Social Media Data Mining in Python

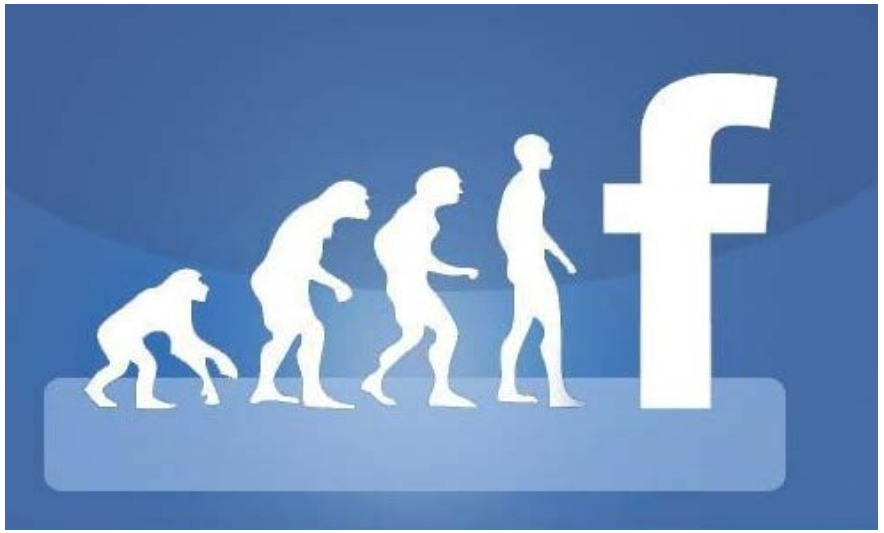
Dr. Kokil Jaidka *jaidka@sas.upenn.edu*

Why profile people from social media?

- **Social media measurement is efficient**
  - Unobtrusive and cheap
  - Can be used for communities

- **Social media gives insight**
  - Visualization is key!

- **Social media has vast potential**
  - Health, mental health risk
  - Advertising
  - **Winning the US elections**



Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski, David Stillwell and Thore Graepel

Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski, David Stillwell and Thore Graepel

Power through 'Us': Leaders' Use of We-Referencing Language Predicts Election Victory

Niklas K. Steffens, S. Alexander Haslam

Published: October 23, 2013 • https://doi.org/10.1371/journal.pone.0077952

NEWS & CULTURE

Why I Quit Using the Word Just in My Emails

BY CHELSEA STONE

OCTOBER 12, 2016 9:00 AM

PHOTO: STOCKSY

e

guage

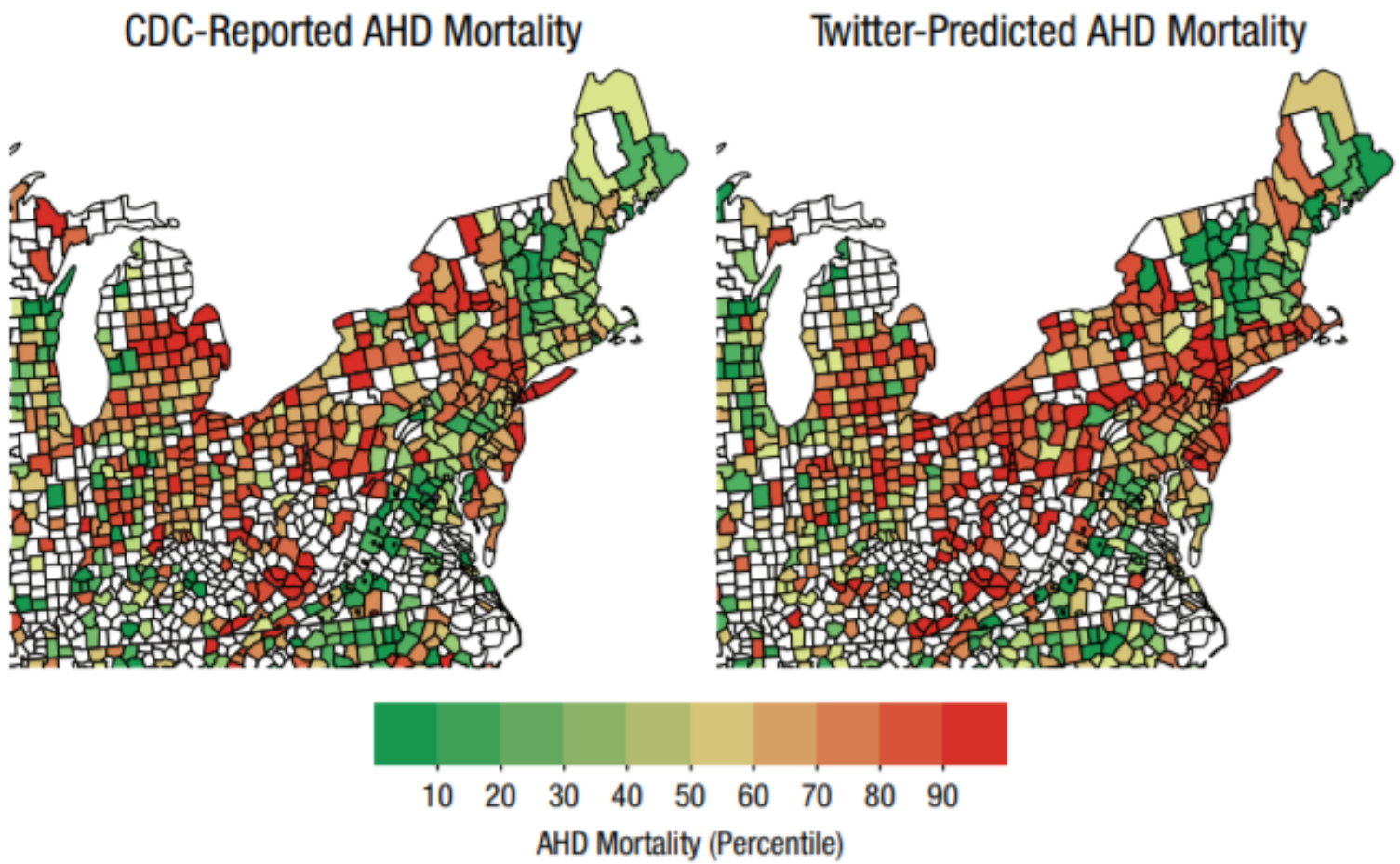
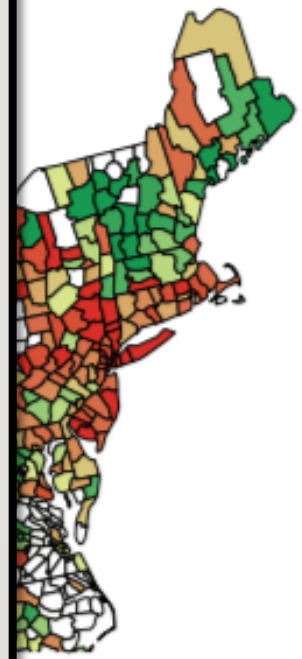


