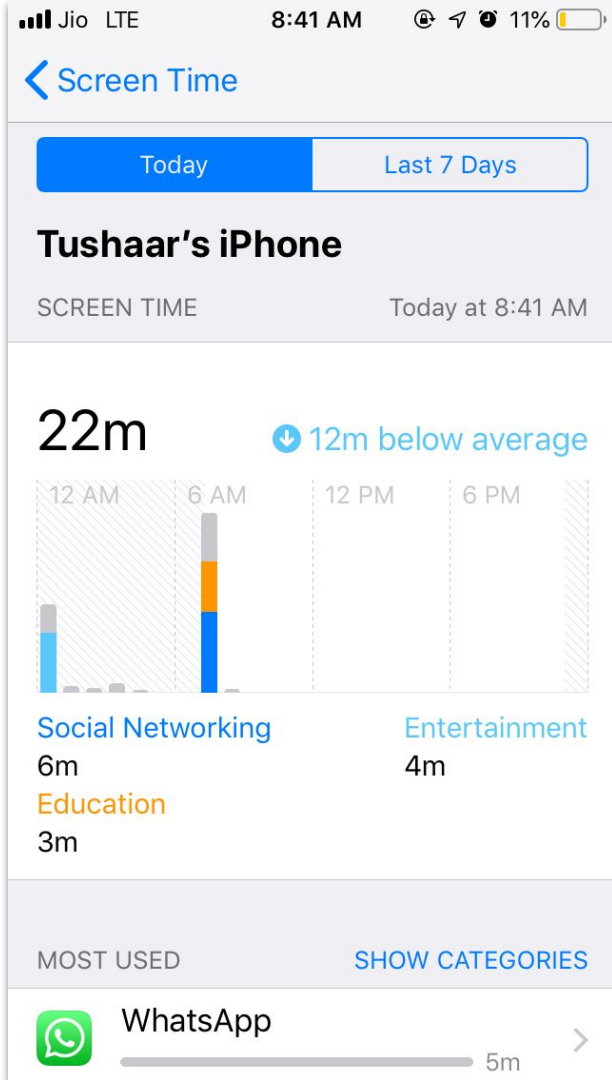
9,821 5:14 AM - May 18, 2017

1,757 people are talking about this

 **Bunmi Laditan**  
@HonestToddler

Receptionist at pediatrician's office: Child's birth date and year?

Me, mother of 3: Wow ok I didn't know there was going to be maths \*nervous laughter\* let's see he's four, it was late April or May, rainy I think, he's a classic Gemini if that helps, this isn't in his file?



# Why Social-Media Data?

Social-media is an effective platform of *self-expression, communication* and *social participation*! More often than not, the news of things around the world first appears on social-media!

As a result of the widespread use of smartphones today, audience spends 22% of their time on social-media\*!



How much do we *socialize*?

\* Facebook, YouTube and Wikipedia

“Sometimes, a handful of *social media* posts (tweets) can be at the heart of a story!”

### Damning new evidence that Dr Kelly DIDN'T commit suicide

The official explanation was that the distinguished weapons expert had taken his own life by overdosing on painkillers and cutting his left wrist.

dailymail.co.uk

Jan 12, 2019



**Daniel Preda**  @MisterPreda · Jan 11

Logan Paul joking about being gay “for one month” while countless LGBT+ around the world are killed & committing **suicide** for their sexuality, is disgusting. He continues to be an awful representation of the YouTube community & shows he truly has learned NOTHING over the last year

 2.4K  39K  162K 

[Show this thread](#)



**Charlotte Clymer**   @cmclymer · Jan 11

In 26 states, your employer can fire you for being LGBTQ.

LGBTQ children are as much as 7x more likely than their cisgender, heterosexual peers to die by **suicide**.

Nearly a third of LGBTQ children report being bullied on school grounds.

Cool prank, Logan Paul.

 794  11K  41K 



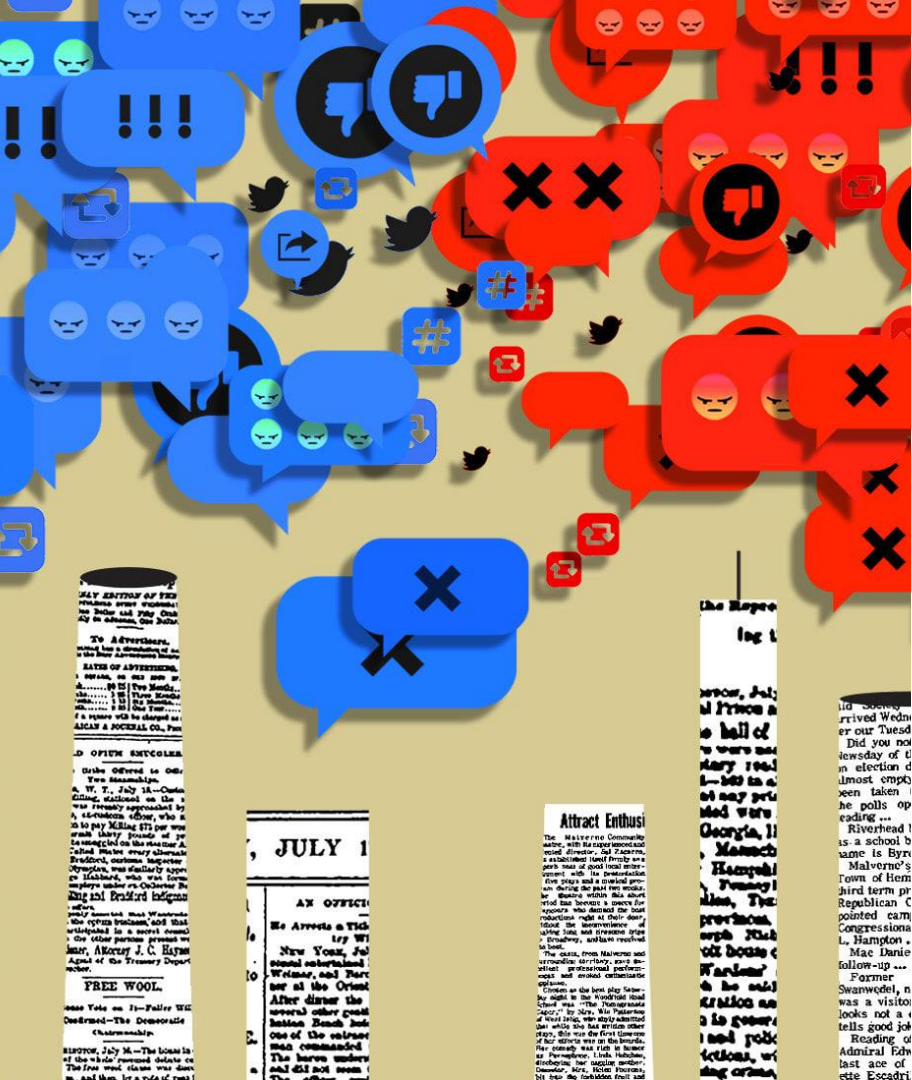
**Daniel Preda**  @MisterPreda · Jan 11

Perhaps for the month of March instead of baiting the gay community you can use your massive platform as a voice for those LGBT+ who have no voice and resort to **suicide** and self-harm. Just a suggestion @LoganPaul

