Social Media & Text Analysis

lecture 2 - Twitter API

CSE 5539-0010 Ohio State University

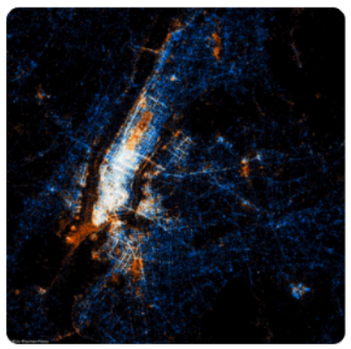
Instructor: Alan Ritter

Website: socialmedia-class.org

Course Website

socialmedia-class.org

Social Media & Text Analytics Syllabus Twitter API Tutorial Homework



A visualization showing the location of Twitter messages (blue) and Flickr photos (orange) in New York City by Eric Fischer

Social media provides a massive amount of valuable information and shows us how language is actually used by lots of people. This course will give an overview of prominent research findings on language use in social media. The course will also cover several machine learning algorithms and the core natural language processing techniques for obtaining and processing Twitter data.

Instructor

Wei Xu is an assistant professor in the Department of Computer Science and Engineering at the Ohio State University. Her research interests lie at the intersection of machine learning, natural language processing, and social media. She holds a PhD in Computer Science from New York University. Prior to joining OSU, she was a postdoc at the University of Pennsylvania. She is organizing the ACL/COLING Workshop on Noisy User-generated Text, serving as a workshop co-chair for ACL 2017, an area chair for EMNLP 2016 and the publicity chair for NAACL 2016.

Time/Place new

Fall 2017, CSE 5539-0010 The Ohio State University
Bolz Hall Room 318 | Tuesday 2:20PM – 4:10PM
dual-listed undergraduate and graduate course
[Office Hour] Dreese 495 | Tuesday 4:15PM – 5:15PM

Prerequisites

In order to succeed in this course, you should know basic probability and statistics, such as the chain rule of probability and Bayes' rule. On the programming side, all projects will be in Python. You should understand basic computer science concepts (like recursion), basic data structures (trees, graphs), and basic algorithms (search, sorting, etc).

Course Readings

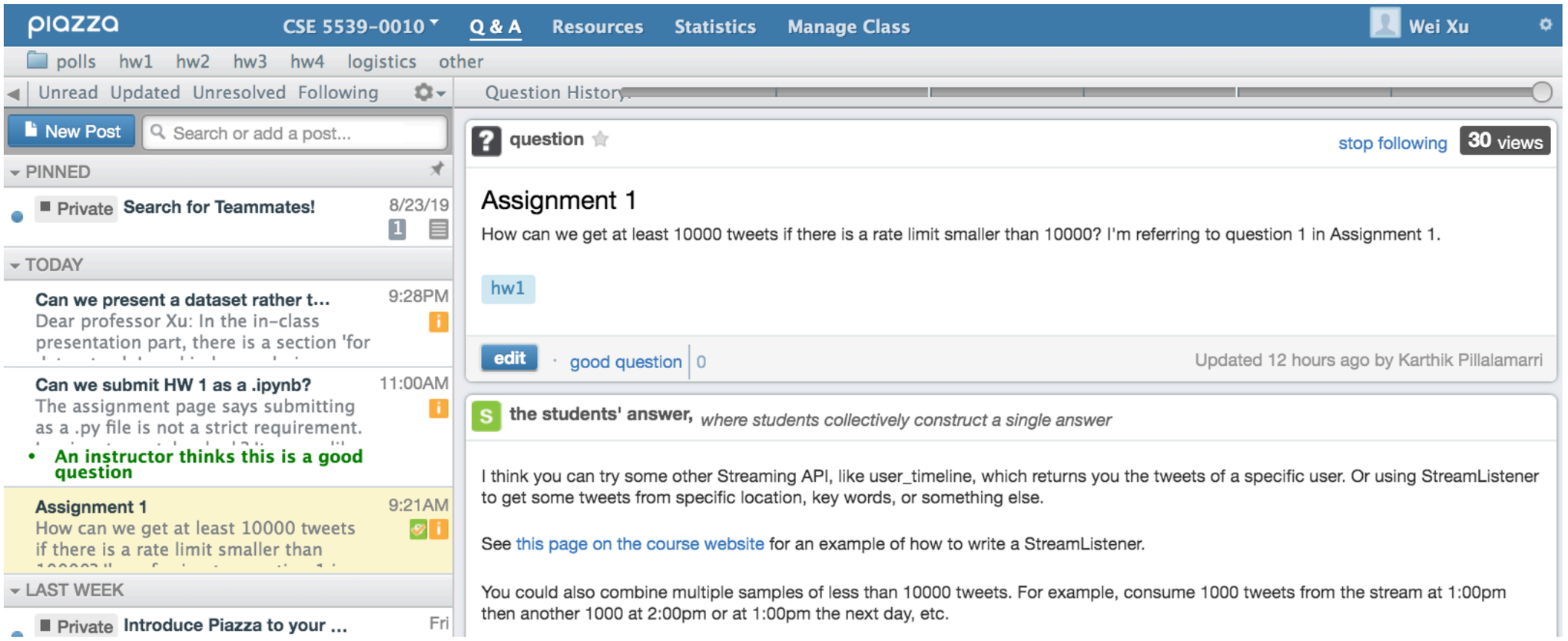
Various academic papers

Discussion Board

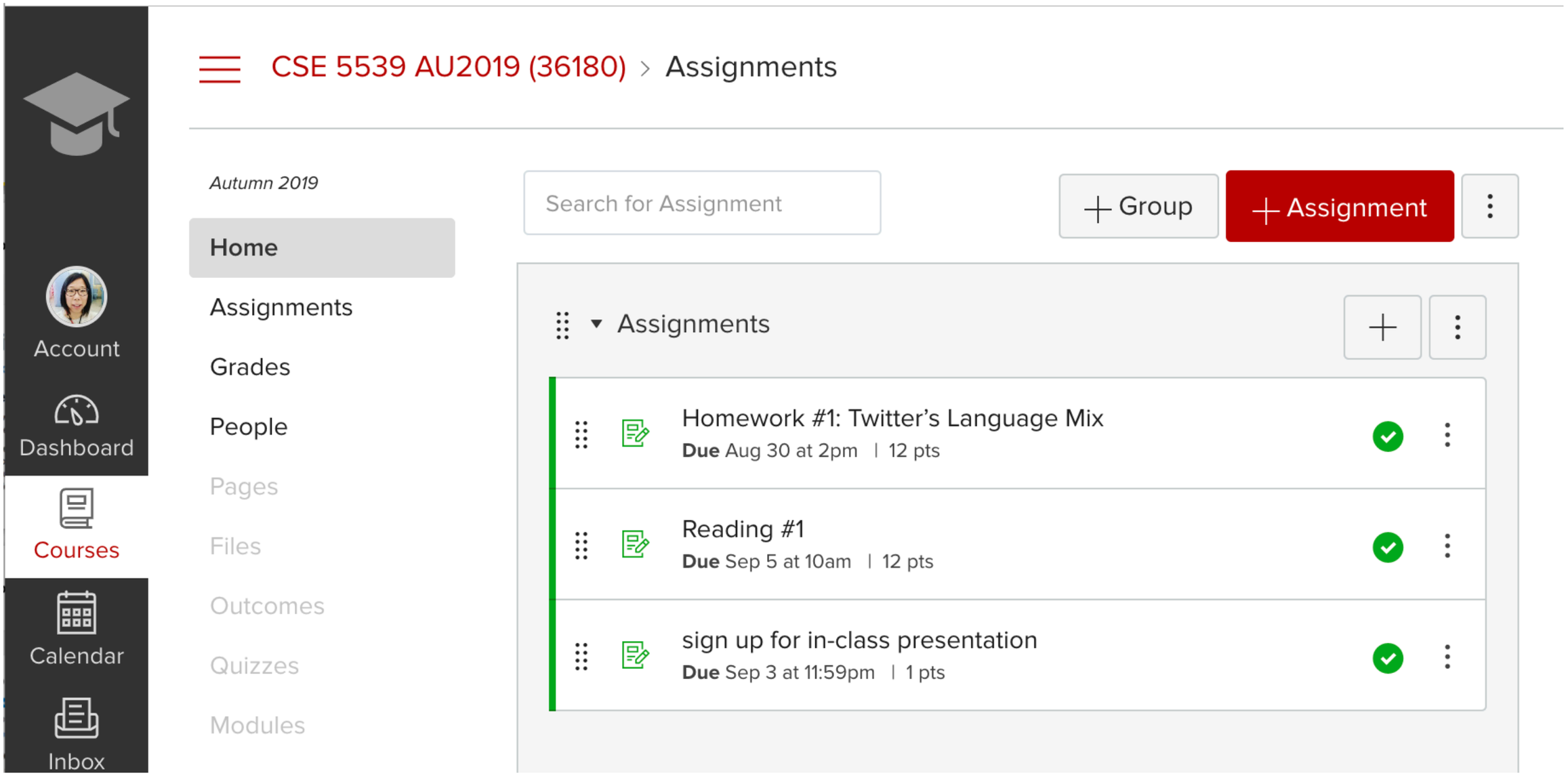
Piazza (TBA)

Have a Question?

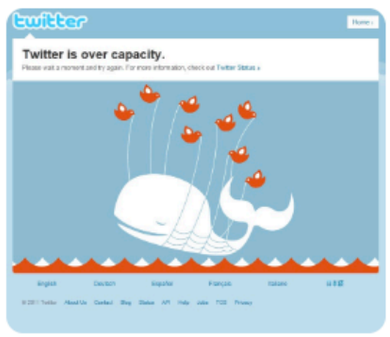
- **Ask in class!**
- **Office Hour:** Fri 4-5 pm, Dreese 495
- **Piazza Q&A Board**



Homework #1


CSE 5539 AU2019 (36180) > Assignments
Autumn 2019
Home
Assignments
Grades
People
Pages
Files
Outcomes
Quizzes
Modules
Account
Dashboard
Courses
Calendar
Inbox
Search for Assignment
+ Group
+ Assignment
Assignments
Homework #1: Twitter's Language Mix
Due Aug 30 at 2pm | 12 pts
Reading #1
Due Sep 5 at 10am | 12 pts
sign up for in-class presentation
Due Sep 3 at 11:59pm | 1 pts

Twitter API Tutorial: socialmedia-class.org



Twitter's 404 error page -- the Fail Whale

Twitter API tutorial

by Wei Xu Follow @cocoweixu and Jeniya Tabassum Follow @JeniyaTabassum (Ohio State University)

Last updated March 20, 2018 (added a script for obtaining all followers of a Twitter user; updated with tweepy p

[download the Jupyter notebook for this tutorial]

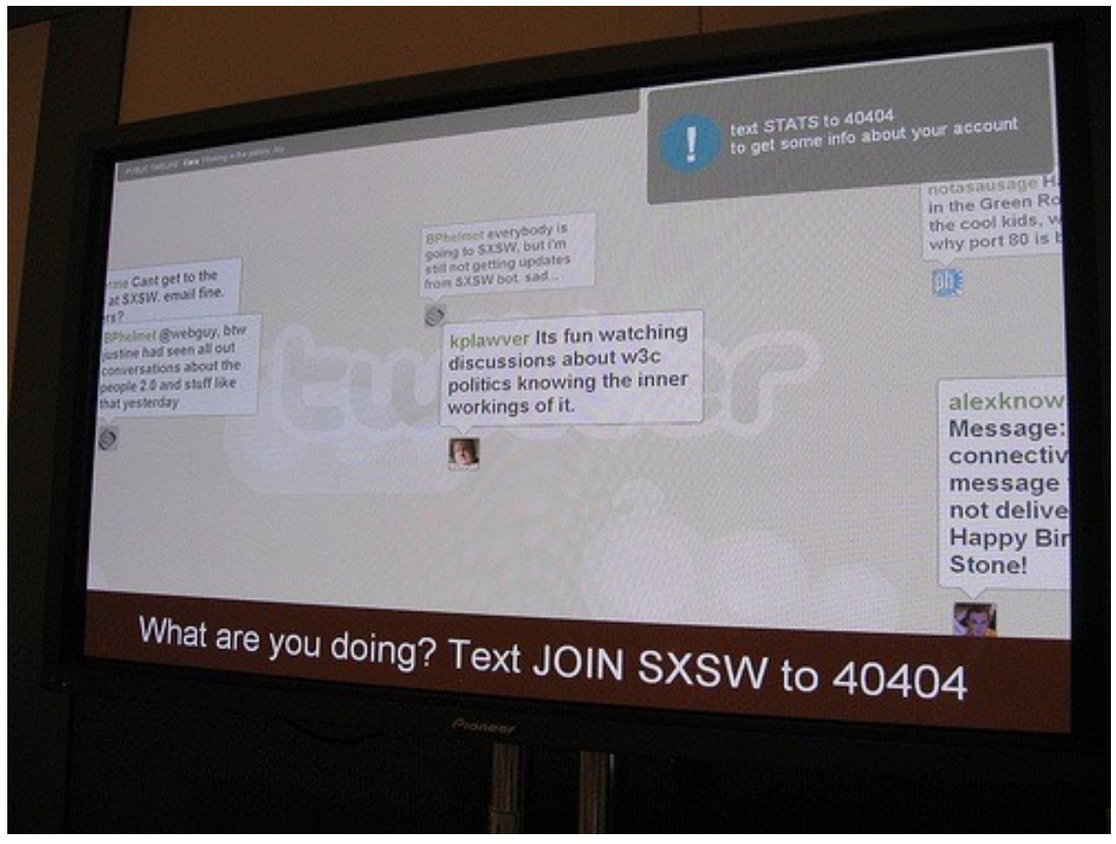
1. Getting Twitter API keys

To start with, you will need to have a Twitter developer account and obtain credentials (i.e. API key, API secret, A and Access token secret) on the to access the Twitter API, following these steps:

- Create a Twitter developer account if you do not already have one from : https://developer.twitter.com/
- Go to https://developer.twitter.com/en/apps and log in with your Twitter user account.
- Click “Create an app”
- Fill out the form, and click “Create”
- A pop up window will appear for reviewing Developer Terms. Click the “Create” button again.
- In the next page, click on “Keys and Access Tokens” tab, and copy your “API key” and “API secret” from `Consumer API keys` section.

Twitter History

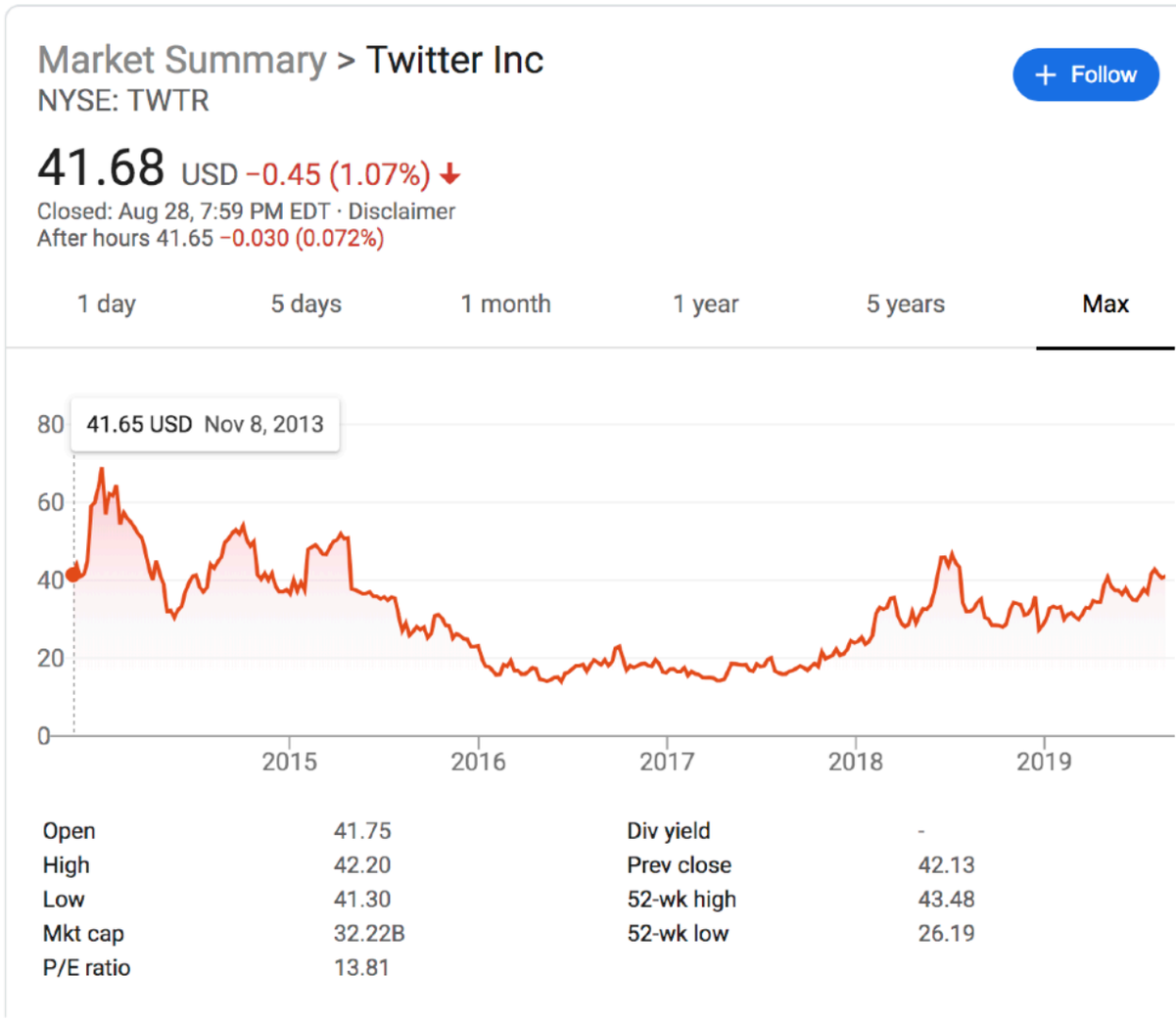
- Jack Dorsey's idea
  (a NYU undergraduate then)
- 1st tweet on March 21, 2006
- exploded at SXSW 2007
  (20k→60k tweets/day)
- 100m tweets/quarter in 2008,
  50m tweets/day in 2010,
  400m tweets/day in 2013
- Huge API usage was
  unexpected as was the rise of
  the @ sign for replies