PHOTO: STOCKSY

uage

The Cambridge Analytica Files

‘I made Steve Bannon’s psychological warfare tool’: meet the data war whistleblower

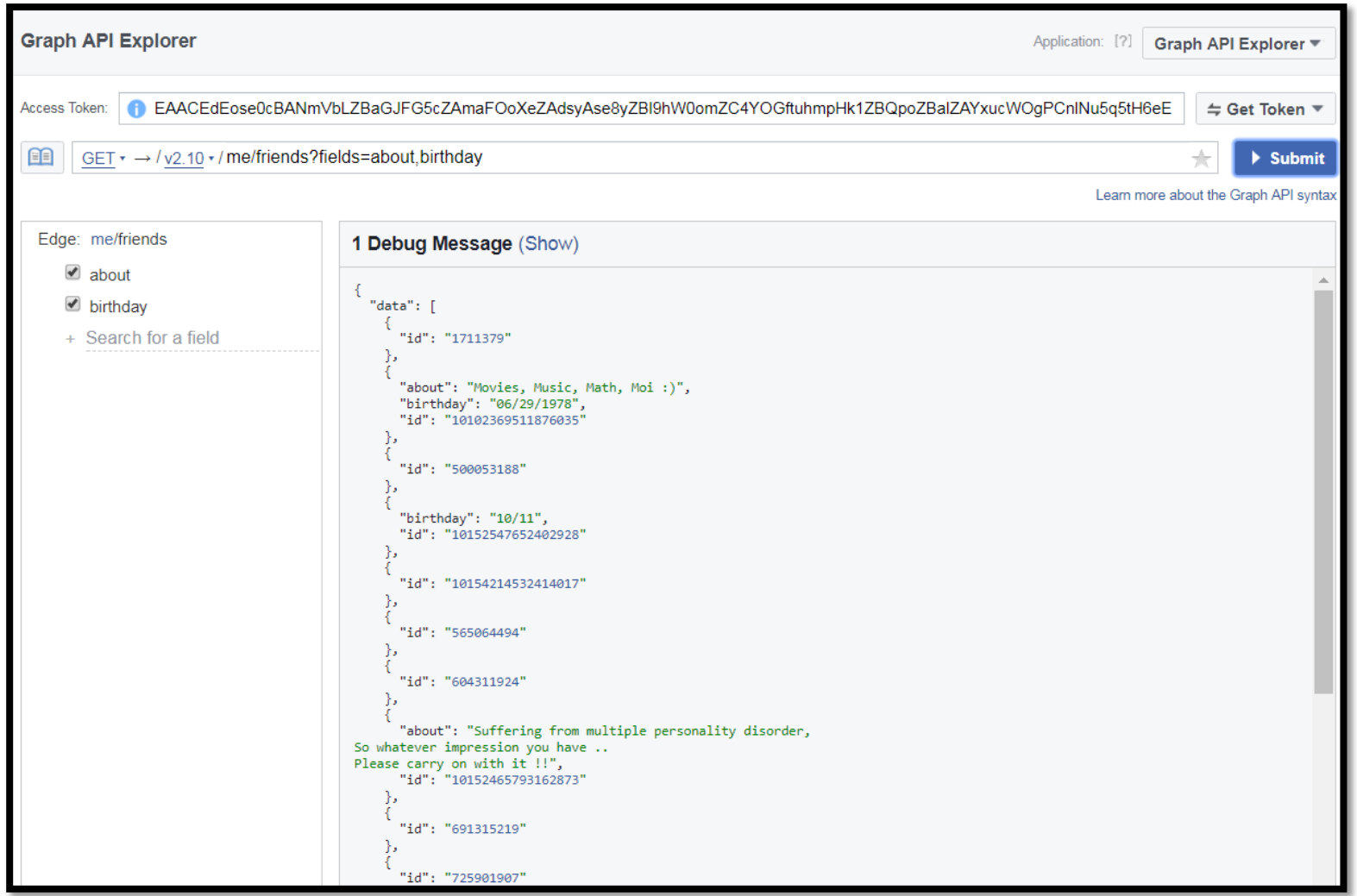


D Mortality

guage

Let’s explore the Graph API explorer

https://developers.facebook.com/tools/explorer/145634995501895/



Takeaways

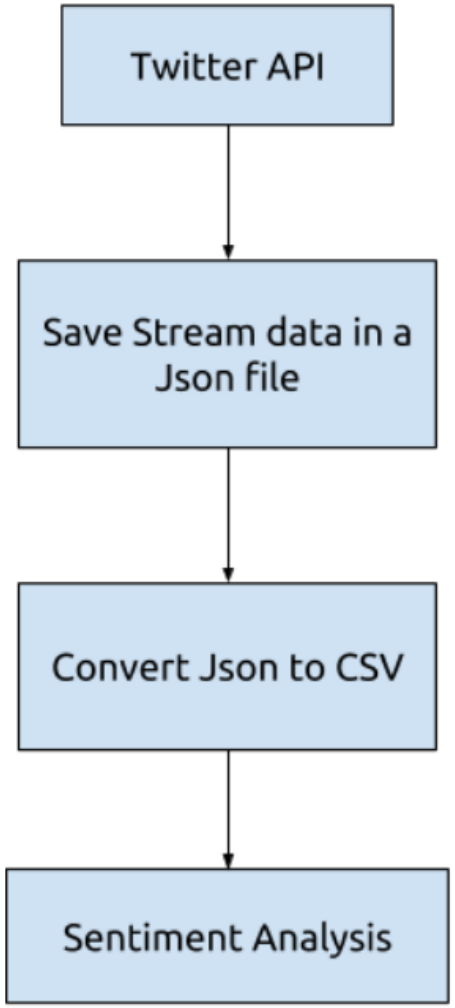
After this lecture, you will

- **Have a general understanding of developer APIs (application programming interface) to collect social media data.**
- **Be able to write a Python script to collect data from Twitter.**
- **Know how to analyse and plot the data you have collected.**

Prerequisites

- **Basic familiarity with Python:**
  - You've set up Anaconda Spyder
  - You know how to install packages: *tweepy*

Data Mining from Twitter



Get the credentials:

- Consumer Key (API Key)
- Consumer Secret (API Secret)
- Access Token
- Access Token Secret

Save the Twitter data in a json File.
To collect the streaming data,
we use Tweepy.

Extract only the fields you
want, into a csv file.

Do sentiment analysis on
the tweet's text using NLTK

https://www.nextplacelab.de

Twitter Authentication

To access and collect Twitter data, you need to have appropriate authentication for the purpose of an application and/or a script.