 406  4.1K  30K 

[Show this thread](#)





# Psychology or Sociology?

Looking at social-media data in bulk can aid us in understanding larger issues!

## Use-Cases:

- ❖ *Analysis of devices used to tweet from @realDonaldTrump's profile!*
- ❖ *What kind of news did Trump tweet?*
- ❖ *What does it feel like to be trolled on social-media?*

What It Feels Like To Be Trolled

8 mentions

2 A.M. on Thursday

# Social Data Storytelling!

“What really shocked me was not knowing that the lines didn't represent coasts or rivers or political borders, but *real human social-media relationships!*”

—Paul Butler—



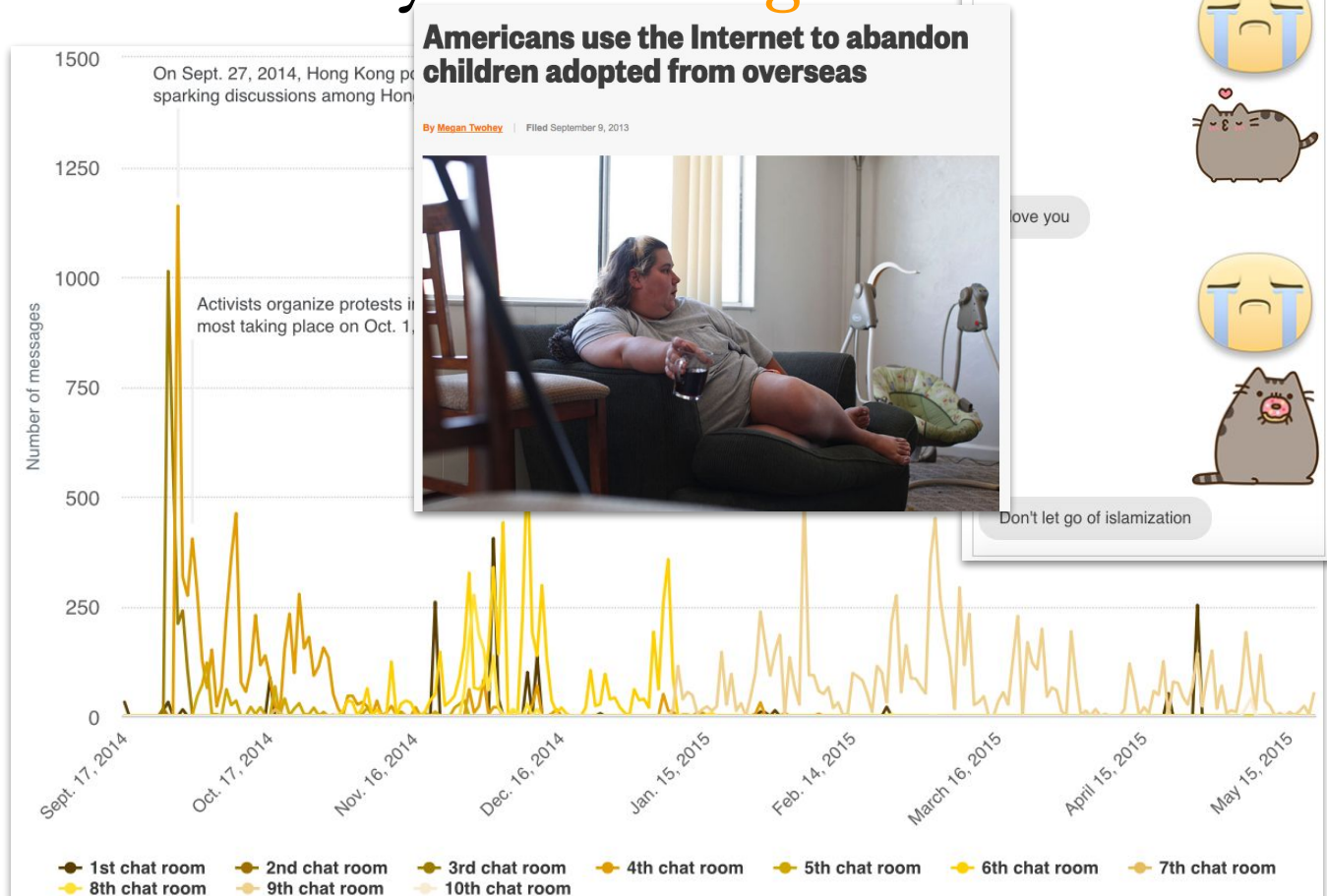
# Who are we, when nobody's **watching**?

## Quantified Selfie:

An *individual's* actions over time can help us understand how the story unraveled!

Precise *documentation* of actions can be more reliable and powerful than memory!

Traces of *behaviour*!





“Without data, you’re just another person  
with an opinion!”

—Edwards Deming—

# APIs: what data is out there?



Limited data streams\* provided by the companies:

- ❖ **Facebook**: Public posts and groups
- ❖ **Twitter**: Searches going back to 7 days and 3,200 latest tweets from a person, ~5,000 friends and followers
- ❖ **Instagram**: Nada

\* Due to security concerns

# Twitter: what can we mine?

Let's now breakdown this tweet:

1. Read this
2. Recall **F·R·I·E·N·D·S** and laugh a little!
3. Then, tell me *what kind of data* this tweet might yield



**FRIENDS** ✓  
@FriendsTV

Following

Don't make Joey beg.



9:02 AM - 8 Jan 2019

544 Retweets 5,226 Likes



19



544



5.2K



Person who posted it:

Picture link  
Display name  
Twitter handle

Tweet text

Tweet media

Timestamp

Tweet engagement:

Retweets  
Likes  
Replies

(further down)



FRIENDS

@FriendsTV

Following

Don't make Joey beg.