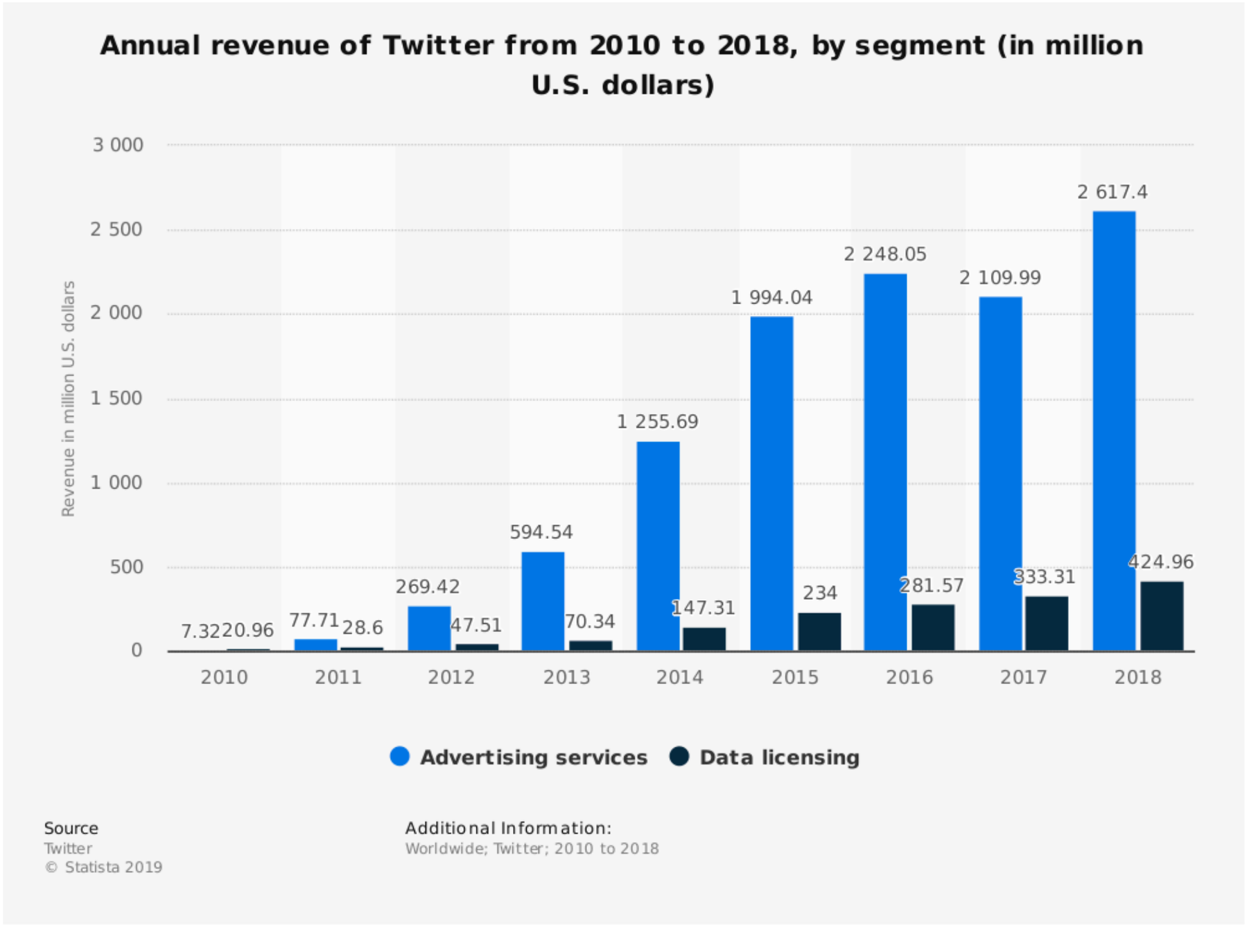
Twitter staff received the festival's Web Award prize with the remark "we'd like to thank you in 140 characters or less. And we just did!"

Twitter History

- IPO in 2013 Q4
- market value $24b, revenue $435m, net loss $162m in 2015 Q1
- CEO Dick Costolo resigned July 1st, 2015
- Dorsey was named permanent CEO of Twitter onOctober 5, 2015



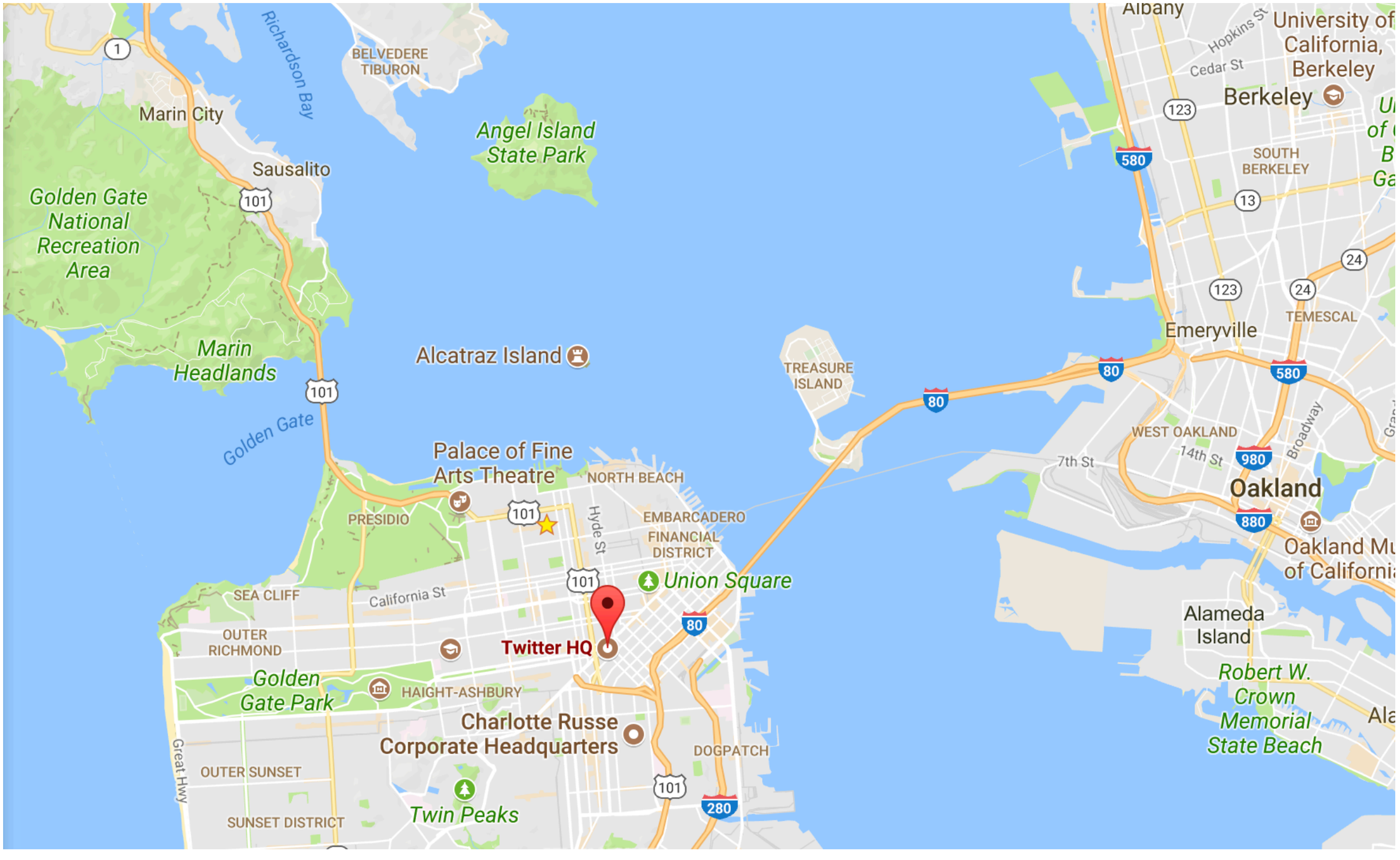
Twitter Revenue



Source
Twitter
© Statista 2019

Additional Information:
Worldwide; Twitter; 2010 to 2018

Twitter HQ (since 2012)

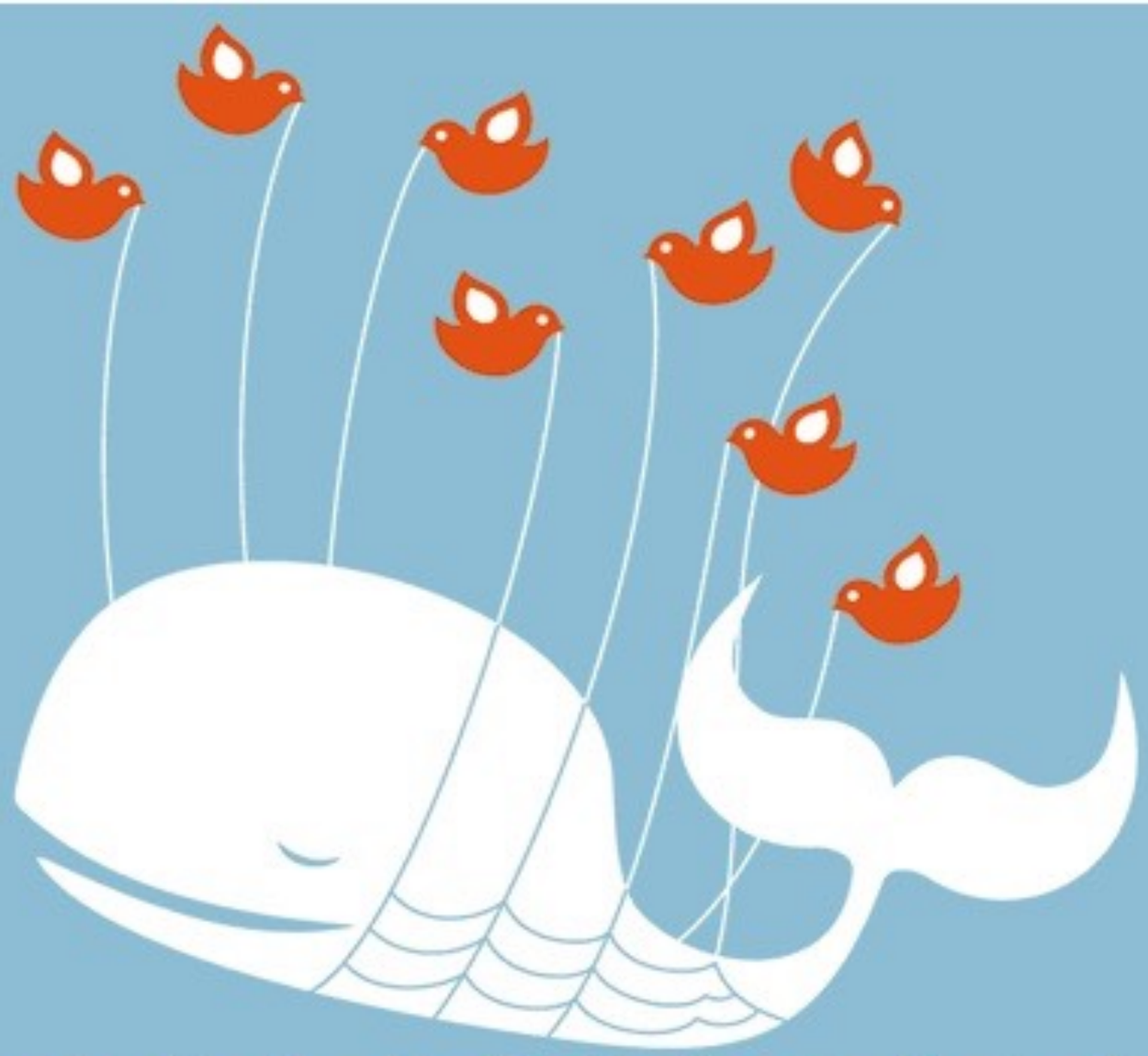


twitter

Home ›

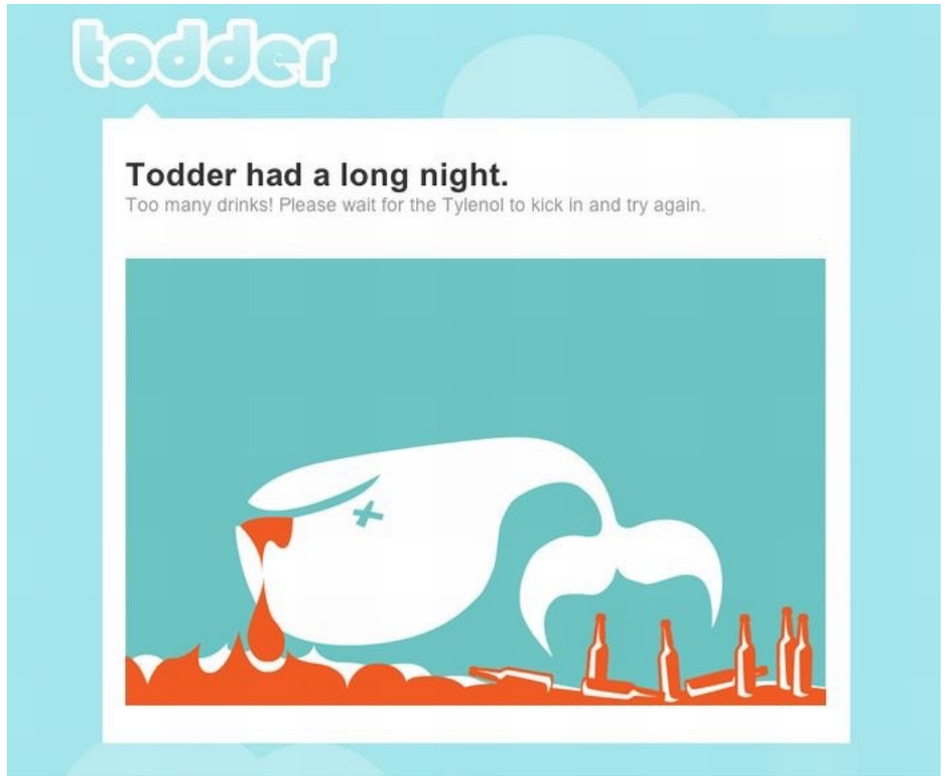
Twitter is over capacity.

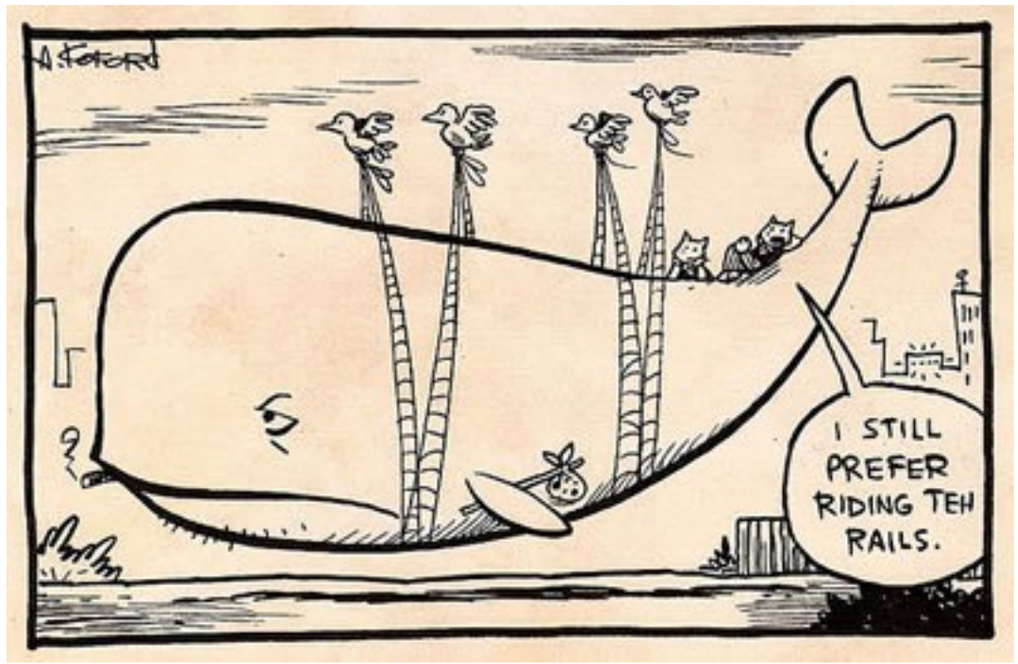
Please wait a moment and try again. For more information, check out Twitter Status »



English Deutsch Español Français Italiano 日本語

 About Us Contact Blog Status API Help Jobs TOS Privacy


todder
Todder had a long night.
Too many drinks! Please wait for the Tylenol to kick in and try again.


I STILL PREFER RIDING TEH RAILS.

Tweets

ChuckGrassley
@ChuckGrassley

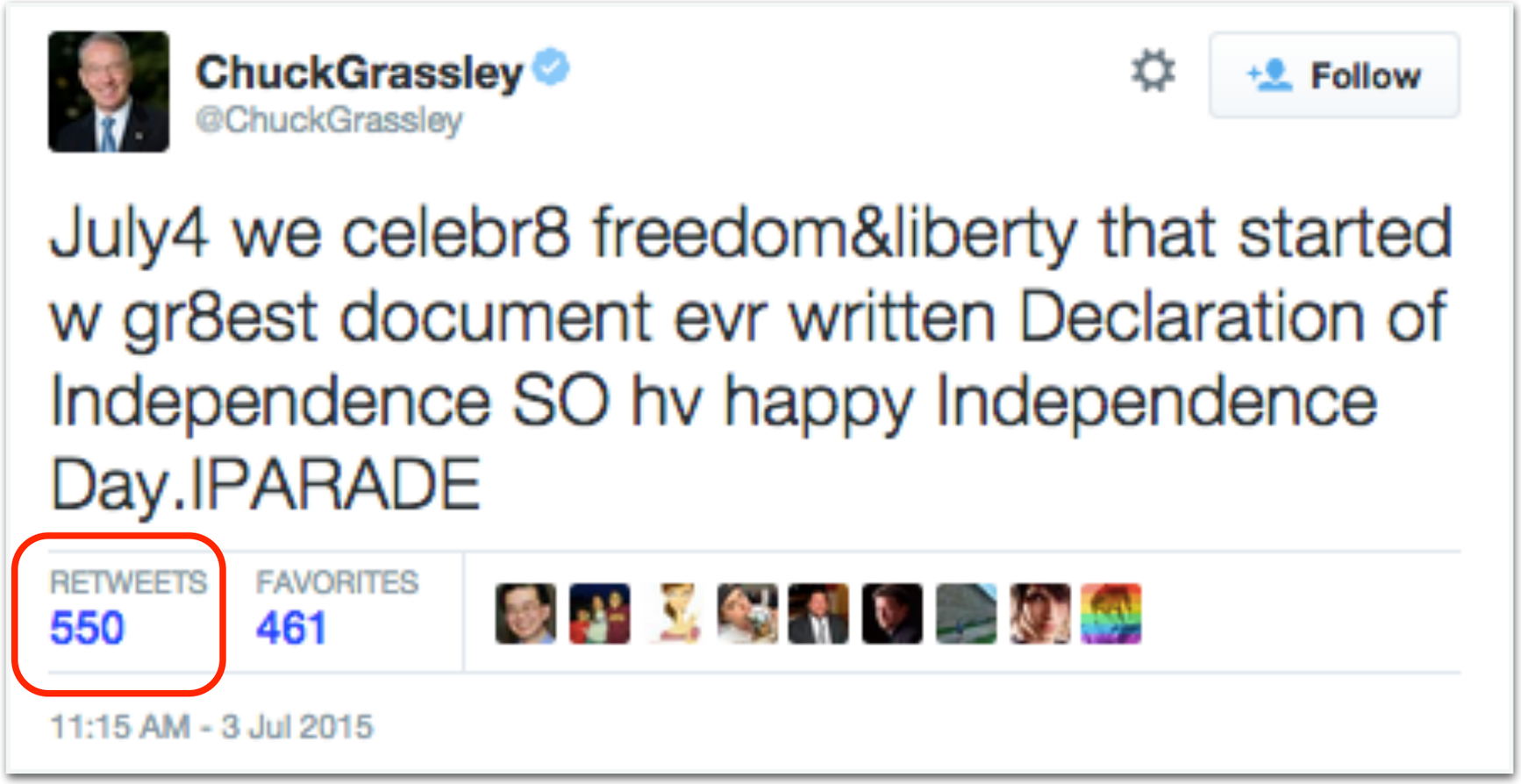
Follow

July4 we celebr8 freedom&liberty that started w gr8est document evr written Declaration of Independence SO hv happy Independence Day.IPARADE