- **Consumer key**
- **Consumer secret**
- **Access token**
- **Access token secret**

You can generate authentication information by creating an app for Twitter in https://apps.twitter.com

Twitter API (Application Programming Interface)

https://developer.twitter.com/en/docs/basics/getting-started#get-started-app

Secure | https://developer.twitter.com/en/docs/basics/getting-started#get-started-app

Developer Use cases Products Docs More Apply

Ads

Metrics

Publisher tools & SDKs

Developer utilities

API reference index

You can display Twitter content in your native mobile app using the Twitter Kit SDK. You can generate Tweet and timeline views for use in an iOS or Android app.

Get started: Build an app on Twitter

Twitter's API platform includes numerous endpoints to help you build an app and solution on Twitter. Our basic endpoints are available for free. As your app or solution needs grow, you'll also find enterprise APIs that include increased levels of access.

Get started with the basic REST and Streaming APIs

Twitter's basic REST and Streaming APIs enable free access to numerous endpoints. To get started, you must first create an app.

1. Create an app

To use an endpoint, you must create an app and use our OAuth-based authorization system. Visit apps.twitter.com to create one.

2. Start using the endpoints!

Once you've setup your account, accessing the endpoint is super simple. Check out the documentation and API reference for additional details about each endpoint. There are many libraries and utilities in different programming languages that can help you to get started.

Have a question? There's a good chance our community has an answer for you. Visit our developer forums to review topics, ask questions, and learn from others.

Get started with Twitter Data objects

The data provided by Twitter APIs are made up of data objects and their attributes rendered in JavaScript Object Notation (JSON). To learn more about Tweet metadata, see this introduction to Tweet JSON objects.

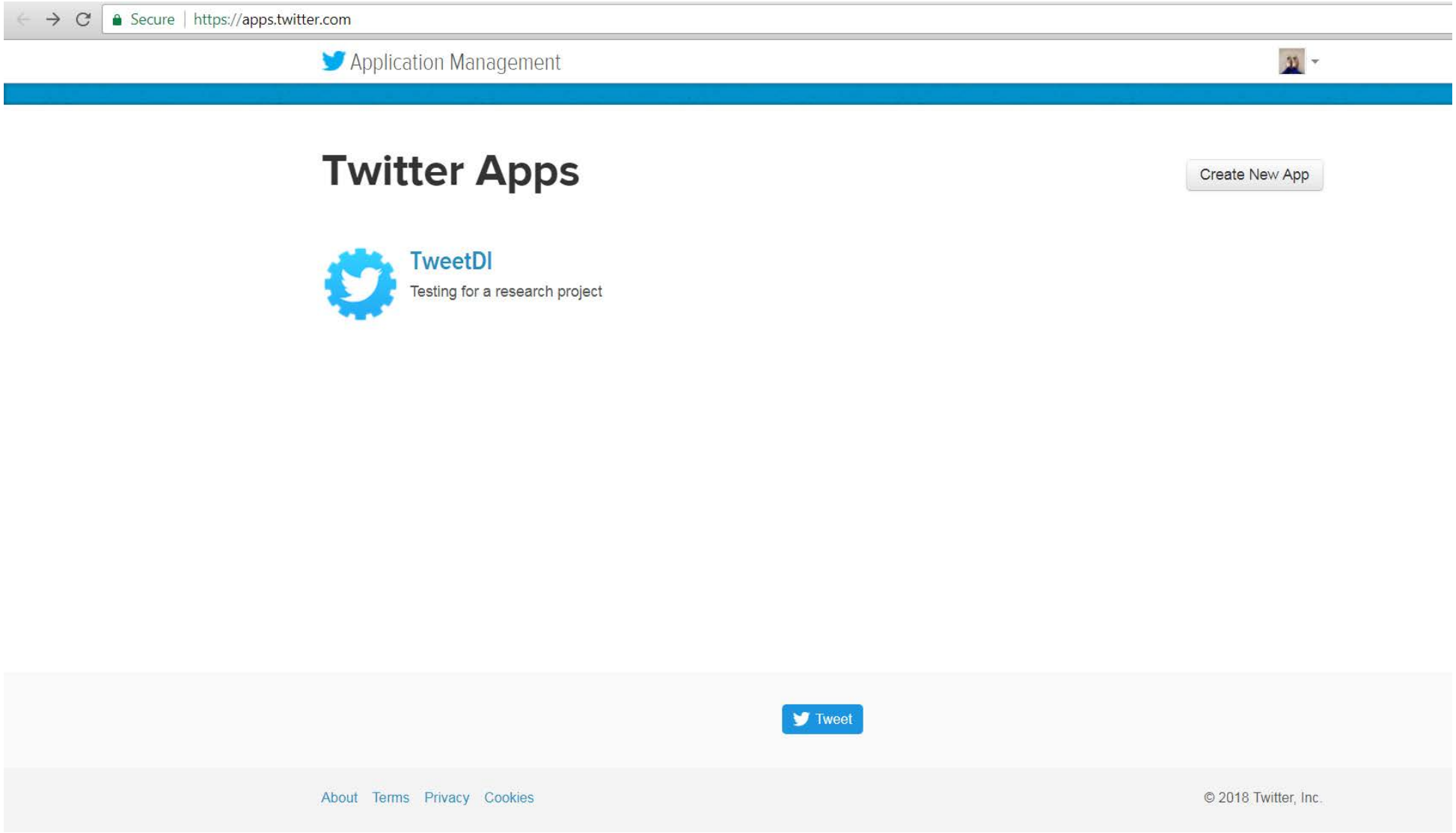
The following documentation provide 'data dictionaries' to help you understand the many attributes that make up Twitter Tweets, Users and other objects.

Tweet object

Tweets are the basic atomic building block of all things Twitter. Click here to learn more about the Tweet object and its data fields.

Twitter API (Application Programming Interface)

https://apps.twitter.com/



Twitter API (Application Programming Interface)

https://apps.twitter.com/

TweetDl

Test OAuth

Details | Settings | Keys and Access Tokens | Permissions

Testing for a research project

http://www.placeholder.com

Organization

Information about the organization or company associated with your application. This information is optional.