9:02 AM - 8 Jan 2019

544 Retweets 5,226 Likes



19 544 5.2K





# Twitter: things to look out for!

## Authentication:

- ❖ Need-based and is different for each company
- ❖ Instagram needs an app\* while Twitter and Facebook don't!

## Rate Limits:

- ❖ How *often* can you grab data from an API per minute?

Read the *Documentation!*

## Scope:

- ❖ The *kind of access* you're getting to the data
- ❖ Twitter's historical searches go back *7 days*
- ❖ Facebook gives *your data*, data from *public groups*

\* Not the app created on the *developer* website

<https://apps.twitter.com/>

### App details

The following app details will be visible to app users and are required to generate the API keys needed to authenticate Twitter developer products.

#### App name (required) ?

Predictive Analytics (PAA 2019)

Maximum characters: 32

#### Application description (required)

Share a description of your app. This description will be visible to users so this is a good place to tell them what your app does.

Predictive analytics to social-media can aid us researchers harness its power to detect psychological traits!

Between 10 and 200 characters

#### Website URL (required) ?

<https://linkedin.com/in/tushaargvs/>

Allow this application to be used to sign in with Twitter

[Learn more](#)

☐ Enable Sign in with Twitter

#### Callback URLs ?

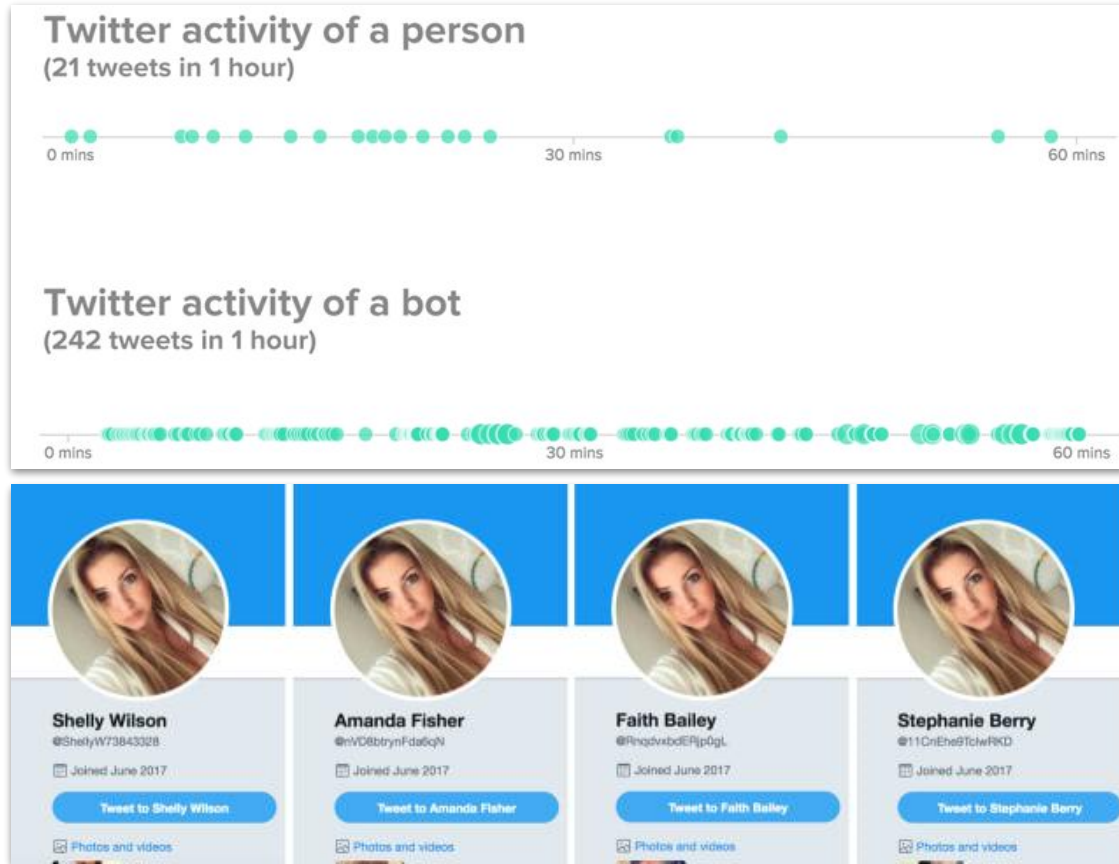
# Is technology distorting the Information Channels?

“ Tweeting 72 times a day is suspicious, but 144 times a day is very suspicious! ”

—DFRL—

Bot armies can amplify voices and make a group of 5 people appear like 45,000 people!

To what extent are automated accounts *problematic* and to what extent are they a way for an individual (programmer) to *express themselves*?



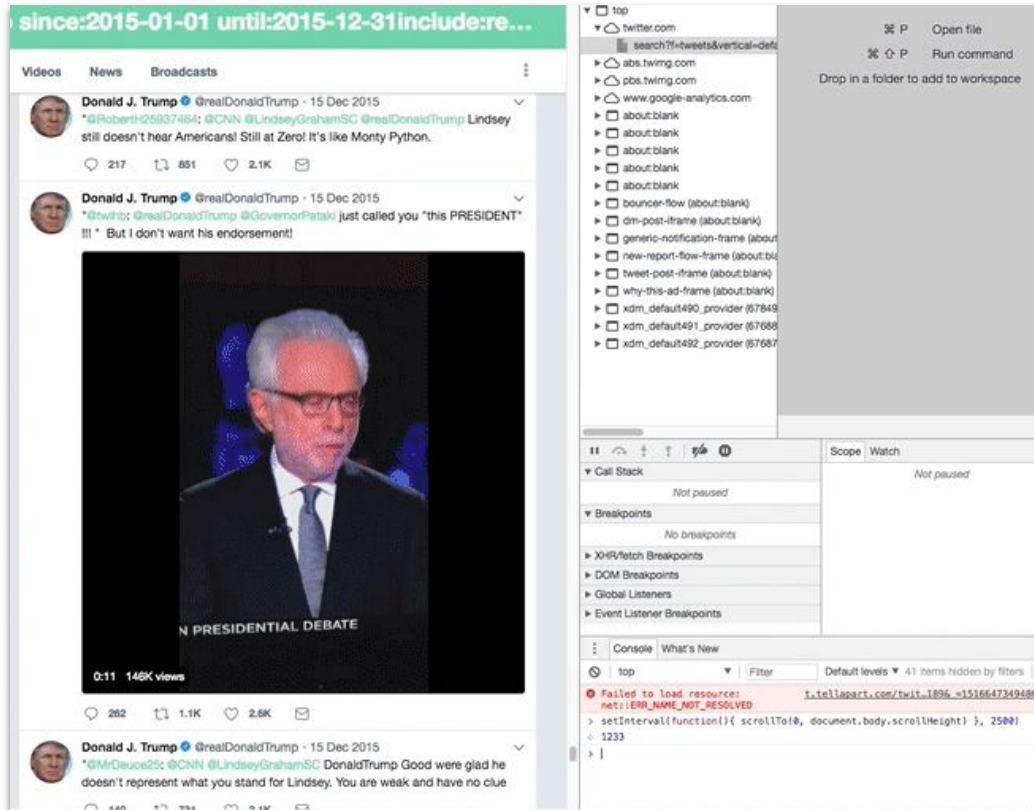
# The data you own: Personal Archives!