RETWEETS 550 FAVORITES 461

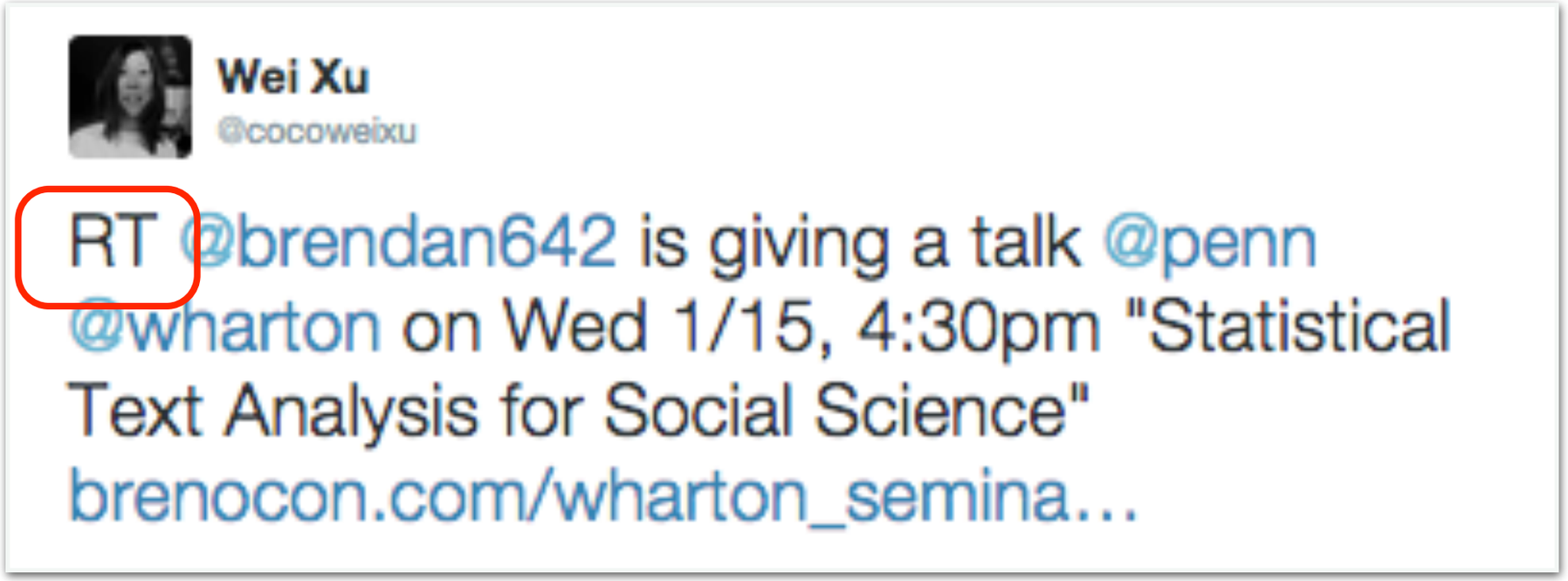
11:15 AM - 3 Jul 2015

ReTweets



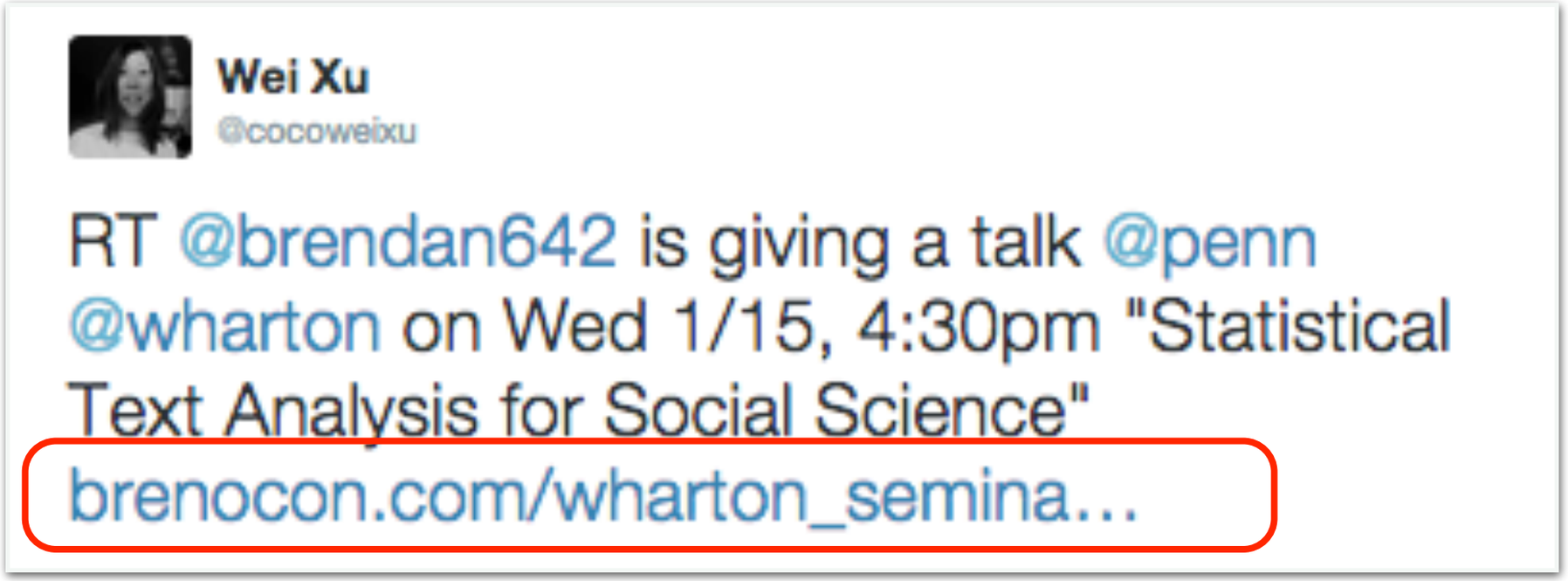
a re-posting of someone else's Tweet

ReTweets



- not an official Twitter feature
- often signifies quoting another user
- sometimes creates problems for data analytics

Embedded Links



- shortened for display

Embedded Links

- can provide extra external information for text processing

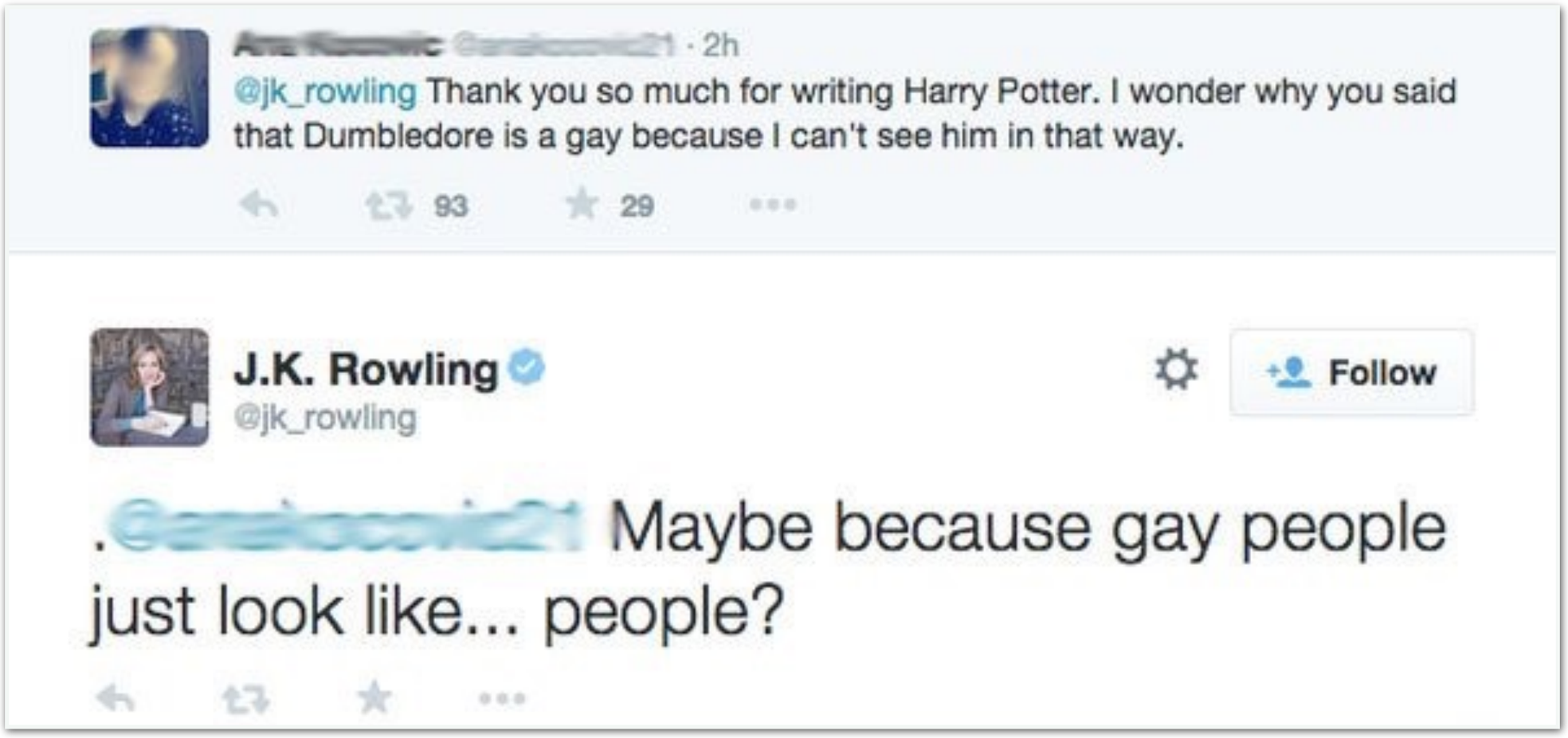
Mentions

Wei Xu
@cocoweixu

RT @brendan642 is giving a talk @penn @wharton on Wed 1/15, 4:30pm "Statistical Text Analysis for Social Science" brenocon.com/wharton_semina…

- user's @username anywhere in the body of the Tweet

Replies/Conversations



- Tweet starts with a @username

Replies/Conversations

- can have multi-round conversations

What are the top forums or discussion websites where leading researchers in the field of Natural Language Processing interact?

1 Answer

Jordan Boyd-Graber, answering questions on Quora because the stakes are so low
Answered Mar 10

It seems to be Twitter (and to a lesser extent, Facebook). Follow your favorite researchers and often technical questions come up.

A random sampling of people I follow on Twitter (as sorted by Twitter):

- Alex Smola (@smolix) | Twitter
- Forough (@fpoursabzi) | Twitter
- Alice Zheng (@RainyData) | Twitter
- Thomas G. Dietterich
- Aaron Clauset (@aaronclauset) | Twitter
- UMD CLIP lab (@umdclip)
- Hugo Larochelle (@hugo_larochelle) | Twitter
- Russ Salakhutdinov
- Tom M Mitchell (@tommmitchell)
- Karl Moritz Hermann
- Edward Grefenstette
- Bert Huang (@berty38) | Twitter
- Tim Vieira (@xtimv) | Twitter
- Yoav Artzi (@yoavartzi) | Twitter
- Omer Levy (@omerlevy_) | Twitter
- Wei Xu (@cocoweixu) | Twitter
- Anima Anandkumar
- Naomi Saphra (@nsaphra) | Twitter
- Dirk Hovy (@dirk_hovy) | Twitter

Jason Eisner

computer science professor at Johns Hopkins

You can learn more about me and my research at http://cs.jhu.edu/~jason. On Quora, I typically answer technical questions about natural language processing and machine learning. Sometimes I also... (more)

Follow 23.2k · Turn On Notifications · Ask Question · •••

Credentials & Highlights More

- Professor at Johns Hopkins University
  2001-present
- Studied at University of Pennsylvania
- Lives in Baltimore
- 2.7m answer views
  37.7k this month
- Top Writer
  2017 and 2016

Feeds

- Answers 216
- Questions 0
- Activity
- Posts 0
- Blogs 0
- Followers 23,283
- Following 5
- Topics 46
- Edits 1,269

216 Answers

Most Recent / 30-Day Views

What are the topics in computer science?

Jason Eisner, computer science professor at Johns Hopkins
Answered Jul 24

You're off to a good start, but yes, there's plenty more! To get a sense of the breadth of CS, you can have a look through the ACM's curriculum guidelines for undergraduate CS education (last updat... (more)

Upvote 75 · Downvote

What are the things I should know as a new CS PhD student?

Jason Eisner, computer science professor at Johns Hopkins
Answered Jun 15, 2015

[A2A] There's lots of advice on the web. Search for "how to be a good grad student" to get some of it.

How to be a Successful Graduate Student, by Mark Dredze (my colleague) and Hanna Wallach, is a good guide with a long list of links at the end, including a link to my own advice page.

2.4k Views · 24 Upvotes · Answer requested by Hao WU

Upvote 24 · Downvote

Knows About

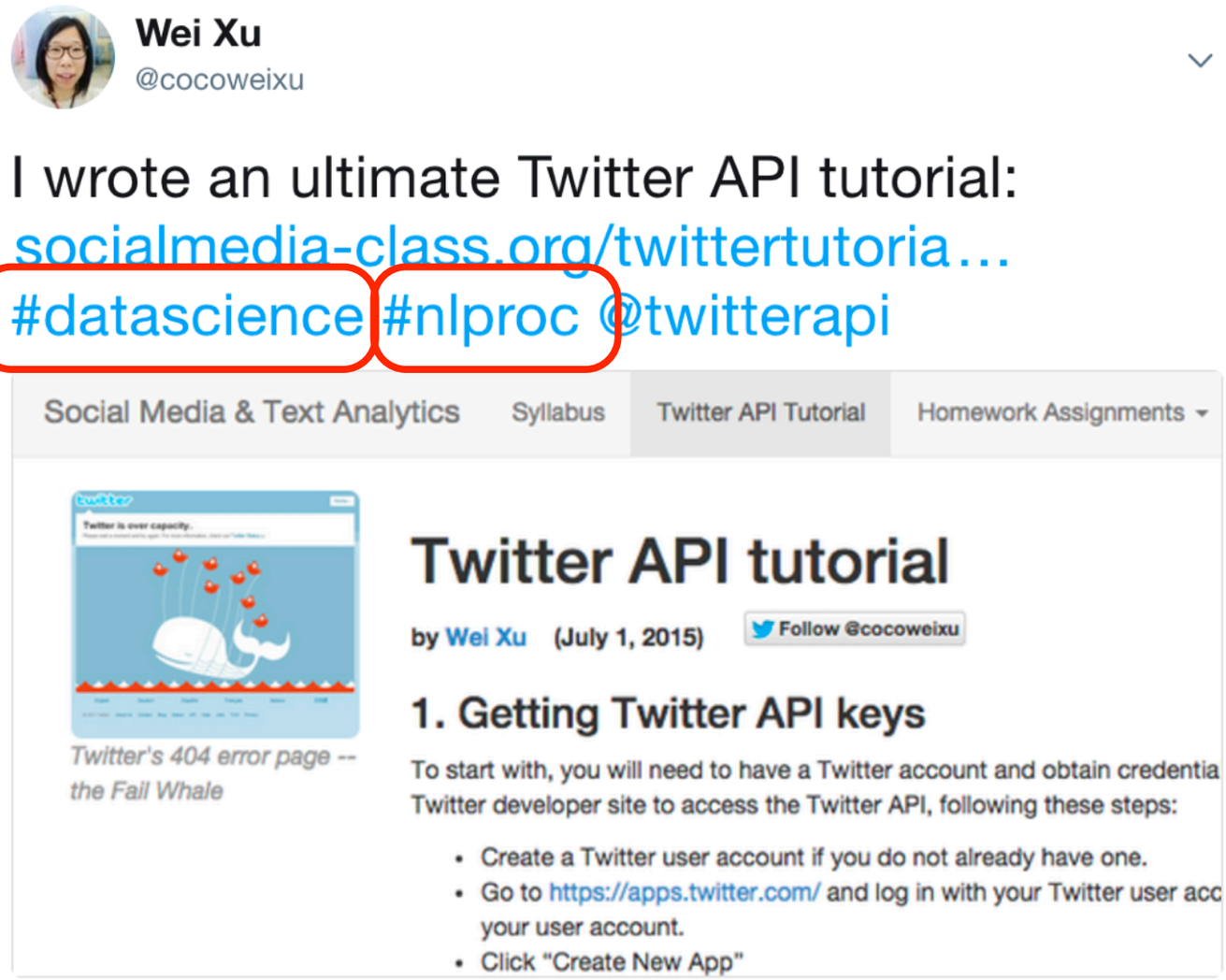
- Graduate School Education
  40 answers
- Academia
  28 answers
- Higher Education
  21 answers
- Machine Learning
  19 answers
- Natural Language Processing
  18 answers

View More

Images

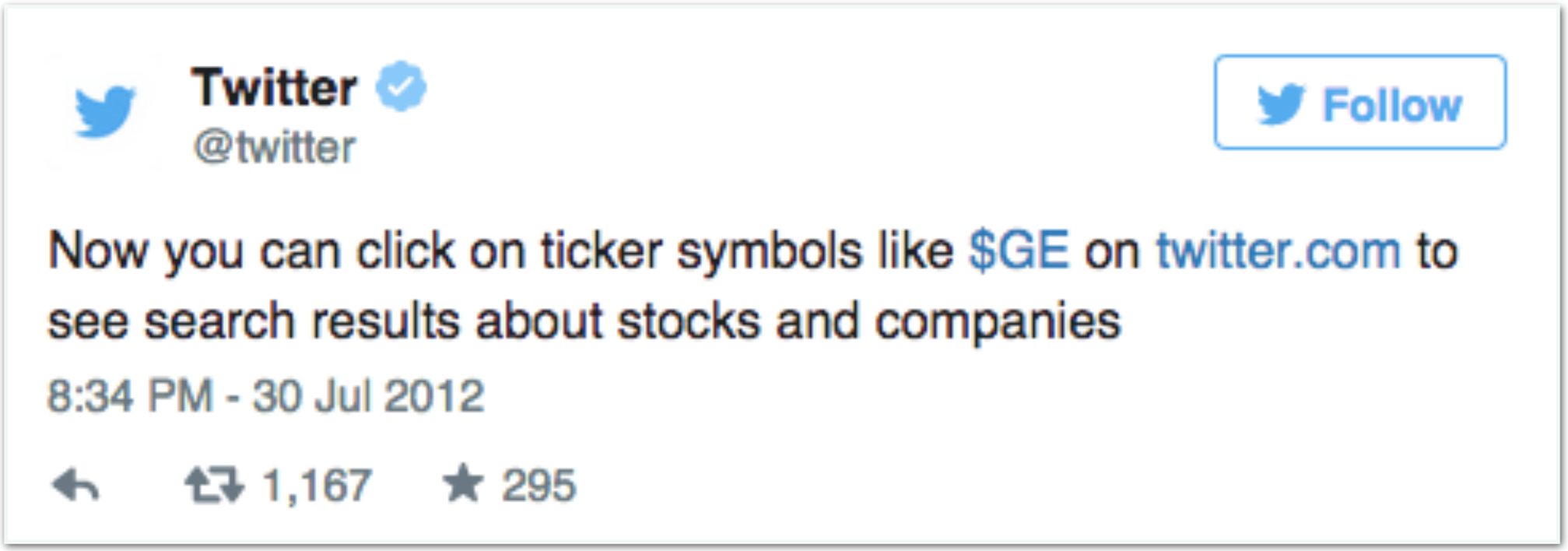

Wei Xu
@cocoweixu
I wrote an ultimate Twitter API tutorial:
socialmedia-class.org/twittertutoria…
#datascience #nlproc @twitterapi
Social Media & Text Analytics
Syllabus
Twitter API Tutorial
Homework Assignments
Twitter is over capacity.
Twitter's 404 error page -- the Fail Whale
Twitter API tutorial
by Wei Xu (July 1, 2015)
Follow @cocoweixu
1. Getting Twitter API keys
To start with, you will need to have a Twitter account and obtain credentia
Twitter developer site to access the Twitter API, following these steps:
Create a Twitter user account if you do not already have one.
Go to https://apps.twitter.com/ and log in with your Twitter user acc
your user account.
Click "Create New App"
11:55 AM - 2 Jul 2015
51 Retweets
105 Likes
6
51
105