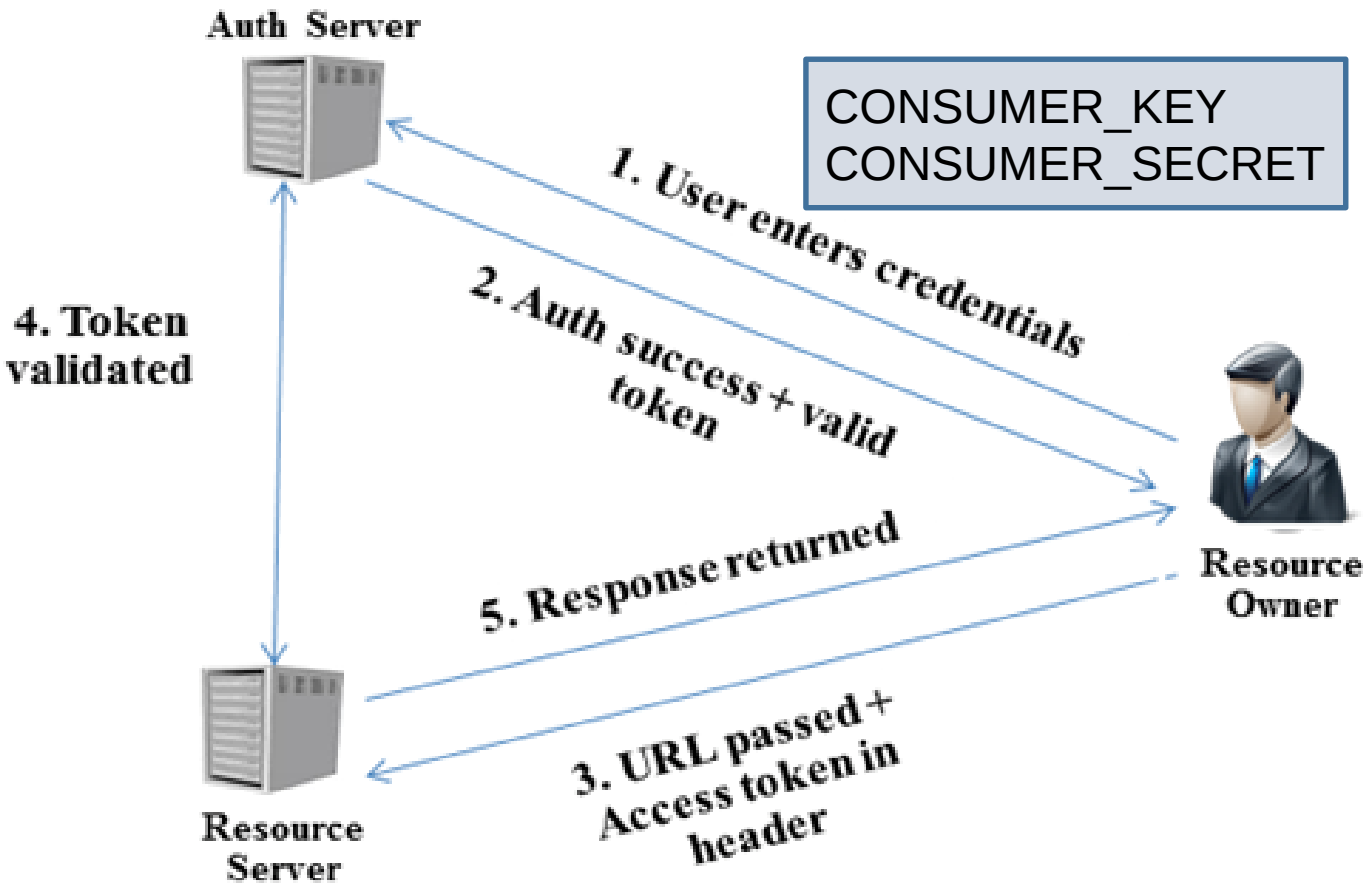
Organization	None
Organization website	None

Application Settings

Your application's Consumer Key and Secret are used to authenticate requests to the Twitter Platform.

Access level	Read-only (modify app permissions)
Consumer Key (API Key)	[redacted] (manage keys and access tokens)
Callback URL	None
Callback URL Locked	No
Sign in with Twitter	No
App-only authentication	https://api.twitter.com/oauth2/token
Request token URL	https://api.twitter.com/oauth/request_token
Authorize URL	https://api.twitter.com/oauth/authorize
Access token URL	https://api.twitter.com/oauth/access_token

Oauth dance



https://hsc.com

Example code

We are now ready to make an authenticated call to Twitter using tweepy.

```
CONSUMER_KEY
CONSUMER_SECRET
ACCESS_TOKEN
ACCESS_TOKEN_SECRET
```

```
import tweepy

auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

api = tweepy.API(auth)

#now, use the api object to search, get data from your account, post new status messages etc.
```

What kind of data can you pull?
https://apigee.com/console

Penn
UNIVERSITY of PENNSYLVANIA

Providers API Resources Console Switch to...

API: Twitter
Service: https://api.twitter.com/1.1
Authentication: twitter-feedkoko
Feedba

API Console service will be turned down on April 14, 2018. learn more

Select an API method

Search methods...

- GET /statuses/home_timeline.json

Tweets

- GET /statuses/retweets/{id}.json
- GET /statuses/show/{id}.json
- POST /statuses/destroy/{id}.json
- POST /statuses/update.json
- POST /statuses/retweet/{id}.json
- POST /statuses/update_with_media.json
- GET /statuses/oembed.json

Search

- GET /search/tweets.json

Help

- GET /help/configuration.json
- GET /help/languages.json
- GET /help/privacy.json

ang=en&count=1 Send

```
count=1 HTTP/1.1

_signature_method=&quot;HMAC-
nce=&quot;3502770084&quot;,oauth_ver
ature=&quot;uMaKnm7fxTOqbOGV5sinrdS4
```

Response Snapsh

```
HTTP/1.1 200 OK
x-frame-options: SAMEORIGIN
x-rate-limit-remaining: 178
last-modified: Sat, 31 Mar 2018 19:44:36 GMT
status: 200 OK
Content-Length: 5629
x-response-time: 211
Connection: keep-alive
x-transaction: 0019cdee0047fee2
Server: tsa_b
pragma: no-cache
cache-control: no-cache, no-store, must-revalidate, pre-check=0, post-check=0
x-connection-hash: 6cdc178394e60653c5aba086b443c9b8
x-xss-protection: 1; mode=block; report=https://twitter.com/i/xss_report
x-content-type-options: nosniff
x-rate-limit-limit: 180
expires: Tue, 31 Mar 1981 05:00:00 GMT
Date: Sat, 31 Mar 2018 19:44:36 GMT
set-cookie: personalization_id=&quot;v1_u8cSenkmhaY055/ezhKyHQ==&quot;; Expires=Mon, 30
Mar 2020 19:44:36 UTC; Path=/; Domain=.twitter.com
set-cookie: guest_id=v1%3A152252547656227498; Expires=Mon, 30 Mar 2020 19:44:36 UTC;
Path=/; Domain=.twitter.com
x-rate-limit-reset: 1522526333
```

Summary: Data Mining from Twitter

- **Has an API**
- **Needs one to create an app: https://apps.twitter.com**
- **OAuth needs keys – and an access token**
- **There's a Python wrapper for the API: *tweepy***
  http://docs.tweepy.org/en/v3.5.0/getting_started.html#models
- **There is a way to try it out: https://apigee.com/console**
- **The response is in JSON format.**

Takeaways

After this lecture, you will

- ~~**Have a general understanding of developer APIs (application programming interface) to collect social media data.**~~
- **Be able to write a Python script to collect data from Twitter.**
- **Know how to analyse and plot the data you have collected.**

What would this code do?

```
api.update_status('tweepy + oauth!')
```

What would this code do?

```
api.update_status('tweepy + oauth!')

public_tweets = api.home_timeline()
for tweet in public_tweets:
    print(tweet.text)
```