Social-Media	What Information	Format of Information
Facebook	The archive Messages, timeline activities	.html files (HTML web pages) and .json files
Google	Google allows us to download our own data from any google services we may be using. This includes our email archive (delivered as an mbox file that email clients recognize), our calendars, our photos and our google maps locations, if applicable	Various depending on the product. Email, for instance, comes as an .mbox file that can be opened in an email client like Thunderbird. Other formats include: XML, HTML, PowerPoint, Word, JSON, GeoJSON
Instagram	The company doesn't seem to allow for a download but it seems to be available via 3 <sup>rd</sup> party apps	
LinkedIn	LinkedIn allows us to request a download an archive of our account, which includes our activities, our profile and our contacts	.csv files (Spreadsheets)
Tumblr	The company doesn't seem to allow for a download but it seems to be available via 3 <sup>rd</sup> party apps	
Twitter	Twitter's archive download includes a web site that allows us to browse through our tweets as well as a spreadsheet of our activity	.html files, .csv files
WhatsApp	Our WhatsApp chats can be requested as a text file to be sent to our email	.txt documents

Data compilations of *your own actions* is provided by the company! All you have access to, is *your own account*!



# The data that's out there: Web Scraping!



Every piece of content present *online* is a potential trove of information to be collected.

Scraping data from websites can happen via *scripts* or any other methods!

# Filter Bubbles

Algorithmic creation of filter bubbles leads to *segregation* of information. Already existing differences just get exacerbated\*!



\* Facebook as a Research Tool for the Social Sciences.  
M. Kosinski et al. American Psychologist. 2015.

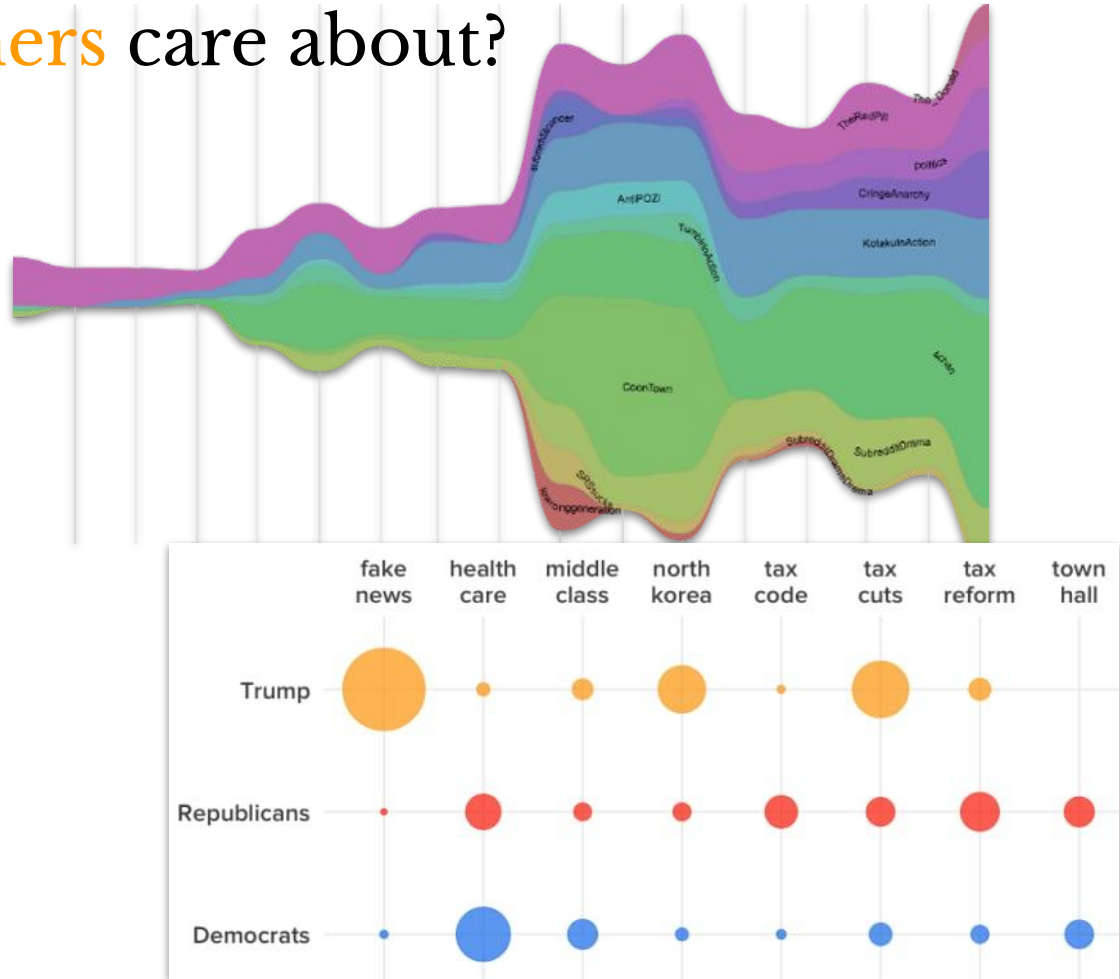
“ The goal is to turn data into information and  
information into insight! ”

—Carly Fiorina—

# What do we researchers care about?

We can use social-media to:

- ❖ Serve as a *proxy* to understand real people
- ❖ Look at the *online ecosystem* that is proliferating!
- ❖ Tracing and recounting *human actions* over time
- ❖ Analyze *information wars* (fought by humans and non-humans)
- ❖ Understand the *skewed information universes*

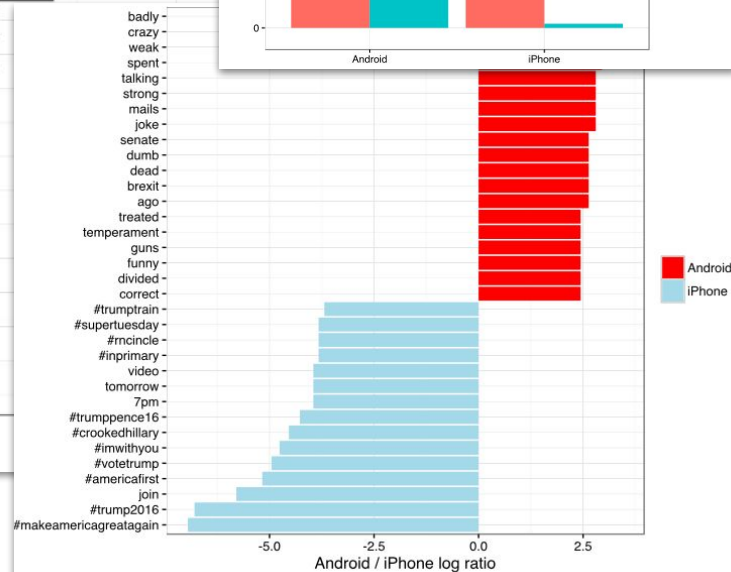
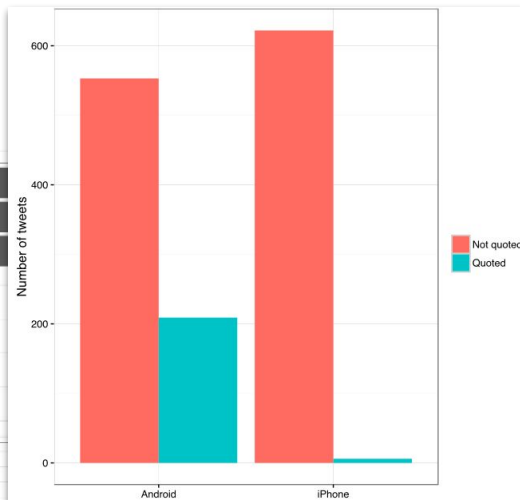
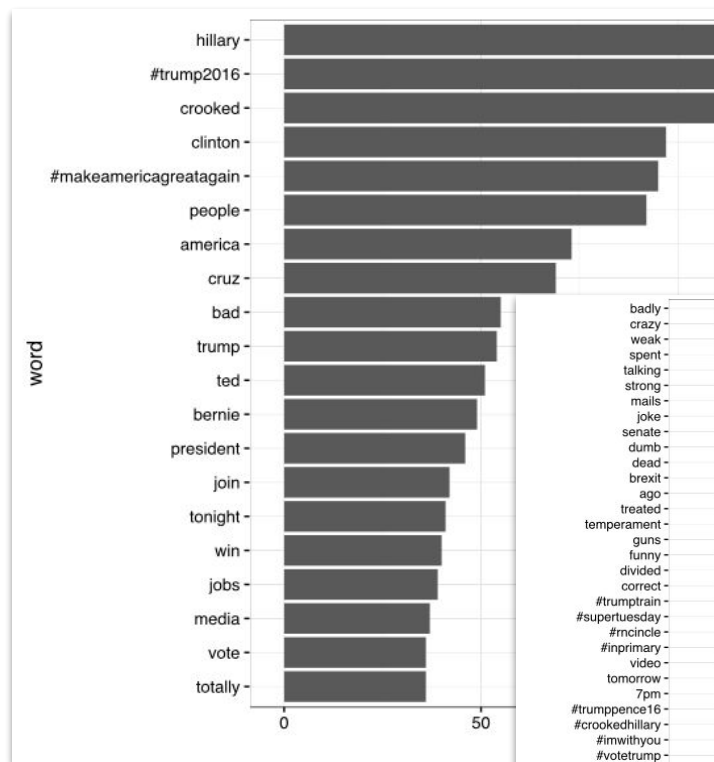




# What can we **analyze** from social data?

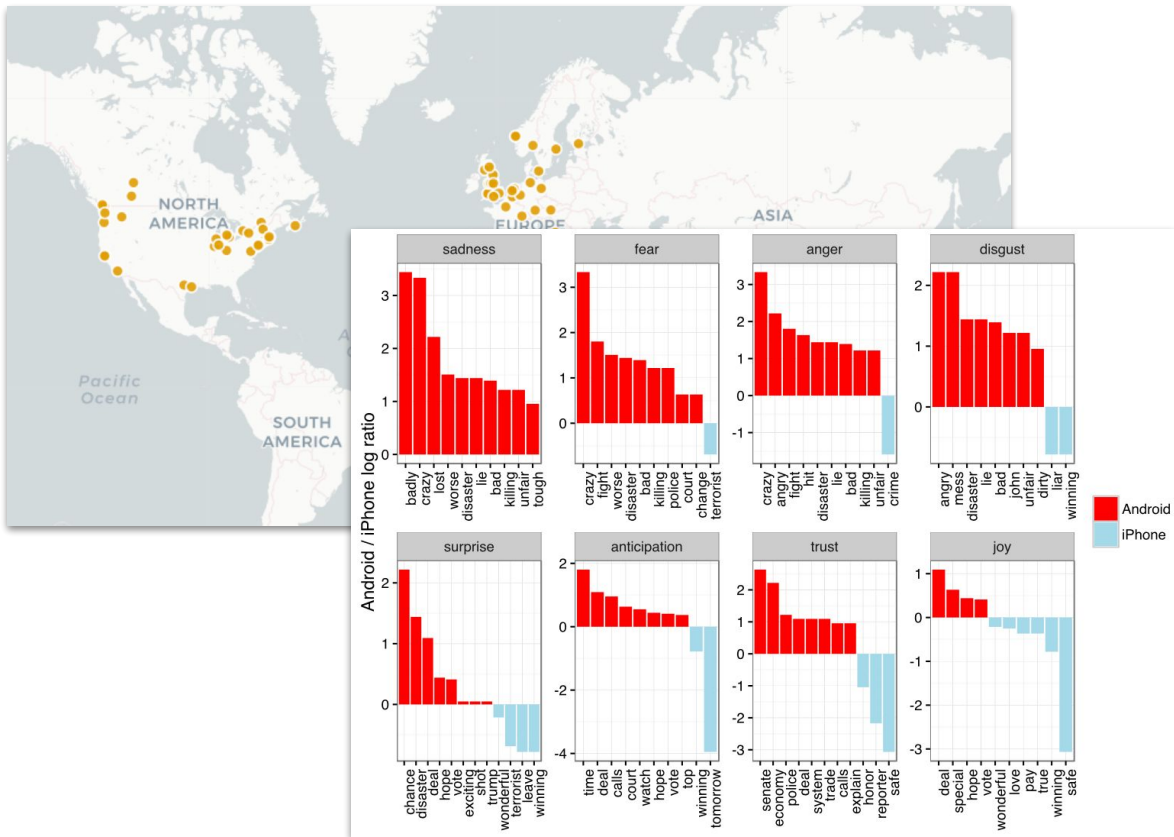
## 1. Finding *Patterns* and *trends* with time!