Hashtags

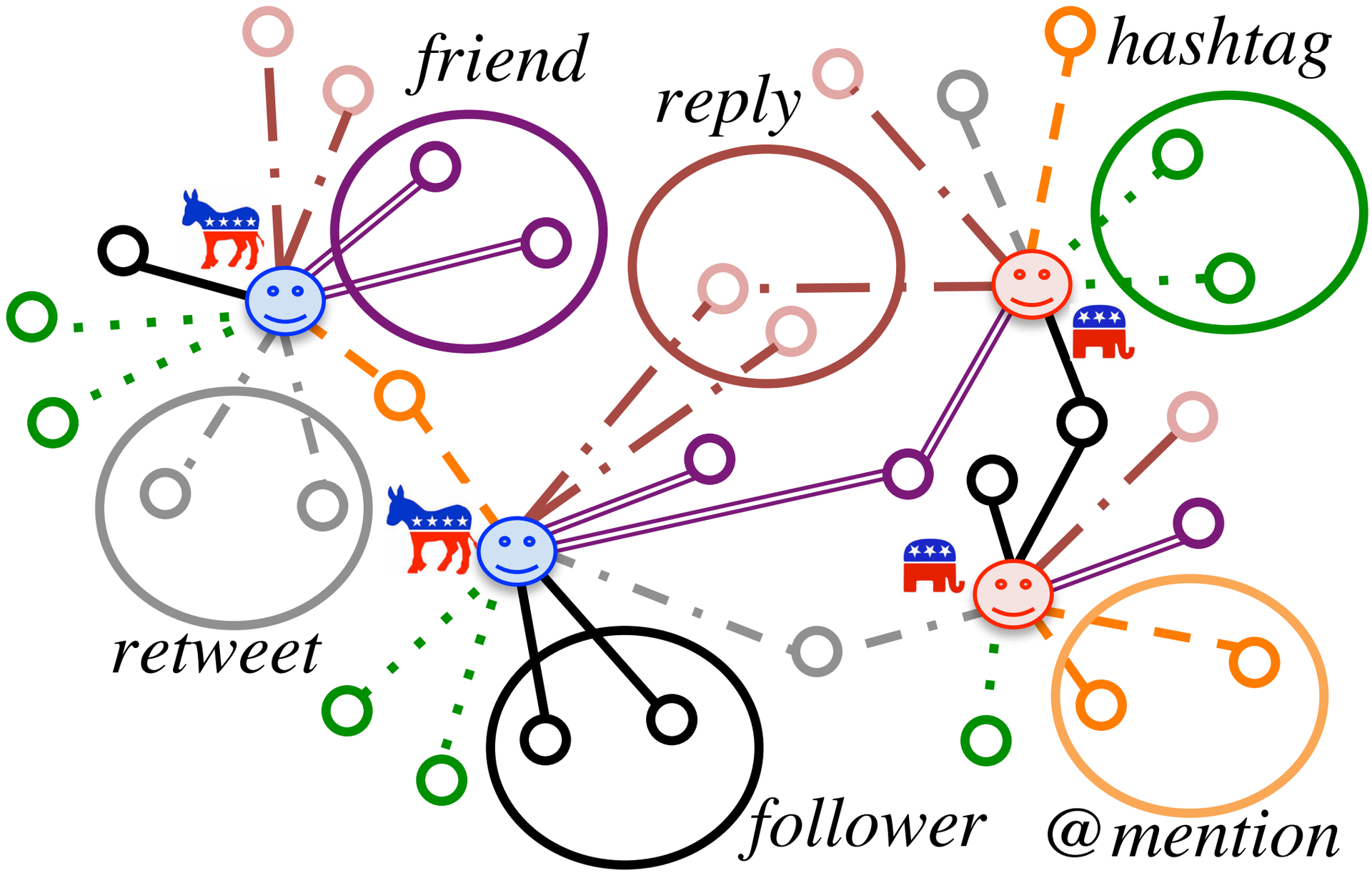

Wei Xu
@cocoweixu
I wrote an ultimate Twitter API tutorial:
socialmedia-class.org/twittertutoria…
#datascience #nlproc @twitterapi
Social Media & Text Analytics
Syllabus
Twitter API Tutorial
Homework Assignments
Twitter is over capacity.
Twitter's 404 error page -- the Fail Whale
Twitter API tutorial
by Wei Xu (July 1, 2015)
Follow @cocoweixu
1. Getting Twitter API keys
To start with, you will need to have a Twitter account and obtain credentia
Twitter developer site to access the Twitter API, following these steps:
Create a Twitter user account if you do not already have one.
Go to https://apps.twitter.com/ and log in with your Twitter user acc
your user account.
Click "Create New App"
11:55 AM - 2 Jul 2015
51 Retweets 105 Likes
6
51
105

hashtags are powerful

Cashtags


Twitter
@twitter
Follow
Now you can click on ticker symbols like $GE on twitter.com to see search results about stocks and companies
8:34 PM - 30 Jul 2012
1,167
295

Twitter's Social Graph



Source: Volkova, Van Durme, Yarowsky, Bachrach
"Tutorial on Social Media Predictive Analytics" NAACL 2015

Twitter API

What is an API?

Application **P**rogramming **I**nterface

API is a set of protocols that specify how software programs communicate with each other.

What is an API?

> **Without API:**
> An app finds the current weather in London by opening http://www.weather.com/ and reading the webpage like a human does, interpreting the content.

> **With API:**
> An app finds the current weather in London by sending a message to the weather.com API (in a structured format like XML). The weather.com API then replies with a structured response.

Source: Chris Beach @ Quora

Twitter API

- Twitter is recognized for having one of the most open and powerful developer APIs of any major technology company.

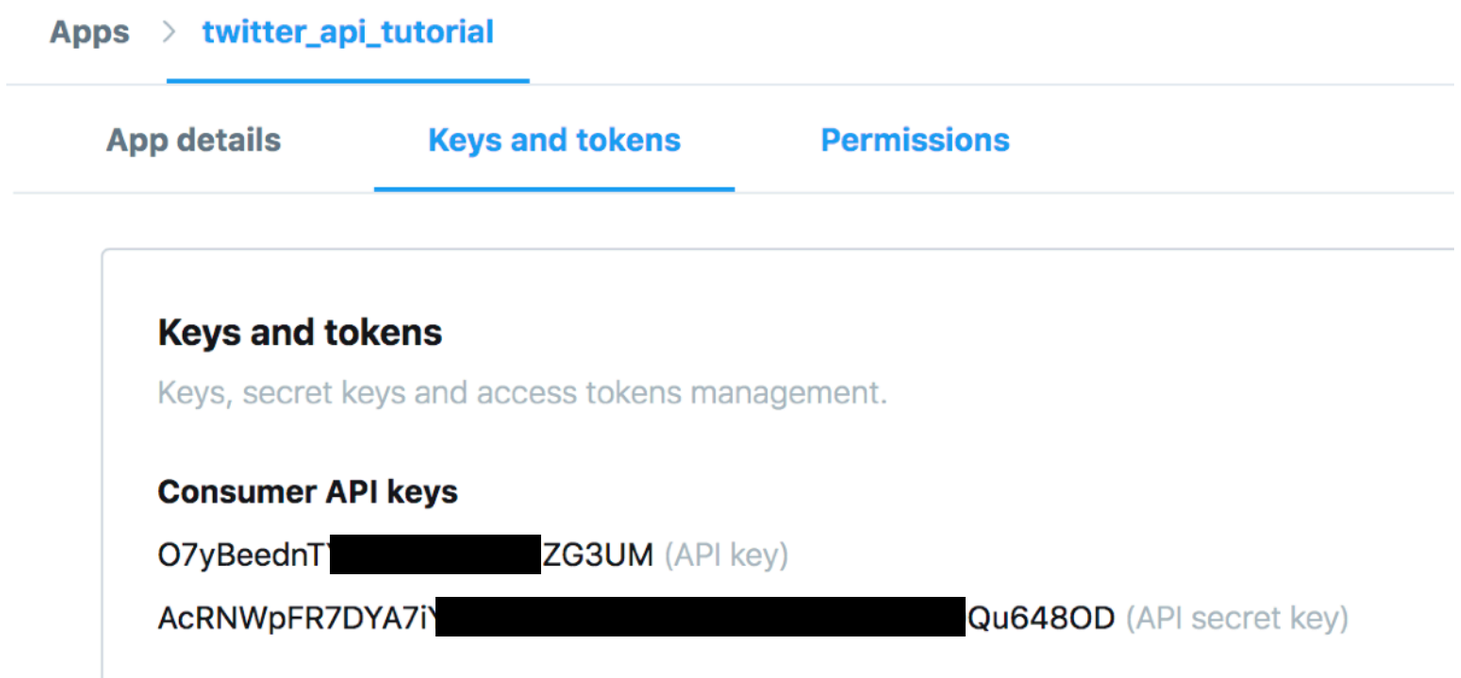
- The first version of its public API was released in September 2006.

Two Most Popular APIs

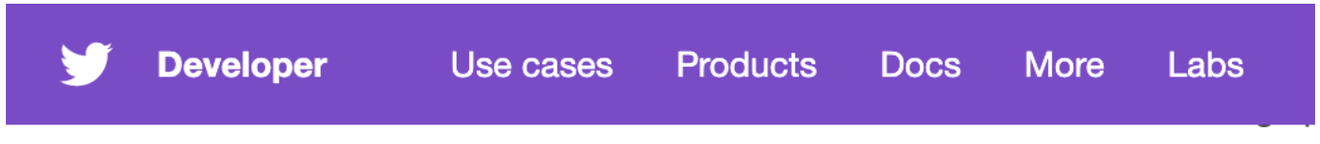
Streaming API	**REST API**
a sample of public tweets and events as they published on Twitter (can specify search terms or users or locations)	- search - trends - read author profile and follower data - post / modify
only real-time data	historical data up to a week
continuous net connection	one-time request
no limit	rate limit (varies for different requests)

OAuth

- Twitter uses OAuth to provide authorized access to its API.
- which means, to start with needs:
- a Twitter account
- OAuth access tokens from developer.twitter.com/



Python Twitter Tools

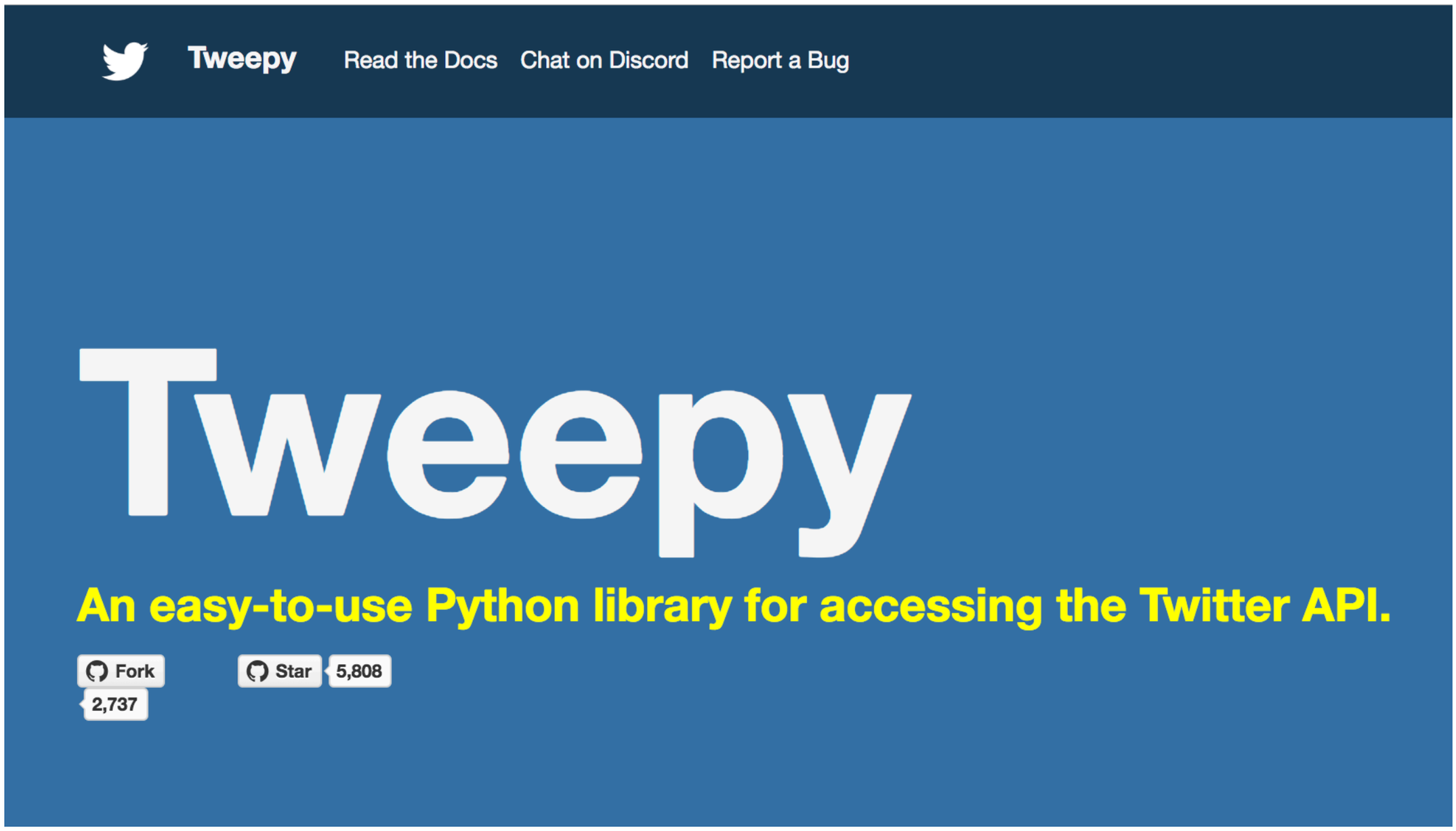


Python

- python-twitter *maintained by @bear* — this library provides a pure Python interface for the Twitter API (documentation)
- tweepy *maintained by @applepie & more* — a Python wrapper for the Twitter API (documentation) (examples)
- TweetPony *by @Mezgrman* — A Python library aimed at simplicity and flexibility.
- Python Twitter Tools *by @sixohsix* — An extensive Python library for interfacing to the Twitter REST and streaming APIs (v1.0 and v1.1). Also features a command line Twitter client. Supports Python 2.6, 2.7, and 3.3+. (documentation)
- twitter-gobject *by @tchx84* — Allows you to access Twitter's 1.1 REST API via a set of GObject based objects for easy integration with your GLib2 based code. (examples)
- TwitterSearch *by @crw_koepp* — Python-based interface to the 1.1 Search API.
- twython *by @ryanmcgrath* — Actively maintained, pure Python wrapper for the Twitter API. Supports both normal and streaming Twitter APIs. Supports all v1.1 endpoints, including dynamic functions so users can make use of endpoints not yet in the library. (docs)
- TwitterAPI *by @boxnumber03* — A REST and Streaming API wrapper that supports python 2.x and python 3.x, TwitterAPI also includes iterators for both API's that are useful for processing streaming results as well as paged results.
- Birdy *by @sect2k* — "a super awesome Twitter API client for Python"

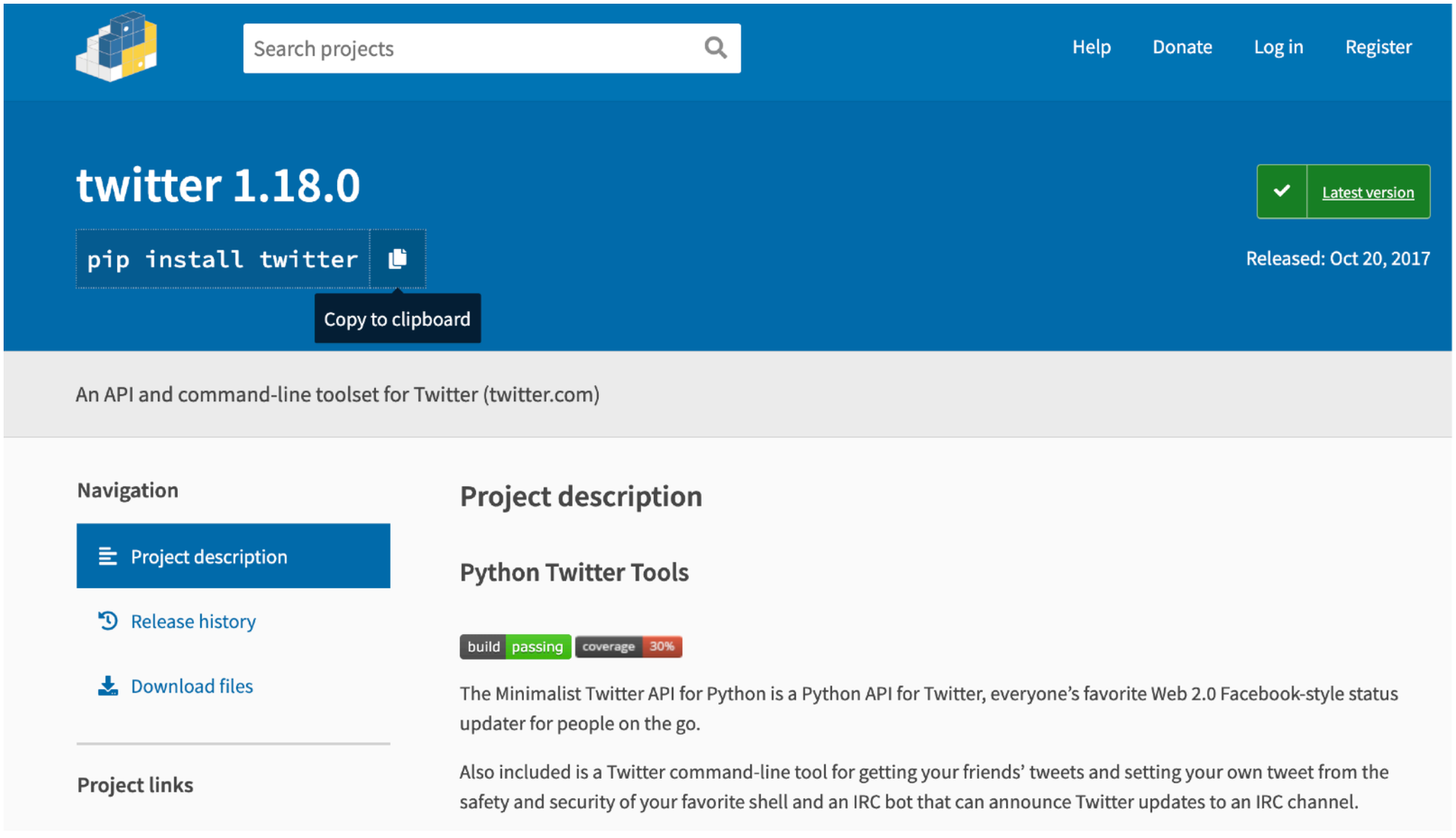
Python Twitter Tools

www.tweepy.org



Python Twitter Tools

https://pypi.org/project/twitter/



OAuth Authentication

Twitter uses OAuth to provide authorized access to the API.

```
[ ]  # Import the tweepy library
     import tweepy
     from tweepy.streaming import StreamListener

     # Variables that contains the user credentials to access Twitter API
     ACCESS_TOKEN = 'YOUR ACCESS TOKEN"'
     ACCESS_SECRET = 'YOUR ACCESS TOKEN SECRET'
     CONSUMER_KEY = 'YOUR API KEY'
     CONSUMER_SECRET = 'ENTER YOUR API SECRET'

     # Setup tweepy to authenticate with Twitter credentials:
     auth = tweepy.OAuthHandler(CONSUMER_KEY, CONSUMER_SECRET)
     auth.set_access_token(ACCESS_TOKEN, ACCESS_SECRET)
```