What would this code do?

```
api.update_status('tweepy + oauth!')

public_tweets = api.home_timeline()
for tweet in public_tweets:
    print(tweet.text)


for tweet in tweepy.Cursor(api.search, q= query).items(50):
    print(tweet.text)
```

Example code

```
data = []
for tweet in tweepy.Cursor(api.search, q = query, count =50).items():
        #extract the fields you want
        #append them as a new row in data[]
```

Example code

```
csvFile = open('result.csv','w')
#Use csv Writer
csvWriter = csv.writer(csvFile)
csvWriter.writerow(["created", "text", "retwc", "hashtag", "followers", "friends"])

data = []
for tweet in tweepy.Cursor(api.search, q = query, count =50).items():
        #extract the fields you want
        #append them as a new row in data[]
        csvWriter.writerow([created, str(text).encode("utf-8"), retwc, hashtag, followers, friends])
csvFile.close()
```

Example code

```
csvFile = open('result.csv','w')
#Use csv Writer
csvWriter = csv.writer(csvFile)
csvWriter.writerow(["created", "text", "retwc", "hashtag", "followers", "friends"])

data = []
for tweet in tweepy.Cursor(api.search, q = query, count =50).items():
        #extract created_at
        #extract text
        #extract retweetcount
        #extract hashtag
        #extract followers
        #extract friends
        csvWriter.writerow([created, str(text).encode("utf-8"), retwc, hashtag, followers, friends])
csvFile.close()
```

Example code

```
csvFile = open('result.csv','w')
#Use csv Writer
csvWriter = csv.writer(csvFile)
csvWriter.writerow(["created", "text", "retw", "hashtag", "followers", "friends"])
data = []
for tweet in tweepy.Cursor(api.search, q = query, count =50).items():
        created = tweet.created_at                          #extract created_at
        text    = tweet.text                                #extract text
        retw    = tweet.retweet_count                       #extract retweet count
        try:                                                #extract hashtag
                hashtag = tweet.entities[u'hashtags'][0][u'text']
       except:
                hashtag = "None"
        followers = tweet.author.followers_count            #extract followers
        friends   = tweet.author.friends_count              #extract friends
        csvWriter.writerow([created, str(text).encode("utf-8"), retwc, hashtag, followers, friends])
csvFile.close()
```

Tweet data dictionary

https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json

Secure | https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json

Developer Use cases Products Docs More Apply

Search all documentation...

Basics

Accounts and users

Tweets

- Post, retrieve and engage with Tweets
- Get Tweet timelines
- Curate a collection of Tweets
- Optimize Tweets with Cards
- Search Tweets
- Filter realtime Tweets
- Sample realtime Tweets
- Get batch historical Tweets
- Rules and filtering
- Premium enrichments
- Tweet data dictionaries

Tweet data dictionaries

Overview Guides

Overview contents

Introduction to Tweet JSON

Jump to on this page

Data Mining from YouTube

- **Has an API: https://developers.google.com/youtube/v3/docs/**
- **Needs one to create an app: https://console.developers.google.com/**
- **OAuth needs keys**
- **There's a Python wrapper for the API: *google-api-python-client, unidecode***
- **There is a way to try it out: https://developers.google.com/apis-explorer/**
- **The response is in JSON format.**

Data Mining from Facebook

- **Has an API: https://developers.facebook.com/docs/graph-api**
- **Needs one to create an app: https://developers.facebook.com/apps**
- **OAuth needs keys**
- **There's a Python wrapper for the API: *facebook-sdk***
- **There is a way to try it out: https://developers.facebook.com/tools/explorer/**
- **The response is in JSON format**

Takeaways

After this lecture, you will

- ~~**Have a general understanding of developer APIs (application programming interface) to collect social media data.**~~
- ~~**Be able to write a Python script to collect data from Twitter**~~**.**
- **Know how to analyse and plot the data you have collected.**

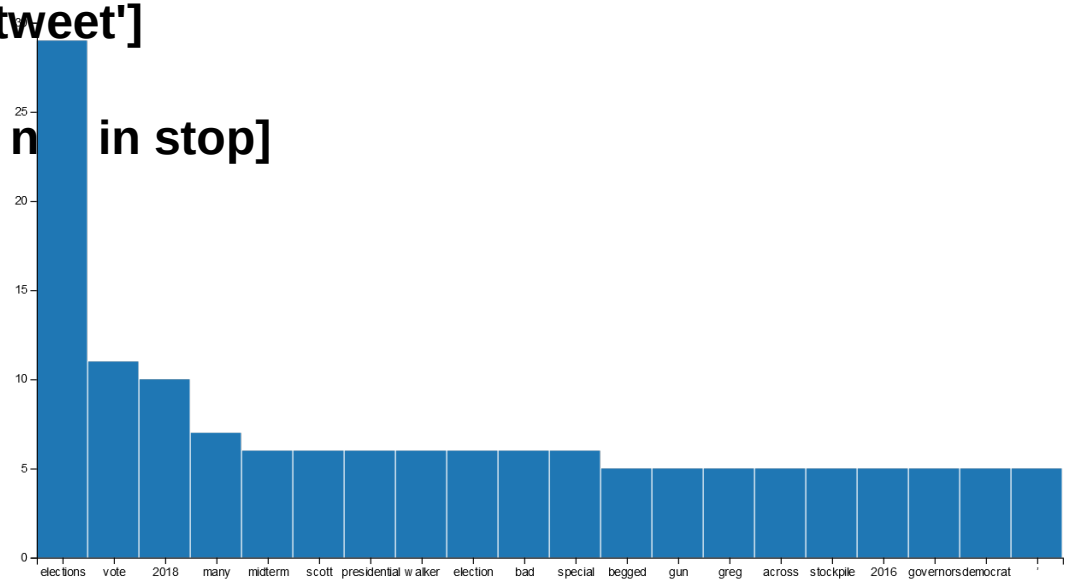
Sentiment analysis: Example code (needs nltk)

```
from nltk.sentiment.vader import SentimentIntensityAnalyzer

#text = text from a single tweet
sid = SentimentIntensityAnalyzer()
ss = sid.polarity_scores(text)

#<class 'dict'>
#{'compound': 0.5267, 'pos': 0.285, 'neu': 0.593, 'neg': 0.122}
```

https://github.com/amueller/word_cloud

Bar plot: Example code (needs nltk, Collections.Counter)