- a.* Sources
- b.* Sentiments
- c.* Locations
- d.* Hashtags
- e.* Alliances
- f.* Words
- g.* Posts' Volume
- h.* Reactions' Volume
- i.* Profiling



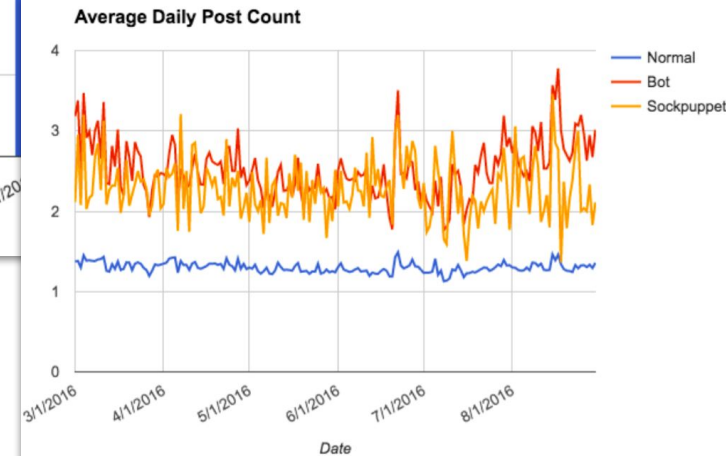
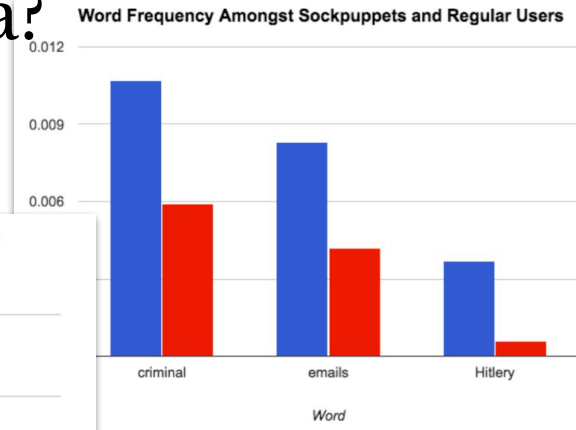
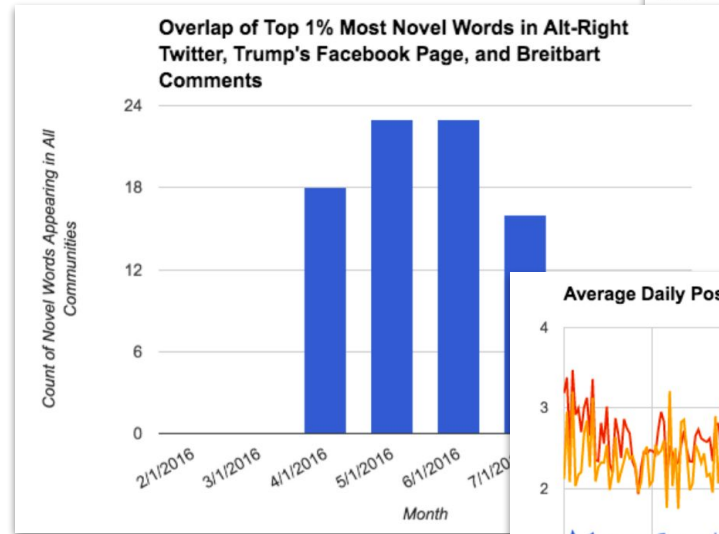
# What can we **analyze** from social data?

1. Finding *Patterns* and *trends* with time!
2. Word analysis to understand how *groups* talk!
  - a. Active Participants
  - b. Clearer Goals
  - c. Linguistics
  - d. Hot Topics
  - e. User Mentions
  - f. Social Network