Streaming API

```
[ ]  # This is a basic listener that just prints received tweets to stdout.
     class StdOutListener(StreamListener):


         def on_data(self, data):
             print(data)
             return True

         #def on_status(self, status):
         #    print(json.dumps(status._json))
         #    return True

         def on_error(self, status_code):
             print(status_code)
             return False


     # tweepy.Stream.sample() will give a live stream (~1% sample) of all public tweets
     # Warning: it will continue to run indefinitely until you stop it.

     listener = StdOutListener()
     twitterStream = tweepy.Stream(auth, listener)
     twitterStream.sample()
```

JSON

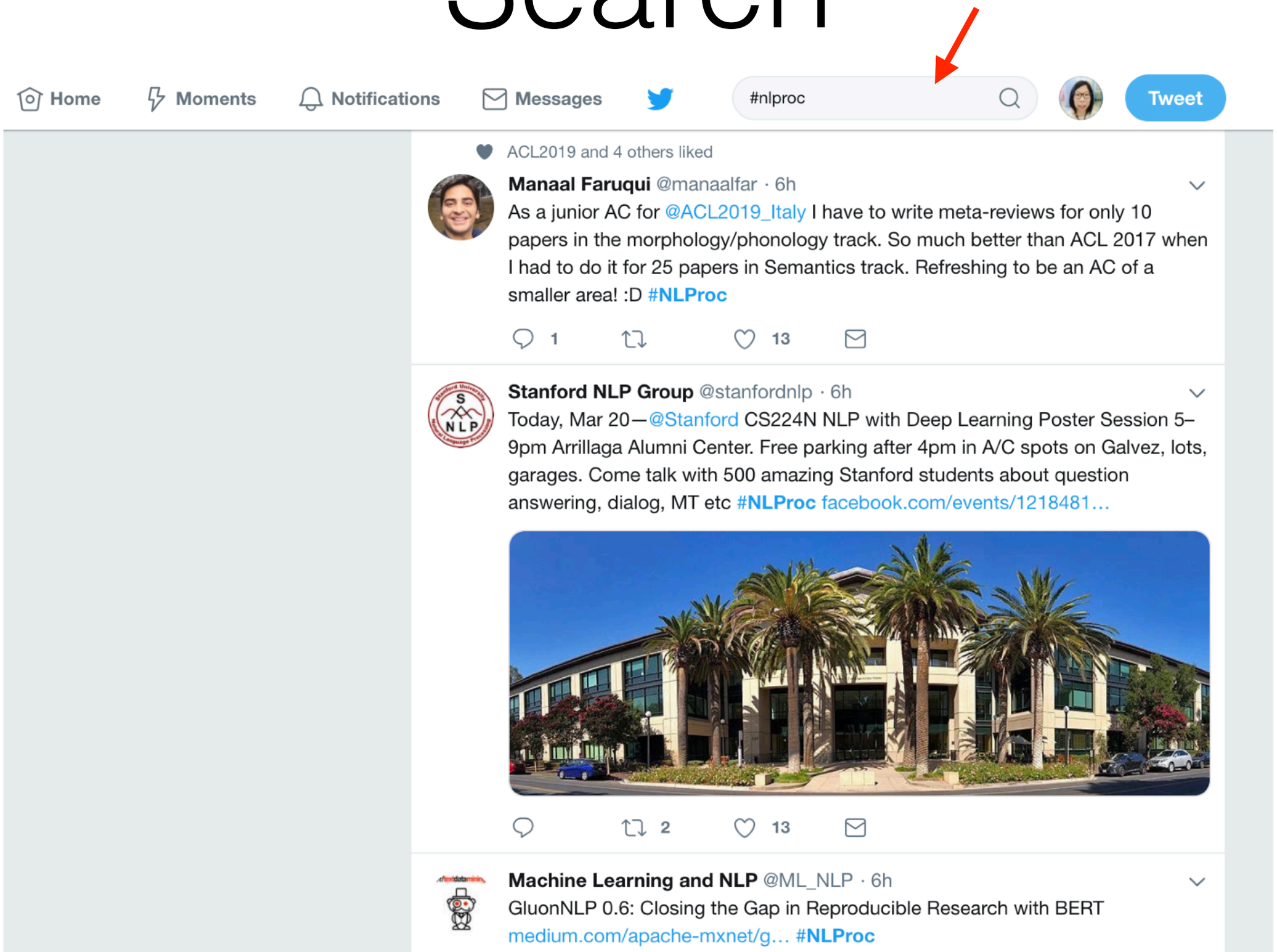
JavaScript Object Notation

JSON is a minimal, readable format for structuring data.

A Tweet in JSON


```
{
   "favorited": false,
   "contributors": null,
   "truncated": false,
   "text": "#CFP Workshop on Noisy User-generated Text at ACL - Beijing 31 July 2015. Papers due: 11 May 2015. http://t.co/rcygyEowqH   #NLProc #WNUT15",
   "possibly_sensitive": false,
   "in_reply_to_status_id": null,
   "user": {
      "follow_request_sent": null,
      "profile_use_background_image": true,
      "default_profile_image": false,
      "id": 237918251,
      "verified": false,
      "profile_image_url_https": "https://pbs.twimg.com/profile_images/527088456967544832/Dn
```

Search


Home
Moments
Notifications
Messages
#nlproc
Tweet
ACL2019 and 4 others liked
Manaal Faruqui @manaalfar · 6h
As a junior AC for @ACL2019_Italy I have to write meta-reviews for only 10 papers in the morphology/phonology track. So much better than ACL 2017 when I had to do it for 25 papers in Semantics track. Refreshing to be an AC of a smaller area! :D #NLProc
1
13
Stanford NLP Group @stanfordnlp · 6h
Today, Mar 20—@Stanford CS224N NLP with Deep Learning Poster Session 5–9pm Arrillaga Alumni Center. Free parking after 4pm in A/C spots on Galvez, lots, garages. Come talk with 500 amazing Stanford students about question answering, dialog, MT etc #NLProc facebook.com/events/1218481…
2
13
Machine Learning and NLP @ML_NLP · 6h
GluonNLP 0.6: Closing the Gap in Reproducible Research with BERT
medium.com/apache-mxnet/g… #NLProc

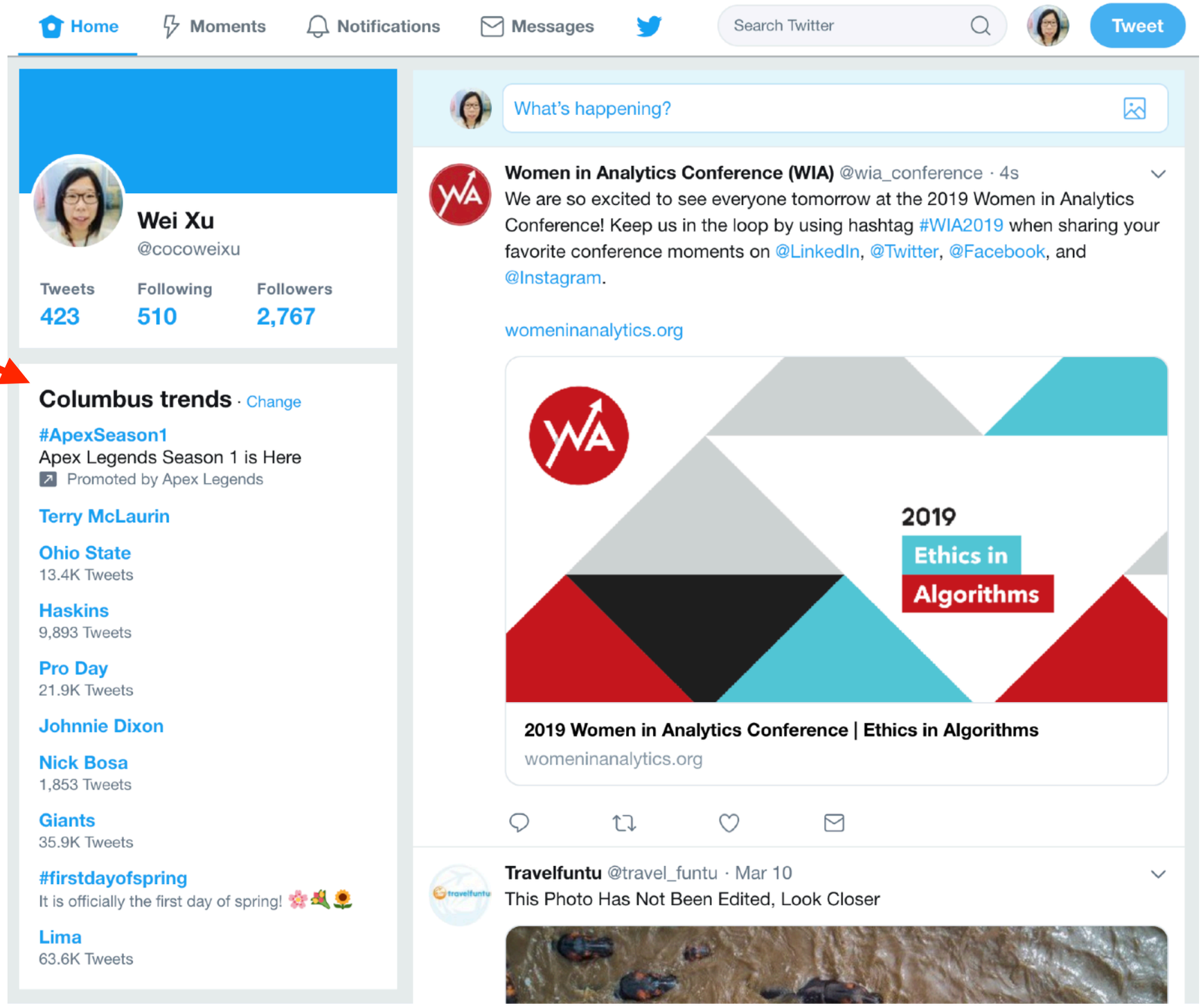
Search API

```
[ ]  # Search for latest tweets about "#nlproc"
     tweets = tweepy.Cursor(api.search, q='#nlproc')

     # Print out the latest 10 tweets that contain "#nlproc" hashtag
     for item in tweets.items(10):
       print(item._json)
```

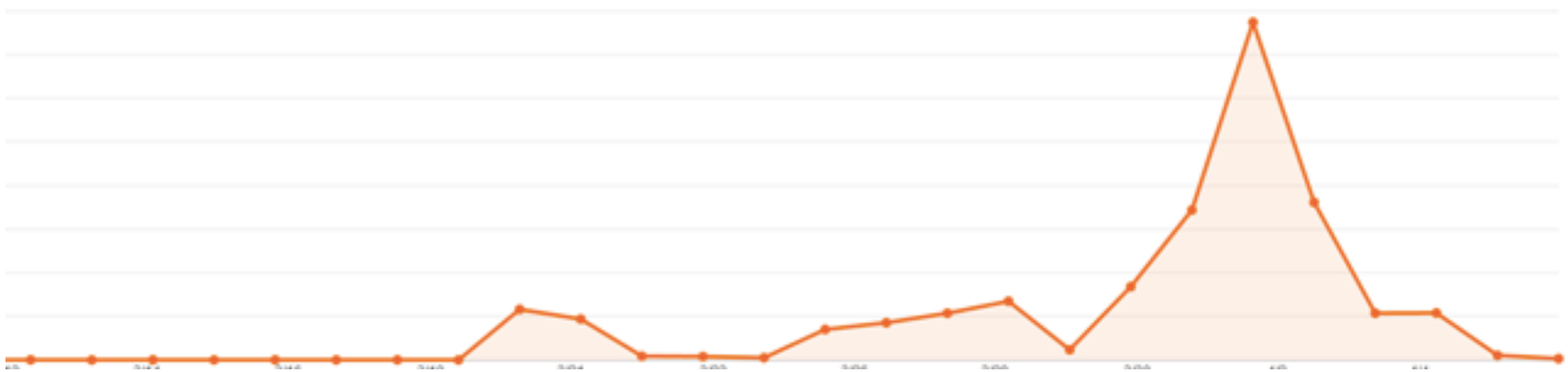

```
{'name': 'Dwayne Haskins', 'url': 'http://twitter.com/search?q=%22Dwayne+Haskins%22', 'promoted_content': None, 'query': '%22Dwayne+Haskins%22', 'twe
{'name': 'McCain', 'url': 'http://twitter.com/search?q=McCain', 'promoted_content': None, 'query': 'McCain', 'tweet_volume': 381900}
{'name': 'Lima', 'url': 'http://twitter.com/search?q=Lima', 'promoted_content': None, 'query': 'Lima', 'tweet_volume': 70085}
{'name': '#firstdayofspring', 'url': 'http://twitter.com/search?q=%23firstdayofspring', 'promoted_content': None, 'query': '%23firstdayofspring', 'tw
{'name': 'Daniel Caesar', 'url': 'http://twitter.com/search?q=%22Daniel+Caesar%22', 'promoted_content': None, 'query': '%22Daniel+Caesar%22', 'tweet_
{'name': '#InternationalDayOfHappiness', 'url': 'http://twitter.com/search?q=%23InternationalDayOfHappiness', 'promoted_content': None, 'query': '%2
{'name': 'AirPods', 'url': 'http://twitter.com/search?q=AirPods', 'promoted_content': None, 'query': 'AirPods', 'tweet_volume': 131265}
{'name': 'Pro Day', 'url': 'http://twitter.com/search?q=%22Pro+Day%22', 'promoted_content': None, 'query': '%22Pro+Day%22', 'tweet_volume': 19808}
{'name': '#SpringEquinox', 'url': 'http://twitter.com/search?q=%23SpringEquinox', 'promoted_content': None, 'query': '%23SpringEquinox', 'tweet_volu
{'name': 'Flume', 'url': 'http://twitter.com/search?q=Flume', 'promoted_content': None, 'query': 'Flume', 'tweet_volume': 37576}
{'name': '#HappinessInOneWord', 'url': 'http://twitter.com/search?q=%23HappinessInOneWord', 'promoted_content': None, 'query': '%23HappinessInOneWor
{'name': '#StrangerThings3', 'url': 'http://twitter.com/search?q=%23StrangerThings3', 'promoted_content': None, 'query': '%23StrangerThings3', 'twee
{'name': 'Eloy', 'url': 'http://twitter.com/search?q=Eloy', 'promoted_content': None, 'query': 'Eloy', 'tweet_volume': None}
{'name': 'Happy Spring', 'url': 'http://twitter.com/search?q=%22Happy+Spring%22', 'promoted_content': None, 'query': '%22Happy+Spring%22', 'tweet_vo
{'name': 'Bill & Ted 3', 'url': 'http://twitter.com/search?q=%22Bill+%26+Ted+3%22', 'promoted_content': None, 'query': '%22Bill+%26+Ted+3%22', 'twee
```

Trends


Home
Moments
Notifications
Messages
Search Twitter
Tweet
Wei Xu
@cocoweixu
Tweets
423
Following
510
Followers
2,767
Columbus trends · Change
#ApexSeason1
Apex Legends Season 1 is Here
Promoted by Apex Legends
Terry McLaurin
Ohio State
13.4K Tweets
Haskins
9,893 Tweets
Pro Day
21.9K Tweets
Johnnie Dixon
Nick Bosa
1,853 Tweets
Giants
35.9K Tweets
#firstdayofspring
It is officially the first day of spring!
Lima
63.6K Tweets
What's happening?
Women in Analytics Conference (WIA) @wia_conference · 4s
We are so excited to see everyone tomorrow at the 2019 Women in Analytics Conference! Keep us in the loop by using hashtag #WIA2019 when sharing your favorite conference moments on @LinkedIn, @Twitter, @Facebook, and @Instagram.
womeninanalytics.org
2019
Ethics in
Algorithms
2019 Women in Analytics Conference | Ethics in Algorithms
womeninanalytics.org
Travelfuntu @travel_funtu · Mar 10
This Photo Has Not Been Edited, Look Closer

Trends

trending topics are determined by an unpublished algorithm, which finds words, phrases and hashtags that have had a sharp increase in popularity, as opposed to overall volume.