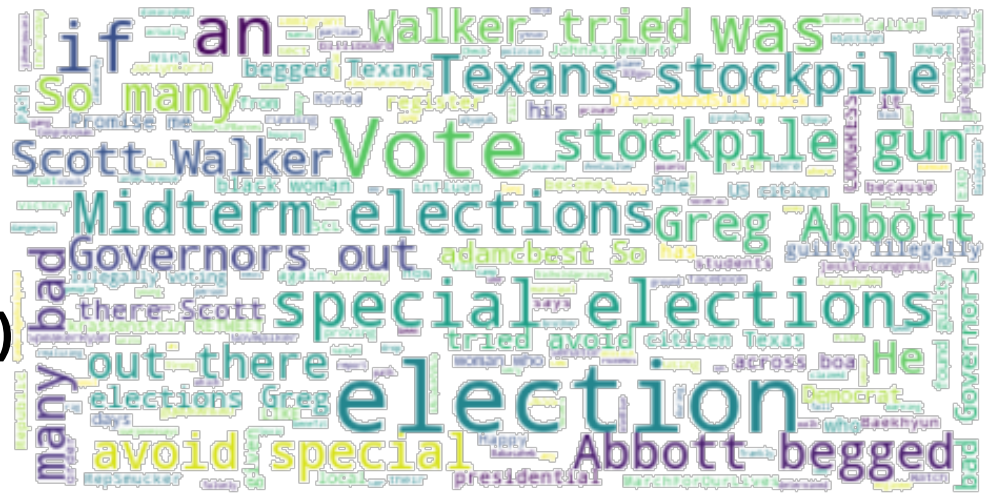
```
punctuation = list(string.punctuation)
count_all = Counter()
stop = stopwords.words('english') + punctuation + ['rt','retweet']
corpus=str.lower(corpus)
terms_stop = [term for term in preprocess(corpus) if term n in stop]
terms_only = [term for term in preprocess(corpus)
            if term not in stop and
            not term.startswith(('#', '@'))]
count_all.update(terms_only)
word_freq = count_all.most_common(20)
labels, freq = zip(*word_freq)
data = {'data': freq, 'x': labels}
bar = vincent.Bar(data, iter_idx='x')
bar.to_json('term_freq.json', html_out=True, html_path='chart.html')
```



https://github.com/amueller/word_cloud

Word cloud: Example code (needs nltk, wordcloud and matplotlib)

```
import matplotlib.pyplot as plt
import random
from wordcloud import WordCloud

#text = all the text from your tweets
wordcloud = WordCloud(#parameters).generate(text)

plt.imshow(wordcloud)
plt.axis("off")
plt.show()
```



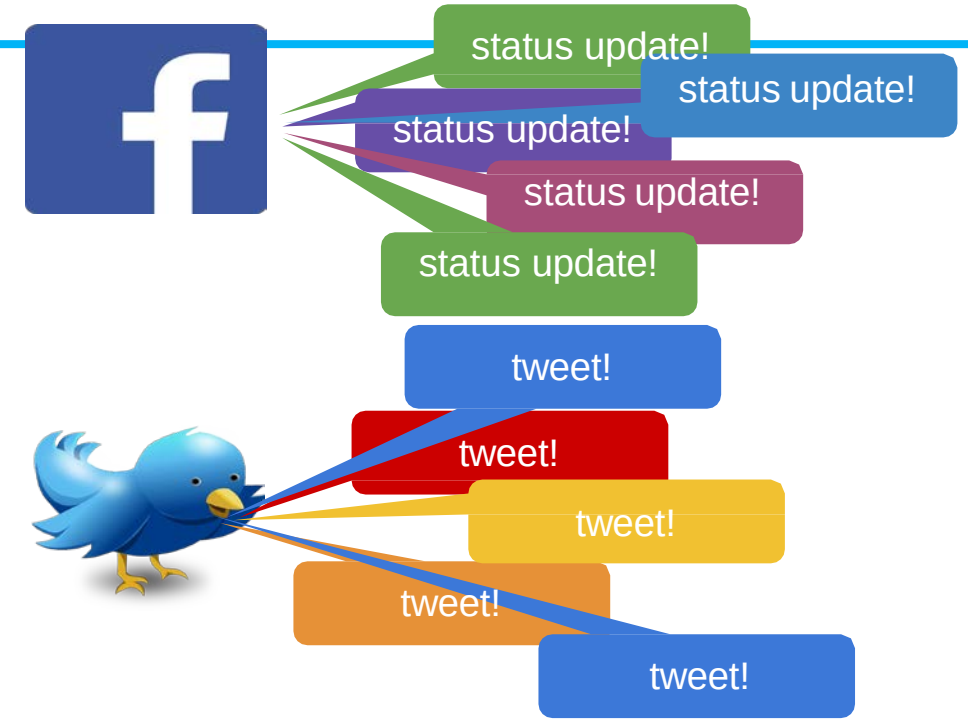
https://github.com/amueller/word_cloud

Takeaways

After this lecture, you will

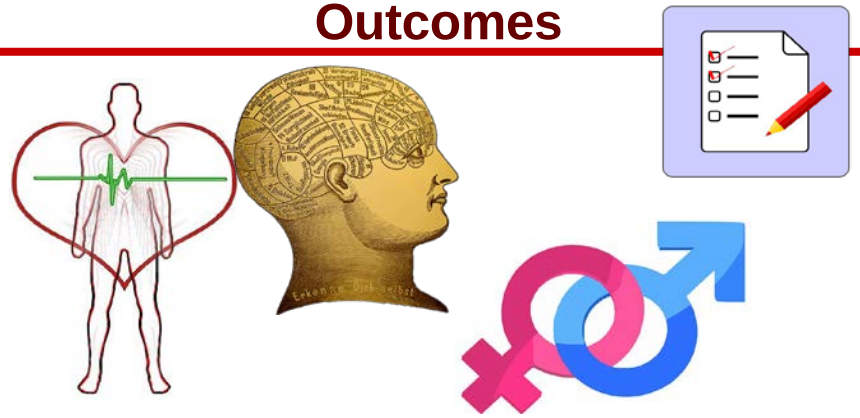
- ~~**Have a general understanding of developer APIs (application programming interface) to collect social media data.**~~
- ~~**Be able to write a Python script to collect data from Twitter**~~**.**
- ~~**Know how to analyze and plot the data you have collected**~~**.**