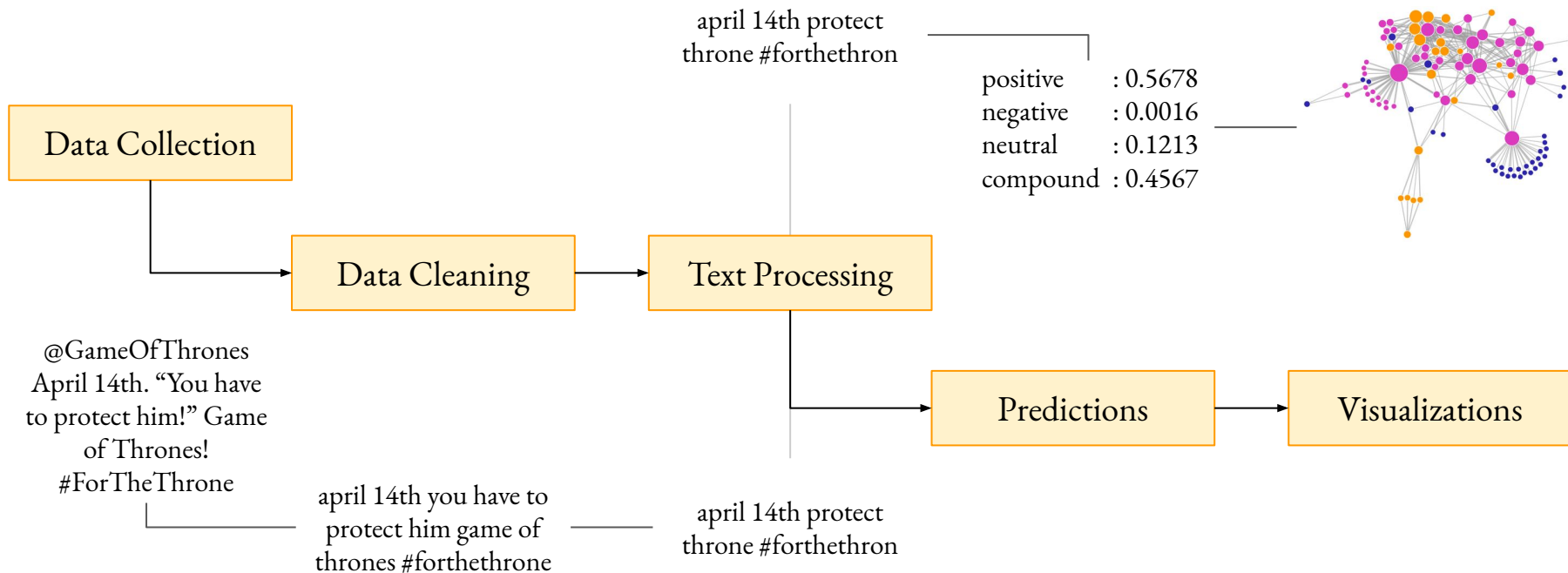


# What can we **analyze** from social data?

1. Finding *Patterns* and *trends* with time!
2. Word analysis to understand how *groups* talk!
3. Understand the *manipulation of information* through bots and sockpuppets!
  - a. Linguistics
  - b. Influence



# How do we **analyze** social data: **NLP** to the rescue\*!



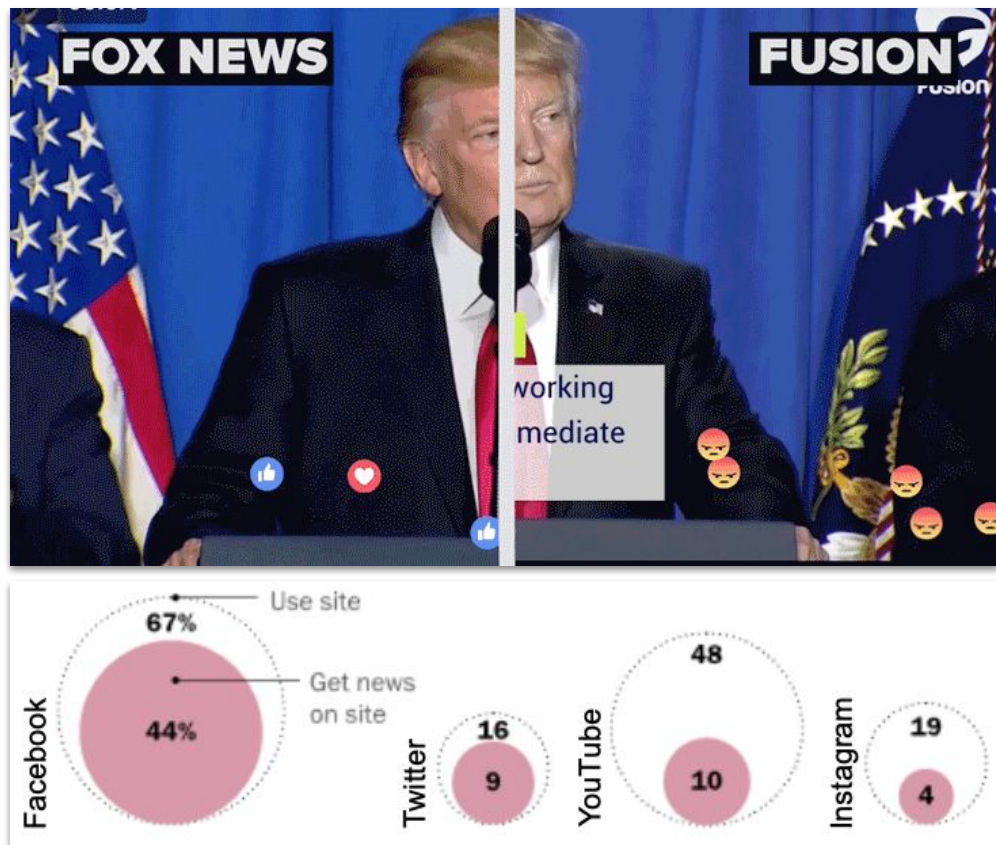
\* We'll limit our discussion to *textual* data and leave multimedia data out!



“ There are knowns, and there are unknowns!  
Then, there are known unknowns, and  
unknown unknowns! ”

—Ryson D'souza—

# Any Data Caveats?



- ❖ Get some context as to who uses what!
  - ❖ *Sources* of data
  - ❖ *Proportions* of the compositions
- ❖ Be wary of the tyranny of the loudest!
  - ❖ *Volumetric* analyses
- ❖ People may not be who they say they are!
  - ❖ *Sockpuppet* and *bot* profiles!