Trends API

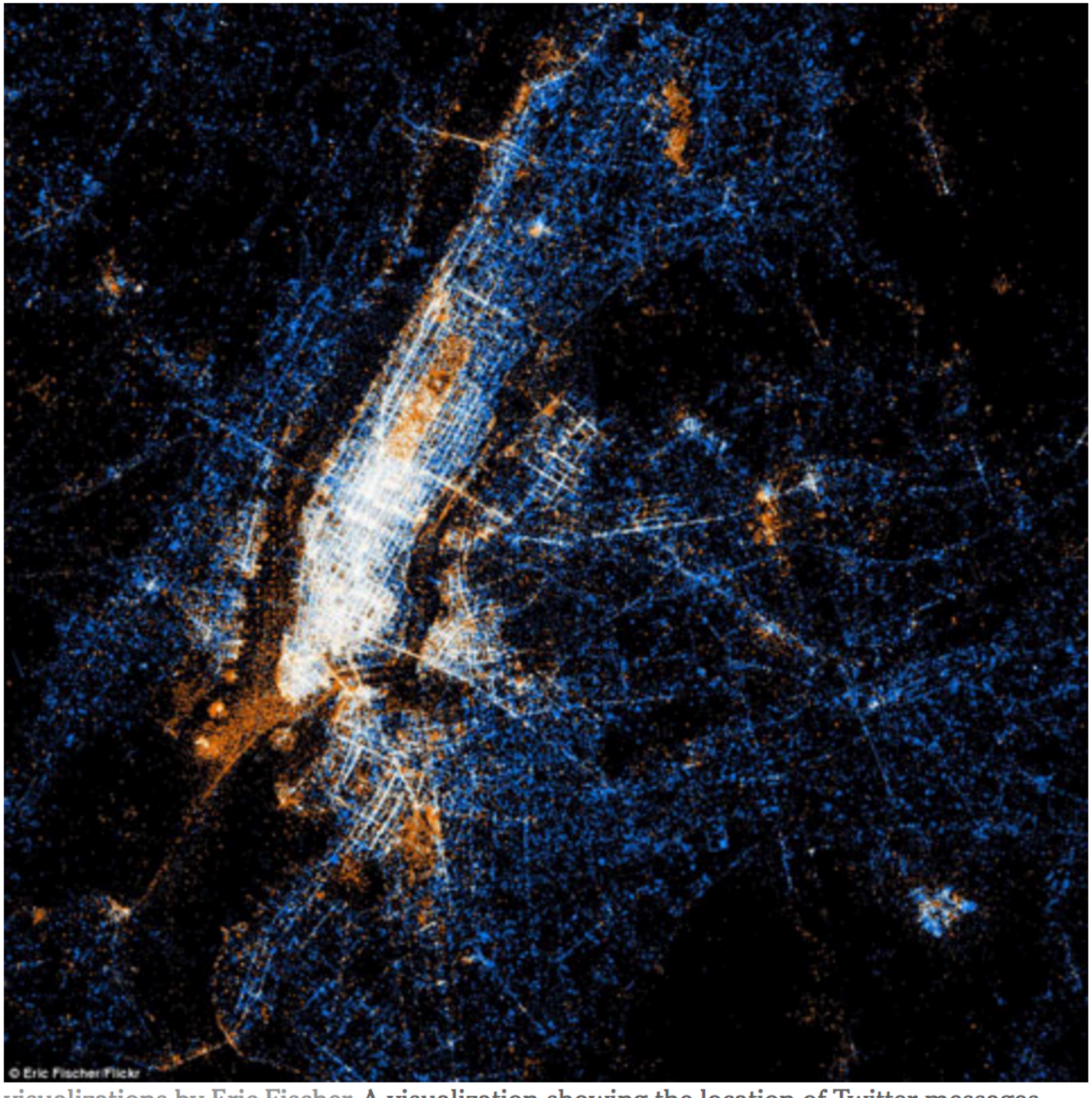
Where On Earth ID

```
# Where On Earth ID for Columbus, Ohio is 2383660.
COLUMBUS_WOE_ID = 2383660

columbus_trends = api.trends_place(COLUMBUS_WOE_ID)

trends = json.loads(json.dumps(columbus_trends, indent=1))

for trend in trends[0]["trends"]:
  print (trend)
```

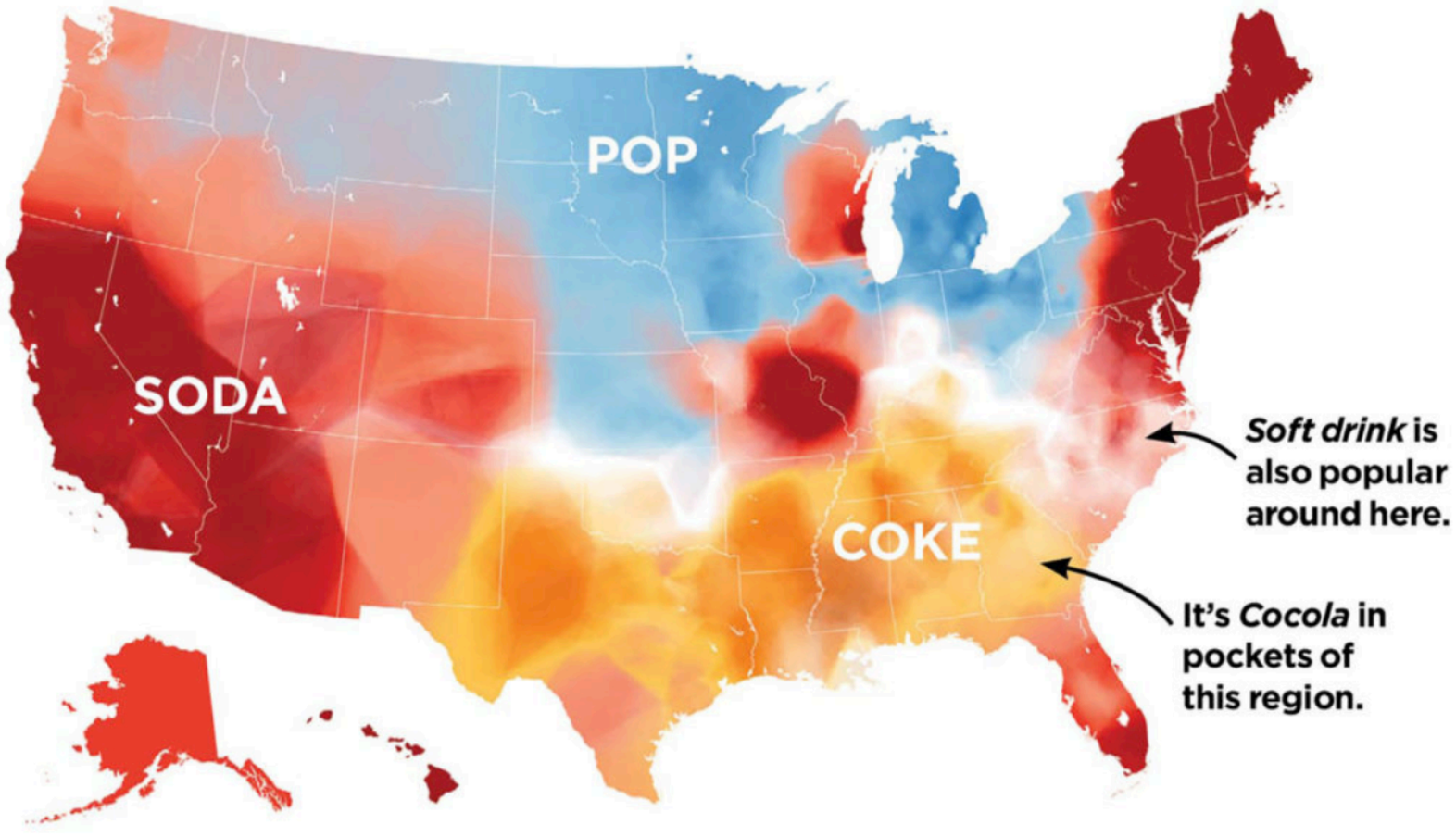


visualizations by Eric Fischer A visualization showing the location of Twitter messages and Flickr photos in New York City.

What do you call carbonated beverages?

- Soda
- Pop
- Coke
- other ways?

What do you call carbonated beverages?



JOSH KATZ

What do you call a sale of household items

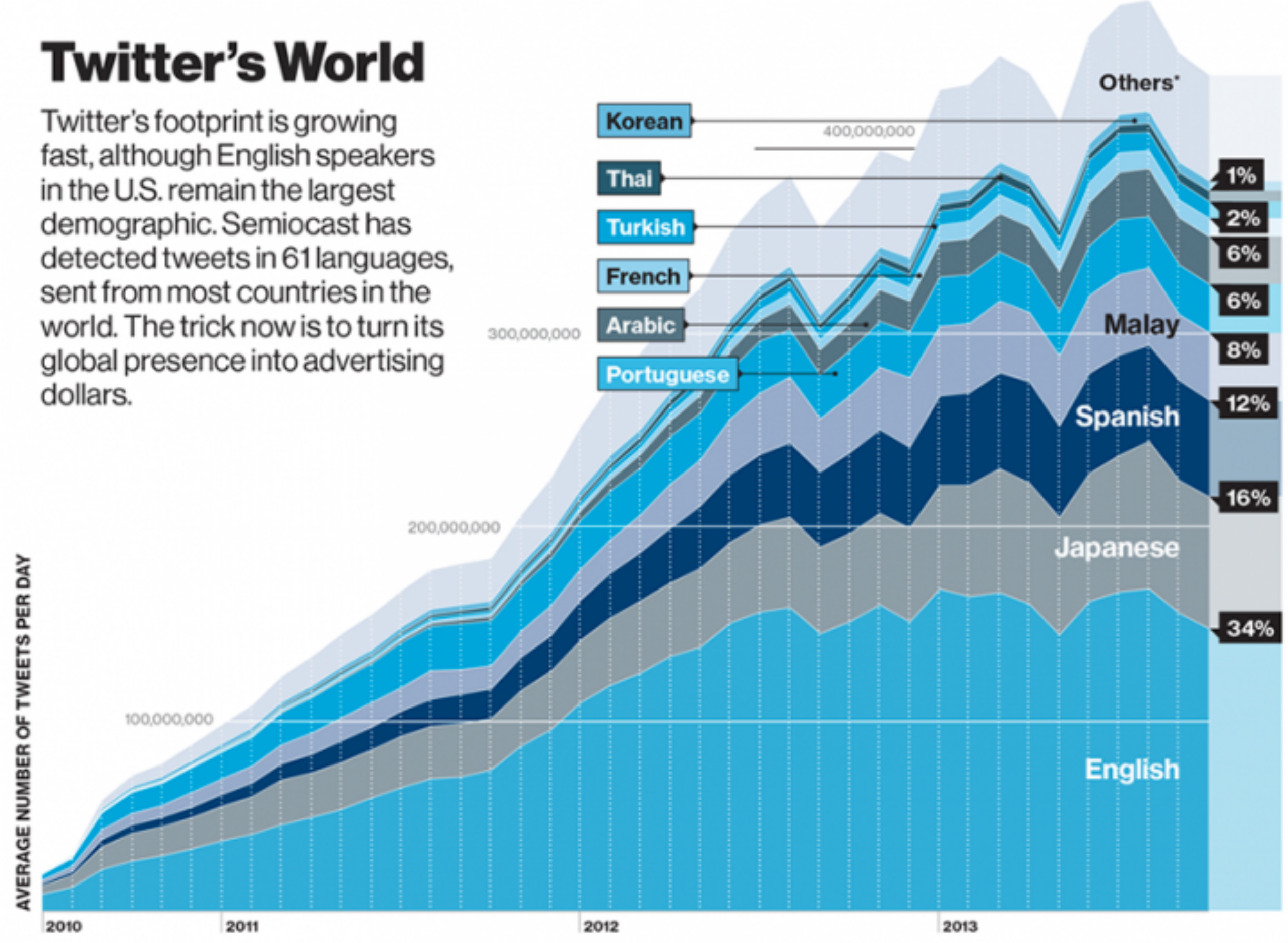
- Garage Sale
- Yard Sale
- other ways?

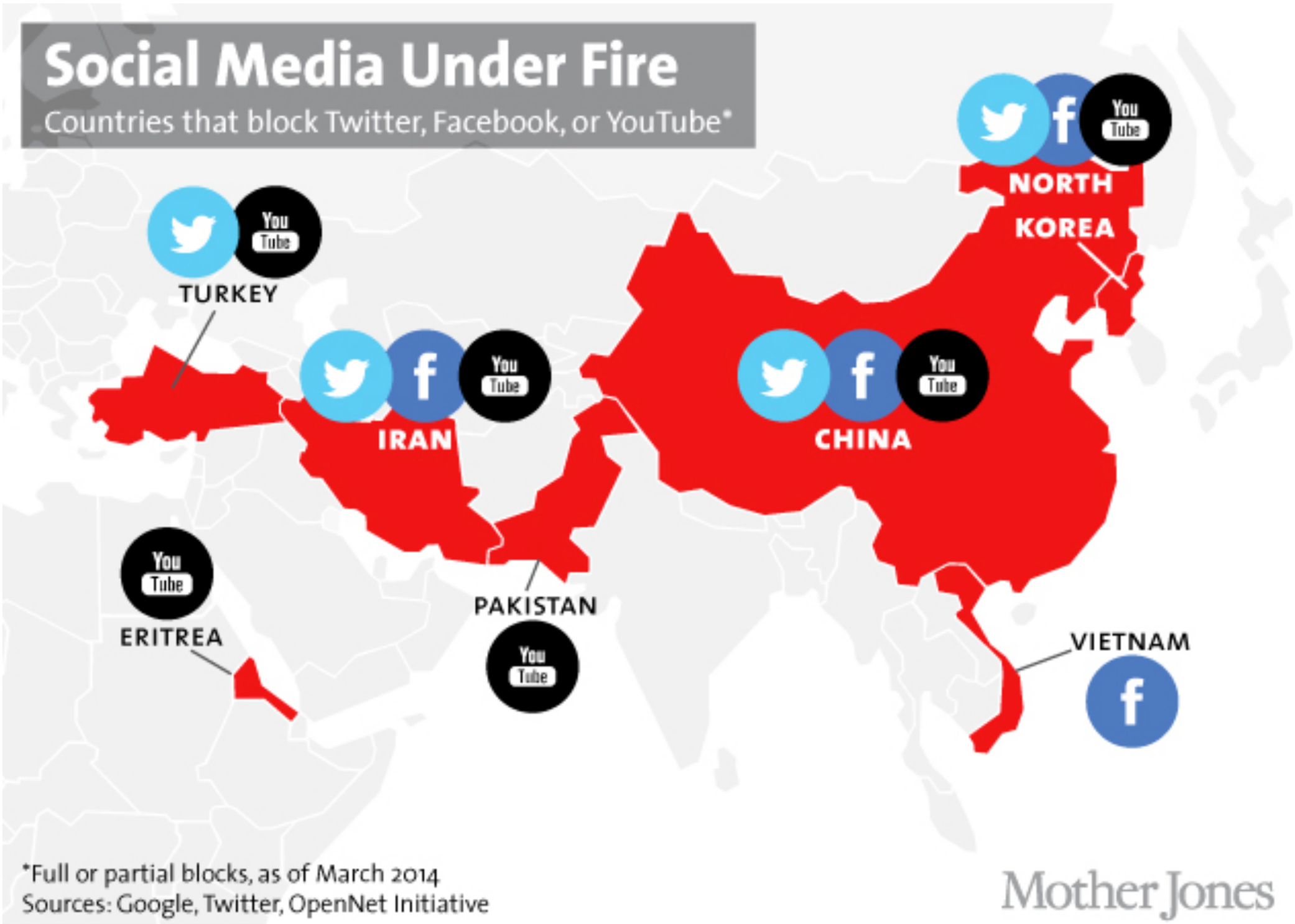
What do you call a sale of household items

JOSH KATZ

Twitter's World

Twitter's footprint is growing fast, although English speakers in the U.S. remain the largest demographic. Semiocast has detected tweets in 61 languages, sent from most countries in the world. The trick now is to turn its global presence into advertising dollars.



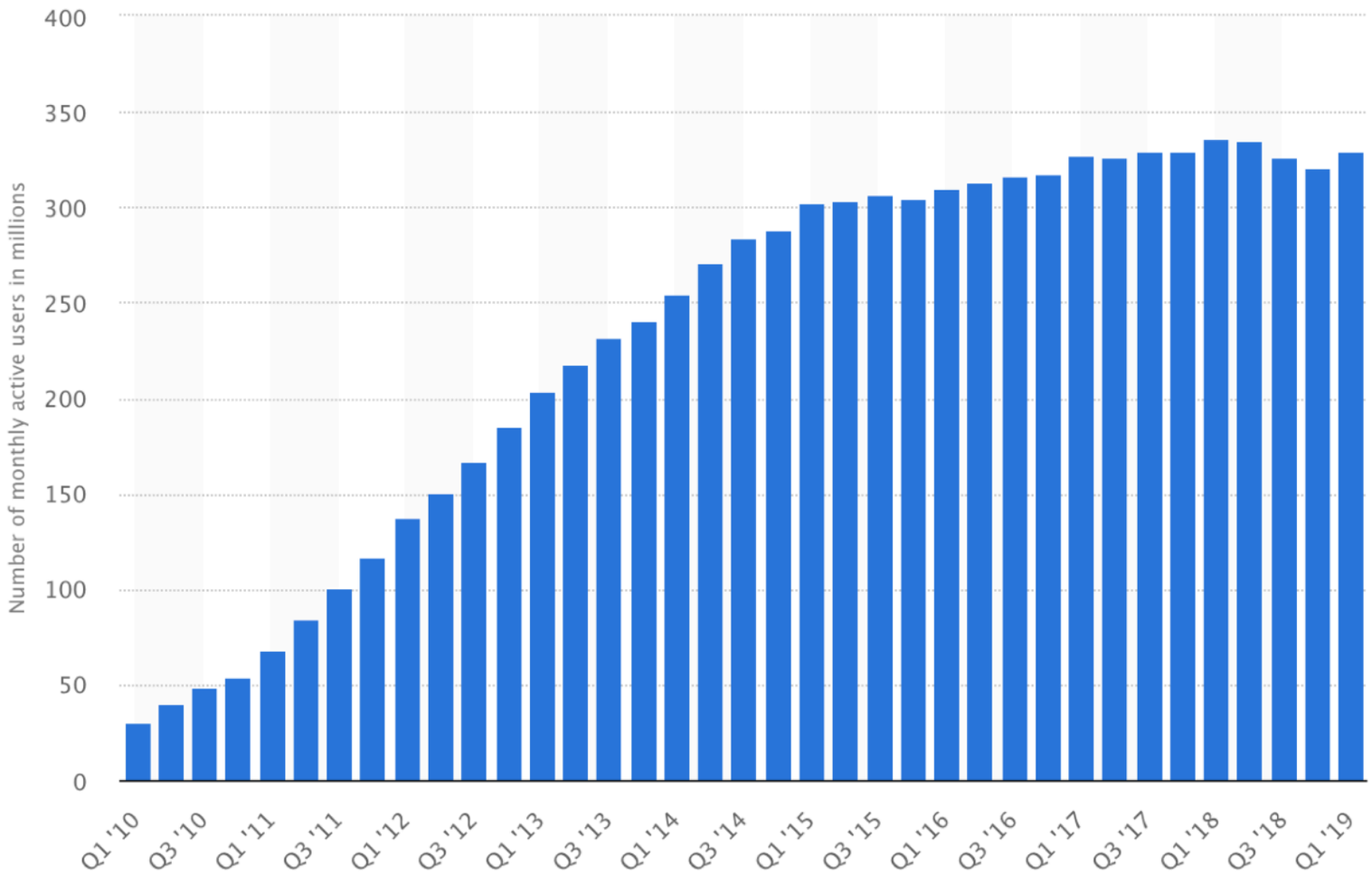


新浪微博
weibo.com

known as the “Chinese Twitter”
120 Million Posts / Day

Twitter Demographics

- As of the 1st quarter of 2019, Twitter averaged 330 million monthly active users, and 139 million monetizable daily active Twitter users worldwide.