Some other applications



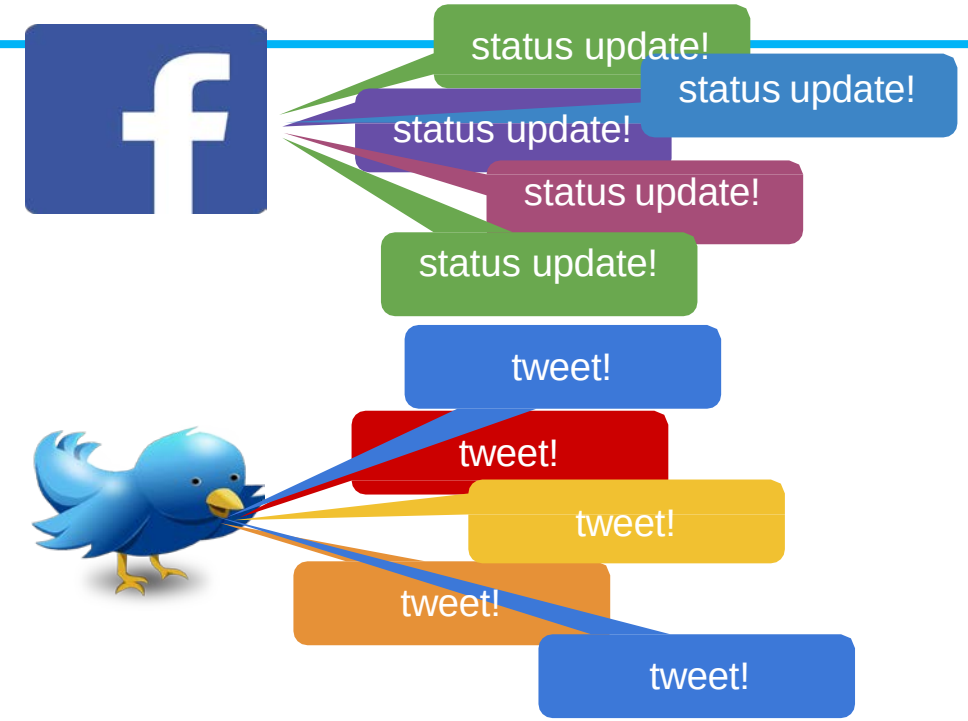
language analysis

Outcomes

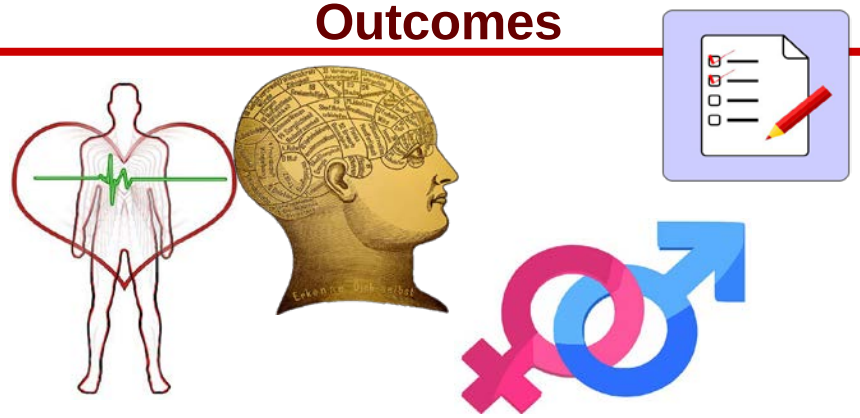


Slide courtesy: Dr. Andrew H. Schwartz

Some other applications



Outcomes



language analysis

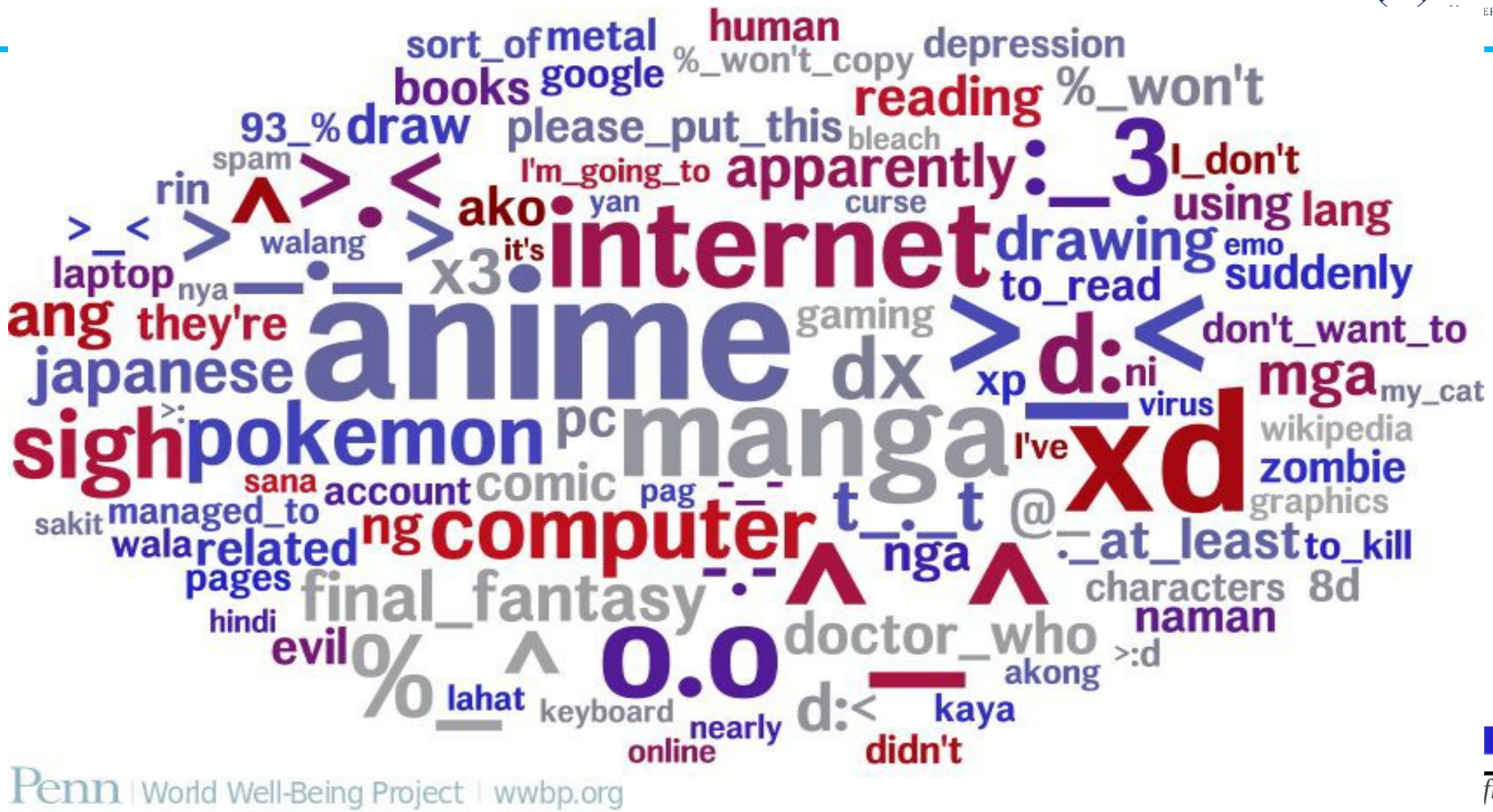
- **prediction (measurement)**
- **insights**

Slide courtesy: Dr. Andrew H. Schwartz

Extraversion -- *sociable, assertive, active, energetic, talkative, outgoing*

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). **Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach**. *In PLOS*