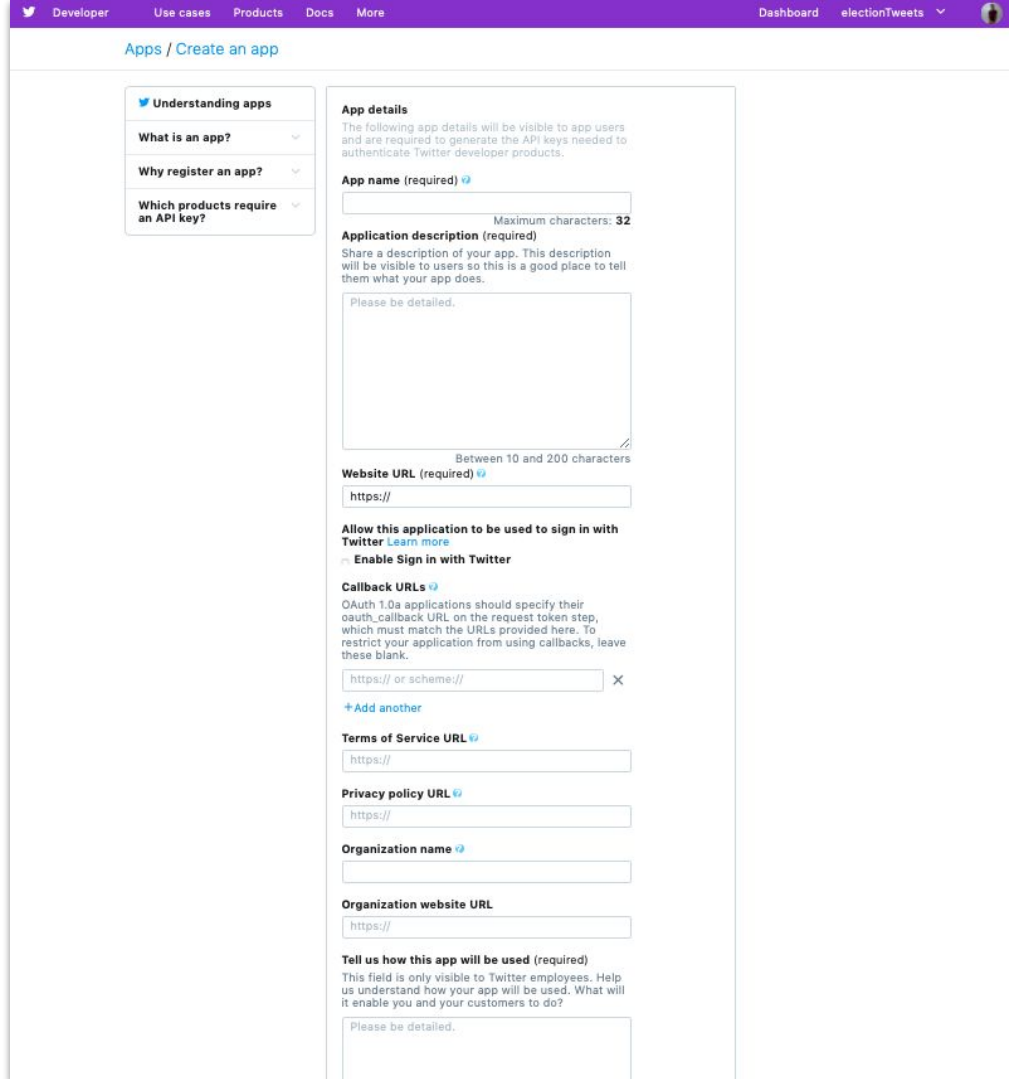
# Further Reading!

- [1] Matthew Russell. *Mining the Social Web, Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites*. O'Reilly Media. <http://shop.oreilly.com/product/0636920010203.do>. 2011.
- [2] Sadilek Adam and John Krumm. *Far Out: Predicting Long-Term Human Mobility*. AAAI. <https://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/download/4845/525>. 2012.
- [3] Derek Ruths and Jürgen Pfeffer. *Social media for large studies of behavior*. Science. [https://www.researchgate.net/profile/Juergen\\_Pfeffer/publication/268879558\\_Social\\_Media\\_for\\_Large\\_Studies\\_of\\_Behaviour/links/55f87ff508ae07629dd77bbb/Social-Media-for-Large-Studies-of-Behaviour.pdf](https://www.researchgate.net/profile/Juergen_Pfeffer/publication/268879558_Social_Media_for_Large_Studies_of_Behaviour/links/55f87ff508ae07629dd77bbb/Social-Media-for-Large-Studies-of-Behaviour.pdf). 2014.
- [4] University of Amsterdam. *The Digital Methods Initiative*. <https://wiki.digitalmethods.net/Dmi/ToolDatabase>.
- [5] *Capturing Stories from the Social Web*. Data Journalism Handbook: Chapter on Social Media. [https://docs.google.com/document/d/1O8q-c\\_fj1LIOSIOOayl\\_wquJXcbVNGCrqR3pcaDCVnw/edit#heading=h.ttv3e2c46o6m](https://docs.google.com/document/d/1O8q-c_fj1LIOSIOOayl_wquJXcbVNGCrqR3pcaDCVnw/edit#heading=h.ttv3e2c46o6m).
- [6] *Finding Stories in Social Media Data*. Book Proposal. <https://docs.google.com/document/d/1gXKdILpTmwzvn5w7mj7NgN55zT668xrM1wNjCYJG3Mw/edit#heading=h.5p4u2elmd8ry>.
- [7] Jason Brownlee. *What is NLP?* <https://machinelearningmastery.com/natural-language-processing/>. 2017

# It's your turn!

Use <https://apps.twitter.com/> to create *your* application!

**P.S.** Fill out the details carefully! It'll be verified and you'll be given access *within a day's time*!



The screenshot shows the 'Create an app' page on the Twitter Developer Portal. The page has a purple header with navigation links: Developer, Use cases, Products, Docs, More, Dashboard, and electionTweets. The main content area is titled 'Apps / Create an app' and features a sidebar with 'Understanding apps' and a list of links: 'What is an app?', 'Why register an app?', and 'Which products require an API key?'. The main form is titled 'App details' and includes instructions for app users. It contains several input fields: 'App name (required)' with a character limit of 32, 'Application description (required)' with a character limit of 10 to 200, 'Website URL (required)', 'Callback URLs' (with an 'Add another' link), 'Terms of Service URL', 'Privacy policy URL', 'Organization name', and 'Organization website URL'. There is also a section for 'Tell us how this app will be used (required)'.

Twitter Developer Portal: Apps / Create an app

**Understanding apps**

- What is an app?
- Why register an app?
- Which products require an API key?

**App details**

The following app details will be visible to app users and are required to generate the API keys needed to authenticate Twitter developer products.

**App name (required)** [?](#)

Maximum characters: 32

**Application description (required)**

Share a description of your app. This description will be visible to users so this is a good place to tell them what your app does.

Please be detailed.

Between 10 and 200 characters

**Website URL (required)** [?](#)

https://

**Allow this application to be used to sign in with Twitter** [Learn more](#)

☐ Enable Sign in with Twitter

**Callback URLs** [?](#)

OAuth 1.0a applications should specify their oauth\_callback URL on the request token step, which must match the URLs provided here. To restrict your application from using callbacks, leave these blank.

https:// or scheme:// [×](#)

[+ Add another](#)

**Terms of Service URL** [?](#)

https://

**Privacy policy URL** [?](#)

https://

**Organization name** [?](#)

**Organization website URL**

https://

**Tell us how this app will be used (required)**

This field is only visible to Twitter employees. Help us understand how your app will be used. What will it enable you and your customers to do?

Please be detailed.

“ Social-media is a network that turns embers to flames, within no time! Harnessing such power only gives us, researchers limitless possibilities! ”