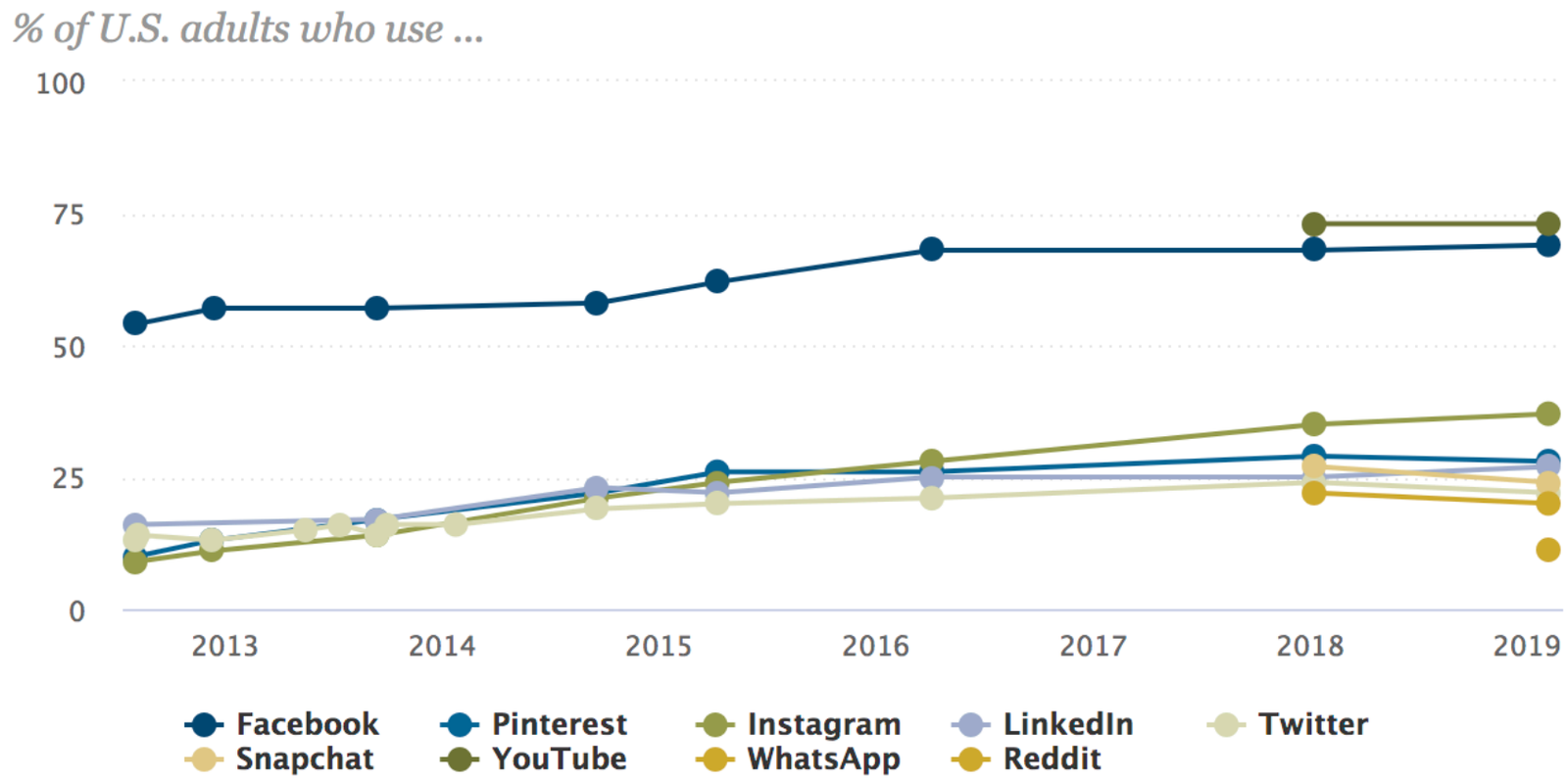
Twitter Demographics

- About 80% (262 million) of all monthly active Twitter users live outside the United States.

- 72 million monthly active users live within the United States.

- The top countries on Twitter outside the U.S. are Japan (50.9 million users), the United Kingdom (18.6 million users), and Saudi Arabia (13.8 million users).

The Most Popular Social Media Platforms



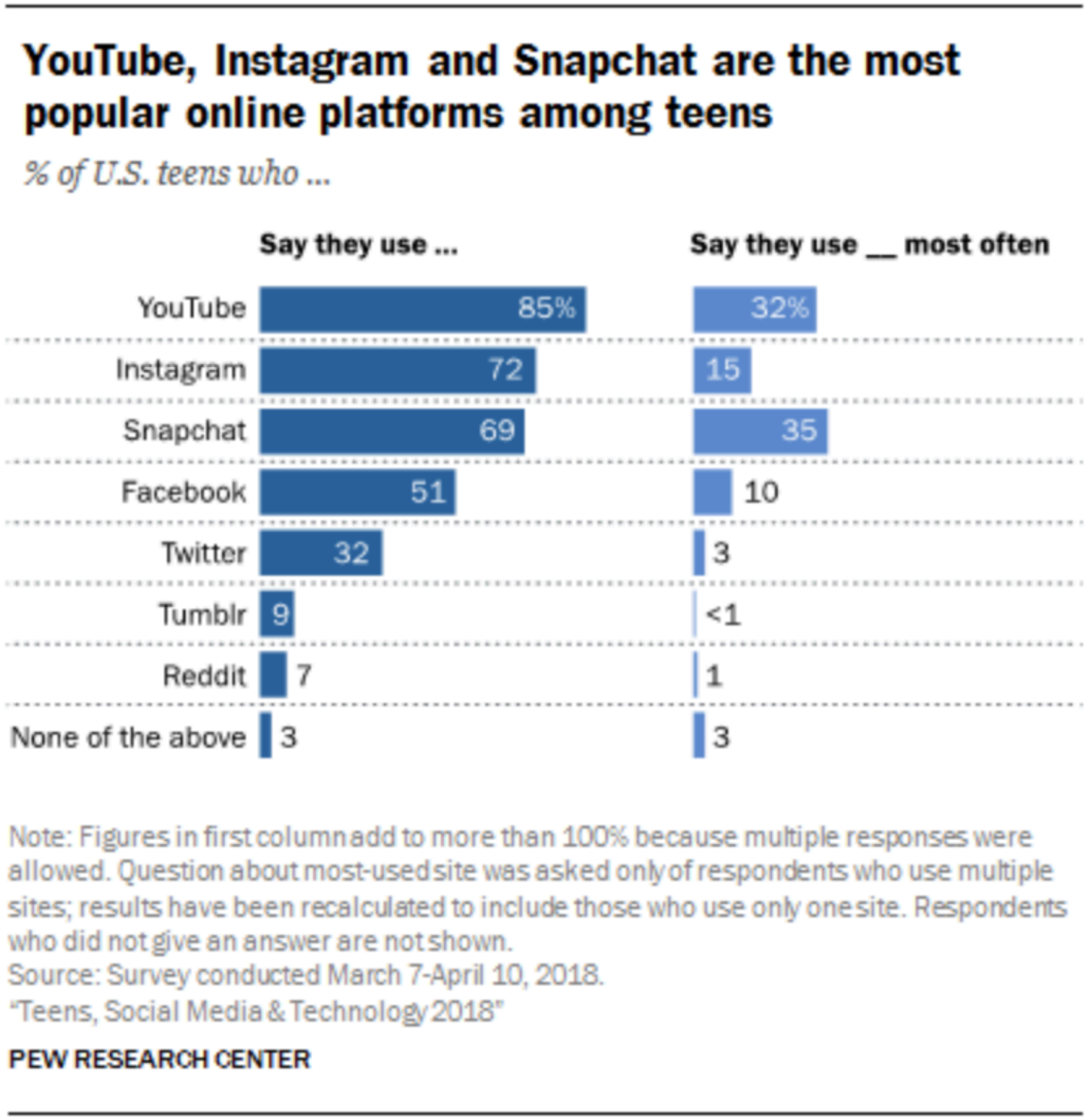
Source: Surveys conducted 2012-2019.

PEW RESEARCH CENTER

https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/

The Most Popular Social Media Platforms

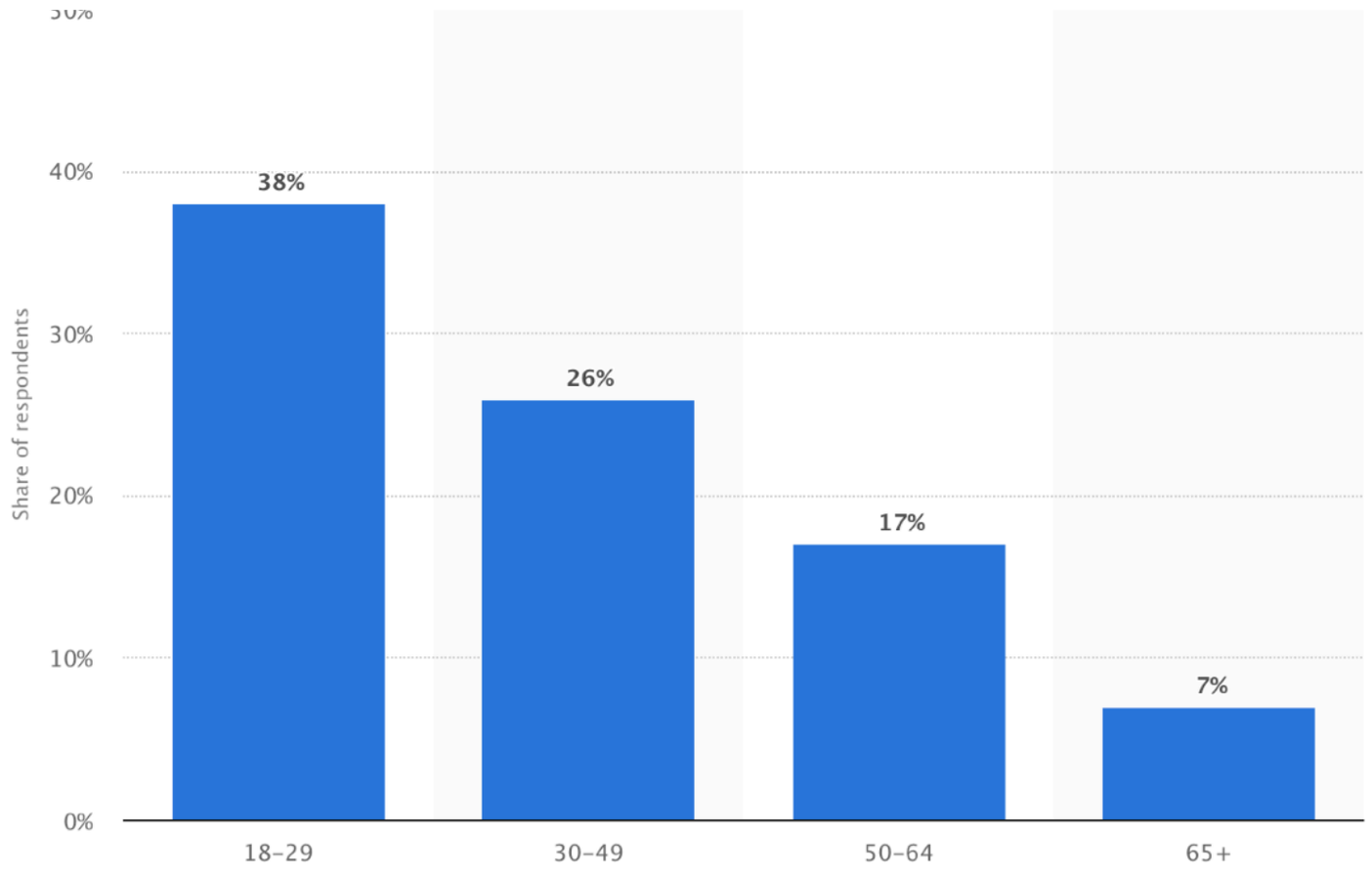
- 95% of teens (age 13-17) now report they have a smartphone or access to one. 45% of teens now say they are online on a near-constant basis.



https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/

Twitter Demographics

- How old are Twitter users? Not very old, it seems. But also not very young.



https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/

Twitter Demographics

- Globally, more men use Twitter than women. But in the United States, more women use Twitter than men.

% of U.S. adults who say they use ...

	Facebook	YouTube	Pinterest	Instagram	Snapchat	LinkedIn	Twitter	WhatsApp
Total	68%	73%	29%	35%	27%	25%	24%	22%
Men	62	75	16	30	23	25	23	20
Women	74	72	41	39	31	25	24	24
White	67	71	32	32	24	26	24	14
Black	70	76	23	43	36	28	26	21
Hispanic	73	78	23	38	31	13	20	49
Ages 18-29	81	91	34	64	68	29	40	27
18-24	80	94	31	71	78	25	45	25
25-29	82	88	39	54	54	34	33	31
30-49	78	85	34	40	26	33	27	32
50-64	65	68	26	21	10	24	19	17
65+	41	40	16	10	3	9	8	6

Note: Whites and blacks include only non-Hispanics. Hispanics are of any race.
Source: Survey conducted Jan. 3-10. 2018
"Social Media Use in 2018"

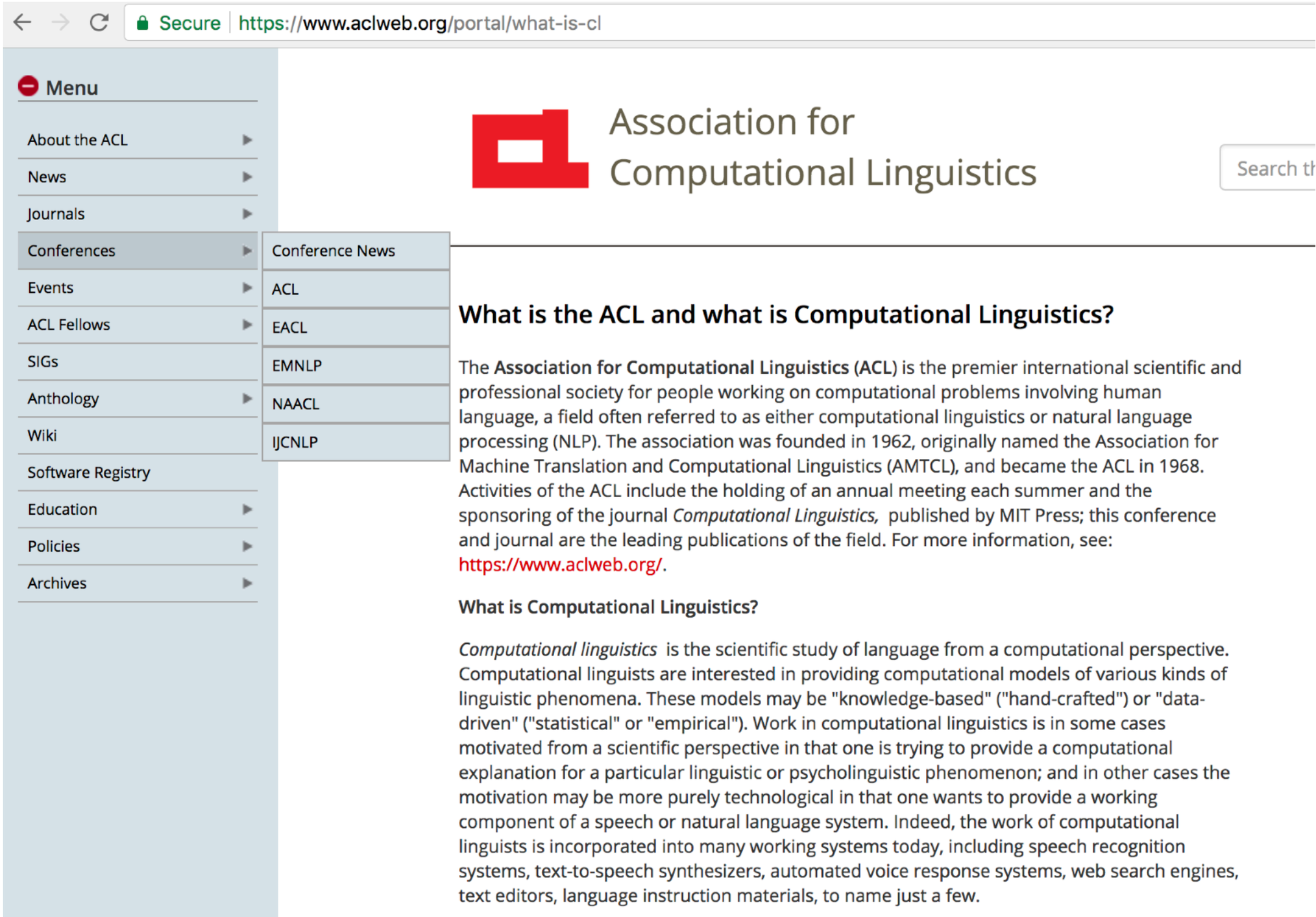
PEW RESEARCH CENTER

The Natural Language Processing Research Community

a.k.a.

- Natural Language Processing (NLP)
- Text Analysis
- Computational Linguistics

ACL


Secure
https://www.aclweb.org/portal/what-is-cl
Menu
About the ACL
News
Journals
Conferences
Events
ACL Fellows
SIGs
Anthology
Wiki
Software Registry
Education
Policies
Archives
Conference News
ACL
EACL
EMNLP
NAACL
IJCNLP
Association for
Computational Linguistics
Search th
What is the ACL and what is Computational Linguistics?
The Association for Computational Linguistics (ACL) is the premier international scientific and professional society for people working on computational problems involving human language, a field often referred to as either computational linguistics or natural language processing (NLP). The association was founded in 1962, originally named the Association for Machine Translation and Computational Linguistics (AMTCL), and became the ACL in 1968. Activities of the ACL include the holding of an annual meeting each summer and the sponsoring of the journal Computational Linguistics, published by MIT Press; this conference and journal are the leading publications of the field. For more information, see: https://www.aclweb.org/.
What is Computational Linguistics?
Computational linguistics is the scientific study of language from a computational perspective. Computational linguists are interested in providing computational models of various kinds of linguistic phenomena. These models may be "knowledge-based" ("hand-crafted") or "data-driven" ("statistical" or "empirical"). Work in computational linguistics is in some cases motivated from a scientific perspective in that one is trying to provide a computational explanation for a particular linguistic or psycholinguistic phenomenon; and in other cases the motivation may be more purely technological in that one wants to provide a working component of a speech or natural language system. Indeed, the work of computational linguists is incorporated into many working systems today, including speech recognition systems, text-to-speech synthesizers, automated voice response systems, web search engines, text editors, language instruction materials, to name just a few.

NLP Publications

- top NLP-specific venues:
  - ACL, NAACL, EACL, EMNLP, COLING (conference)
  - TACL (journal+conference model)
  - CL (journal)

- other venues:
  - NLP: CoNLL, *Sem, WMT, LREC, IJNLP, Workshops …
  - related CS fields: WWW, KDD, AAAI, WSDM, NIPS, ICWSM, CIKM, ICML …
  - related non-CS fields: psychology, linguistics, …

NLP Publications

- ACL Anthology (http://aclweb.org/anthology/)

  all NLP conference and journal papers (free!)

ACL Anthology

Search...

Welcome to the ACL Anthology!

The ACL Anthology currently hosts 52251 papers on the study of computational linguistics and natural language processing.

Subscribe to the mailing list to receive announcements and updates to the Anthology.

The Anthology can archive your poster or presentation! Please submit them in PDF format by **filling out this form**. Attachments will be distributed under the terms of the **CC-BY-4.0 license**.

Full Anthology as BibTeX (7.16 MB)

Give feedback

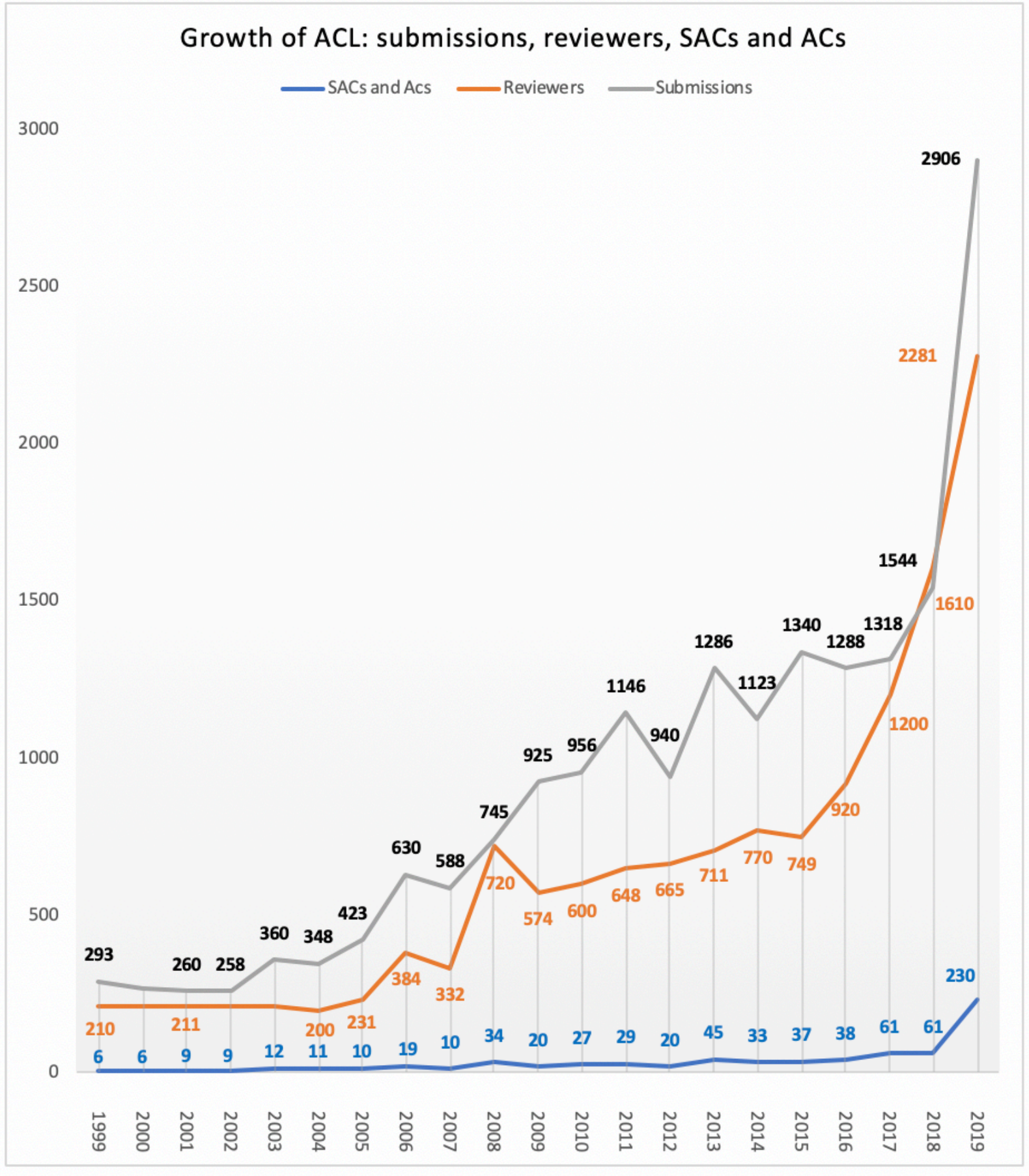
ACL Events

Venue	Present – 2010										2009 – 2000										1999 – 1990										1989 and older										
ACL	19	18	17	16	15	14	13	12	11	10	09	08	07	06	05	04	03	02	01	00	99	98	97	96	95	94	93	92	91	90	89	88	87	86	85	84	83	82	81	80	79	
ANLP																				00			97			94		92				88					83					
CL	19	18	17	16	15	14	13	12	11	10	09	08	07	06	05	04	03	02	01	00	99	98	97	96	95	94	93	92	91	90	89	88	87	86	85	84	83	82	81	80		78
CoNLL		18	17	16	15	14	13	12	11	10	09	08	07	06	05	04	03	02	01	00	99	98	97																			
EACL			17			14		12			09			06			03				99		97		95		93		91		89		87		85		83					
EMNLP		18	17	16	15	14	13	12	11	10	09	08	07	06	05	04	03	02	01	00	99	98	97	96																		
NAACL	19	18		16	15		13	12		10	09		07	06		04	03		01	00																						
***SEMEVAL**	19	18	17	16	15	14	13	12		10			07			04			01			98																					
TACL	19	18	17	16	15	14	13																																			

Conference Rotation

- ACL (and/or NAACL, EACL), EMNLP / COLING

Growth of ACL



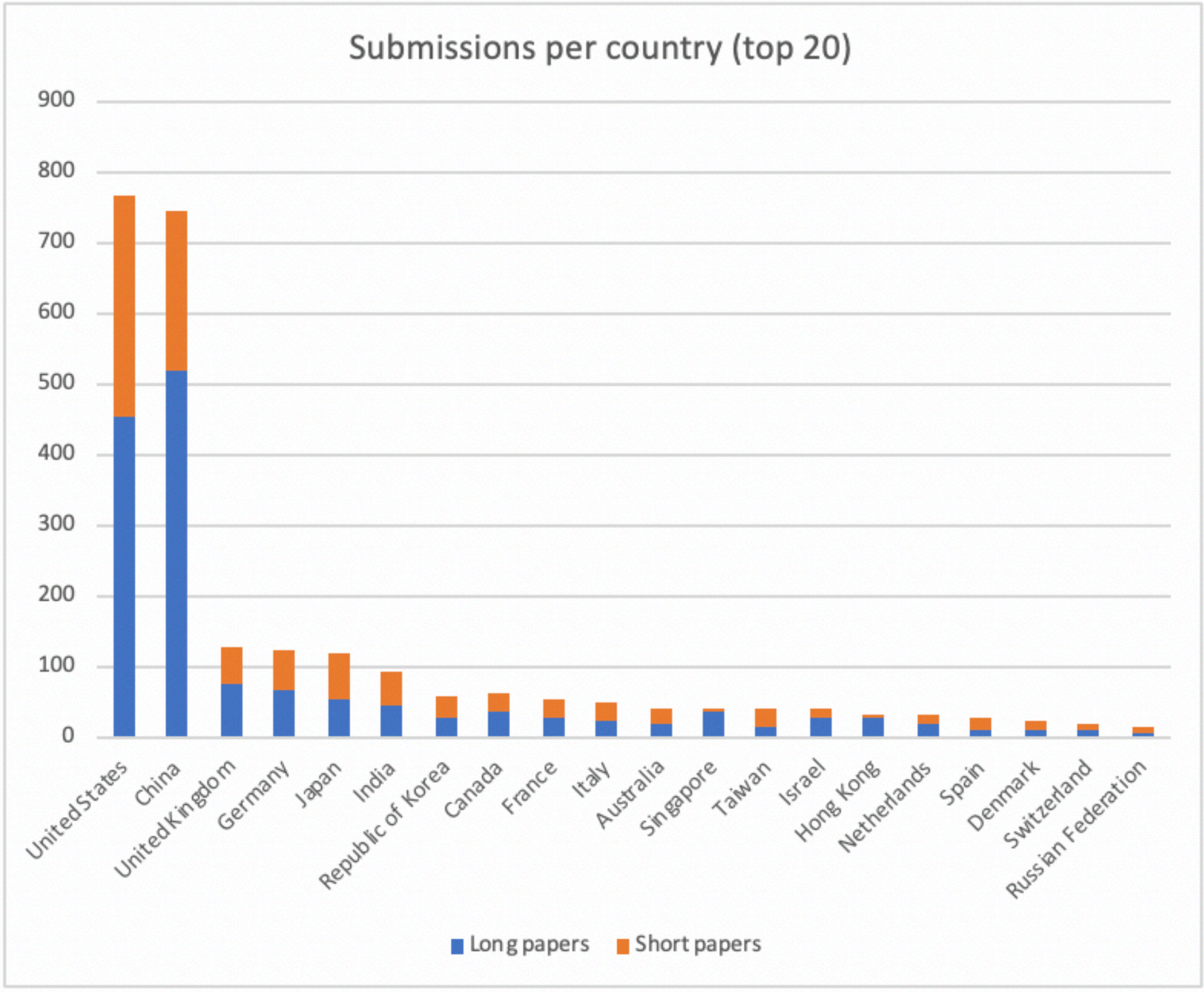
ACL'19 at A Glance

- The Annual Meeting of the Association for Computational Linguistics
- Duration:
  - tutorials (1 day)
  - main conference (3 days)
  - workshops (2 days)
- Attendance of 2000+ people
- Papers:
  - 2,905 submissions
  - 447 long papers and 213 short papers accepted
  - + ?? TACL papers
  - 151 oral and 151 poster presentations

Research Areas

	Area	Long submissions	Accepts	Accept rate (%)
1.	Applications	65	14	28.8
2.	Dialogue and Interactive Systems	126	38	30.2
3.	Discourse and Pragmatics	33	7	21.2
4.	Document Analysis	48	8	16.7
5.	Generation	96	32	33.3
6.	Information Extraction and Text Mining	155	37	23.9
7.	Linguistic Theories, Cognitive Modeling and Psycholinguistics	39	9	23.1
8.	Machine Learning	148	38	25.7
8.	Machine Translation	102	27	26.5
10.	Multidisciplinary and Area Chair COI	69	21	30.4
11.	Multilinguality	43	11	25.6
12.	Phonology Morphology and Word Segmentation	26	7	26.9
13.	Question Answering	99	32	32.3
14.	Resources and Evaluation	70	26	37.1
15.	Sentence-level semantics	69	14	20.3
15.	Sentiment Analysis and Argument Mining	91	24	26.4
17.	Social Media	51	14	27.5
18.	Summarization	48	11	22.9
19.	Tagging Chunking Syntax and Parsing	50	17	34.0
20.	Textual Inference and Other Areas of Semantics	44	16	36.4
21.	Vision Robotics Multimodal Grounding and Speech	56	20	35.7
22.	Word-level Semantics	78	20	25.6

By Country


Submissions per country (top 20)
900
800
700
600
500
400
300
200
100
0
United States
China
United Kingdom
Germany
Japan
India
Republic of Korea
Canada
France
Italy
Australia
Singapore
Taiwan
Israel
Hong Kong
Netherlands
Spain
Denmark
Switzerland
Russian Federation
Long papers
Short papers

By Country

	All submissions			Long submissions			Short submissions		
Country or Region	Sub.	Acc.	Rate (%)	Sub.	Acc.	Rate (%)	Sub.	Acc.	Rate (%)
Australia	46	11	23.9	22	4	18.2	24	7	29.2
Austria	5	0	0.0	2	0	0.0	3	0	0.0
Belgium	8	1	12.5	3	1	33.3	5	0	0.0
Brazil	11	0	0.0	6	0	0.0	5	0	0.0
Canada	74	16	21.6	44	12	27.3	30	4	13.3
Chile	2	0	0.0	2	0	0.0	0	0	N/A
China	817	155	19.0	567	118	20.8	250	37	14.8
Czech Republic	12	2	16.7	5	0	0.0	7	2	28.6
Denmark	25	4	16.0	11	1	9.1	14	3	21.4
Egypt	2	0	0.0	1	0	0.0	1	0	0.0
Estonia	2	0	0.0	2	0	0.0	0	0	N/A
Finland	6	0	0.0	2	0	0.0	4	0	0.0
France	60	11	18.3	32	4	12.5	28	7	25.0
Germany	136	39	28.7	73	26	35.6	63	13	20.6
Greece	7	4	57.1	1	1	100.0	6	3	50.0
Hong Kong	34	10	29.4	26	9	34.6	8	1	12.5
Hungary	7	1	14.3	3	1	33.3	4	0	0.0
India	107	18	16.8	54	16	29.6	53	2	3.8
Iran	3	0	0.0	2	0	0.0	1	0	0.0
Ireland	10	1	10.0	4	1	25.0	6	0	0.0
Israel	41	14	34.1	30	11	36.7	11	3	27.3
Italy	50	6	12.0	25	3	12.0	25	3	12.0
Japan	125	23	18.4	58	13	22.4	67	10	14.9
Luxembourg	2	0	0.0	2	0	0.0	0	0	N/A
Macau	5	1	20.0	3	1	33.3	2	0	0.0
Malta	2	0	0.0	0	0	N/A	2	0	0.0
Mexico	2	0	0.0	0	0	N/A	2	0	0.0
Netherlands	36	9	25.0	22	8	36.4	14	1	7.1
Norway	6	2	33.3	4	1	25.0	2	1	50.0
Pakistan	2	0	0.0	1	0	0.0	1	0	0.0
Peru	2	0	0.0	1	0	0.0	1	0	0.0
Poland	7	1	14.3	5	1	20.0	2	0	0.0
Portugal	8	3	37.5	4	2	50.0	4	1	25.0
Qatar	4	0	0.0	2	0	0.0	2	0	0.0
Republic of Korea	72	7	9.7	36	4	11.1	36	3	8.3
Romania	2	1	50.0	2	1	50.0	0	0	N/A
Russian Federation	14	4	28.6	7	2	28.6	7	2	28.6
Singapore	46	16	34.8	39	13	33.3	7	3	42.9
Slovakia	2	0	0.0	1	0	0.0	1	0	0.0
South Africa	2	1	50.0	1	0	0.0	1	1	100
Spain	29	6	20.7	12	1	8.3	17	5	29.4
Sri Lanka	5	0	0.0	1	0	0.0	4	0	0.0
Sweden	9	0	0.0	4	0	0.0	5	0	0.0
Switzerland	23	4	17.4	12	2	16.7	11	2	18.2
Taiwan	46	6	13.0	18	3	16.7	28	3	10.7
Thailand	2	0	0.0	1	0	0.0	1	0	0.0
Turkey	7	0	0.0	3	0	0.0	4	0	0.0
United Arab Emirates	4	2	50.0	1	1	100.0	3	1	33.3
United Kingdom	138	41	29.7	84	30	35.7	54	11	20.4
United States	820	236	28.8	485	154	31.8	335	82	24.5
Others	18	2		12	0		6	3	
TOTAL	**2905**	**660**	**22.7**	**1737**	**447**	**25.7**	**1168**	**213**	**18.2**

http://acl2019pcblog.fileli.unipi.it/?p=161

How to Do Research?

http://matt.might.net/articles/phd-school-in-pictures/

What is Research?

Goal: Create New Scientific Knowledge

In My Research Lab this Typically Involves:

Designing new NLP/Machine Learning algorithms

- Understanding their performance by comparing to established baselines
- Improving their performance to achieve SOTA Results

Proposing new tasks / applications

- Developing datasets and models
- Show why this is useful (why should anyone care?)

How is Research Different from Taking Courses?

Taking Courses: Instructor tells you exactly what to do

Research: Define **open-ended** problem with an advisor

You (students) take the initiative

Discuss and refine technical approaches

You (students) implement the approach and conduct experiments to verify the idea

How to make good progress on research?

Clearly define the problem you want to solve

Understand the literature (what has been done previously?)

Implement initial prototype / preliminary analysis to demonstrate the **feasibility of your idea**.

After convincing yourself (and your advisor), conduct extensive experiments to **convince top researchers in your field** (much harder).

Why publish research papers?

Publication is the main way that new ideas are communicated and adopted.

In NLP/ML, the most competitive venues are conferences.

Your ideas are more likely to be adopted by a scientific community if they are published in a top-tier venue

Some famous papers only on ArXiv, but mostly from already established researchers.

Peer review is a way to get feedback and improve your work.

What makes a good research paper?

Is the problem new?

Is your approach new?

How good are the results (in comparison to prior work)?

Can you contribute any new datasets / code? Are they likely to be useful to other researchers?

Is the paper well-written?

Research is Hard

Your idea may not work as well as you hoped.

If a problem is easily solved, it's probably already been done.

Designing and performing (good) experiments is not easy.

Deadlines.

Research is Rewarding

You Advanced Science!

Feels good when your first top conference paper is accepted (acceptance rates: 10-30%)

Very nice when other people attend your talk, use your code, read and cite your paper.

Best way to establish yourself as the leading expert on a topic.

Reading #1 is out
Due Sep 5

August 28, 2019 (Wednesday)	AI Seminar by Mounica Maddela • 4:00 -- 5:00pm, Dreese 480	Multi-task Pairwise Neural Ranking for Hashtag Segmentation by Mounica Maddela, Wei Xu, Daniel Preoţiuc-Pietro (ACL 2019)
August 29, 2019	Twitter and Twitter API Tutorial • Brief history of Twitter • Key features of Twitter • Hands-on instructions on obtaining Twitter data via APIs	★ Twitter API Tutorial by Wei Xu ★ What is Twitter, a Social Network or a News Media? by Kwak, Lee, Park and Moon (WWW 2010) How to Do Research with a Professor by Jason Eisner How to Read a Technical Paper by Jason Eisner
September 5, 2019	Language Identification and Naïve Bayes [Reading 1 due] • Language Identification • Supervised Learning • Classification • Naïve Bayes Algorithm + feature selection (Information Gain)	★ Cross-domain Feature Selection for Language Identification by Lui, Baldwin (IJCNLP 2011) Evaluating language identification performance by Mitja T @tm langid.py: An Off-the-shelf Language Identification Tool by Lui, Baldwin (ACL 2012) 6 Easy Steps to Learn Naïve Bayes Algorithm by Sunil Ray Text Classification using Naive Bayes by Hiroshi Shimodaira

In-class Presentation

5539 Presentations (2020SP)

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

100% $ % .0 .00 123 Arial 10

	A	B	C
1	**Date**	**Tentative Plan**	**Student Presentation Group #1**
2	1/10/2020	1st class - no student presentation	------
3	1/17/2020	Twitter API	
4	1/24/2020	LangID & Naive Bayes	
5	1/31/2020	Tagging	
6	2/7/2020	Vector Semantics	Jack DeLano, Zachary Mack
7	2/14/2020	Summarization, Paraphrase	Manikya Swathi Vallahajosyula, Haixin Wang
8	2/21/2020	Information Extraction	Ziwei Jin, Kyle Kottyan
9	2/28/2020	Paraphrase	Celia Tang, Saumya Sahai
10	3/6/2020	Linear Regression	Sandeep Venkatesh And Milt Levy
11	3/13/2020	Spring Break	
12	3/20/2020	Logistic Regression (Reading 8)	Nitish Dashora , Aaron Rehfeldt
13	3/27/2020	RNN & Neural MT (Topic Modeling? - Reading 9)	Biplob Biswas, Yang Zhong
14	4/3/2020	Sentence Pair Modeling (Reading 10)	
15	4/10/2020	CNN & Sentiment Analysis (Reading 11)	Sam Lin
16	4/17/2020	Wrapup	------
17			