Introversion



frequency

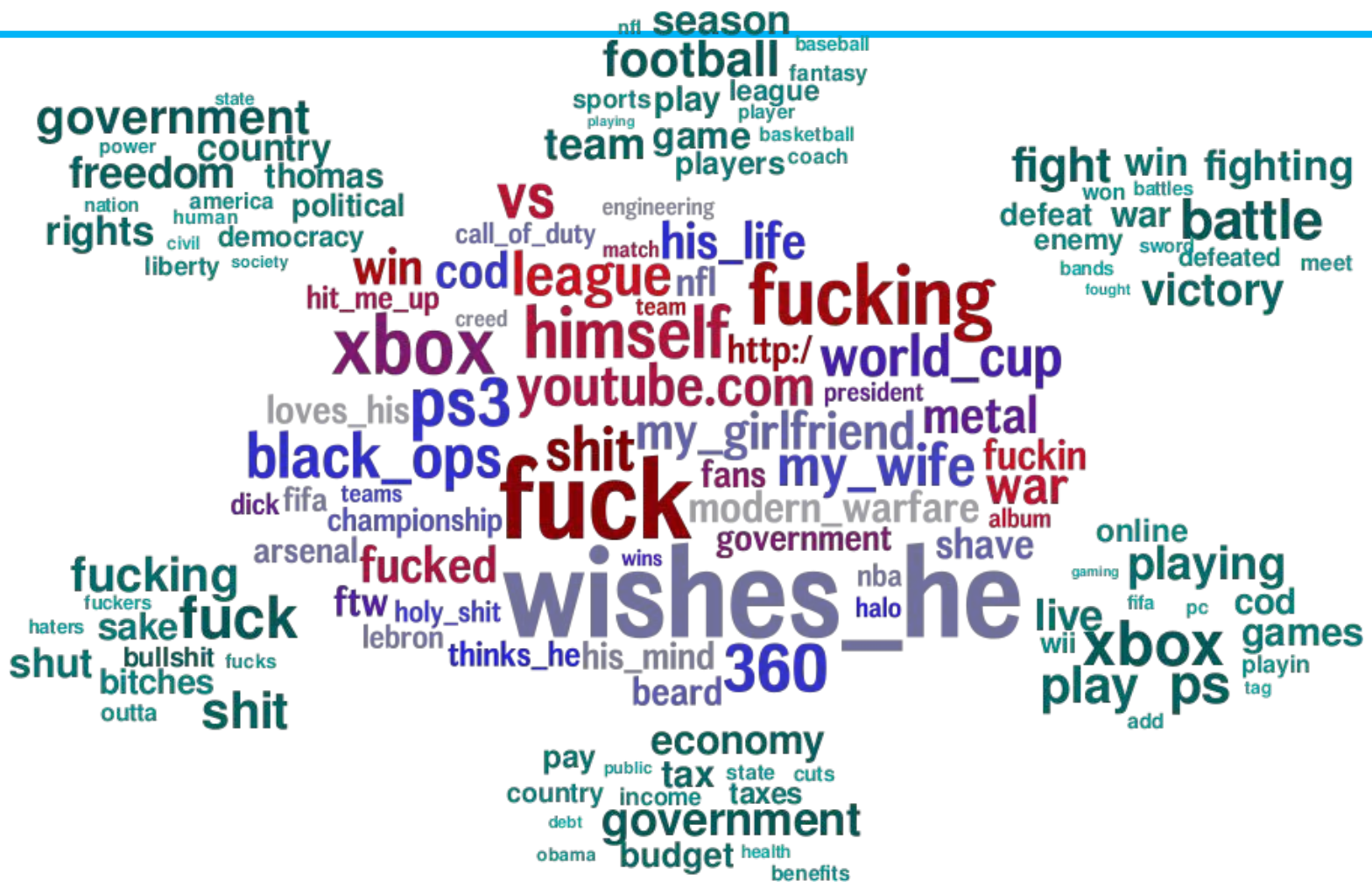
Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., & Ungar, L. H. (2013). **Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach**. *In PLOS*

Gender


super soooo
flippin uber tonight
soooo overly excited
sooooo
tomorrow ridiculously
sooo satisfying excitement yay
hun
girlies love ;D
^.^ <3 inlove
sweetie mucho
sweetheart
^_^
_
i'm_so
her_new
best_friend chocolate her she
shopping
with_my ugh <3 sister
proud_of
mom birthday
my_hair can't_wait
loves_her wonderful
am_so
so_happy lovely
excited baby
soo
hubby boyfriend
much_fun
dress love:)you cute
soooyay_!:(
i_miss mommy girls
supportive loving
herself so_much
yummy like_it_on
wonderful
cleaning wishes_she
amazing
family husband helped
love_him my_family
soooo
omg husband loved
soooo
friends thankful
truely boyfriend
blessed daughter
grateful lucky
bestie :) hacked ar
bestfriend par
ily boyfriend =]
bf
babe besties xoxo
xo <3
cute baby
sweet
sooo adorable
hes cutest babies
soooo puppy he's
lil aww awww
soo
turns nephew
celebrate
happy years
st birthday
wonderful wishing
brother daughter special
sister niece son
a a a
correlation strength
relative frequency
b b b
prevalence in topic

Explicit Language Warning...

Gender


season
nfl
football
baseball
fantasy
sports
play
league
player
playing
team
game
basketball
players
coach
government
state
power
country
freedom
thomas
nation
america
human
political
rights
civil
democracy
liberty
society
vs
engineering
call_of_duty
match
his_life
win
cod
league
nfl
hit_me_up
team
fucking
creed
himself
xbox
http:/
world_cup
president
youtube.com
loves_his
ps3
my_girlfriend
metal
black_ops
shit
fans
my_wife
fuckin
war
dick
fifa
teams
fuck
modern_warfare
championship
album
government
shave
arsenal
fucked
wins
nba
fucking
fuckers
ftw
holy_shit
wishes_he
halo
haters
sake
fuck
lebron
shut
bullshit
fucks
thinks_he
his_mind
360
bitches
beard
outta
shit
fight
win
fighting
won
battles
defeat
war
battle
enemy
sword
defeated
meet
bands
fought
victory
online
gaming
playing
fifa
pc
cod
live
wii
xbox
games
playin
play
ps
tag
add
economy
pay
public
tax
state
cuts
country
income
taxes
debt
government
obama
budget
health
benefits
a
a
a
correlation strength
relative frequency
b
b
b
prevalence in topic

Resources

- Twitter object dictionary: https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json
- Twitter API explorer: https://apigee.com/console/twitter
- Facebook Graph API explorer: https://developers.facebook.com/tools/explorer/
- YouTube API explorer: https://developers.google.com/apis-explorer/
- JSON viewer: http://jsonviewer.stack.hu

Example JSON (paste this into JSON viewer)

{"statuses":[{"created_at":"Sat Mar 31 19:44:21 +0000 2018","id":980168945314484225,"id_str":"980168945314484225","text":"RT @BeaxyExchange: It's important to understand why you want to be involved with crypto. It isn't all about getting rich quick -- It is ab\u2026","truncated":false,"entities":{"hashtags":[],"symbols":[],"user_mentions":[{"screen_name":"BeaxyExchange","name":"Beaxy","id":905959920222314498,"id_str":"905959920222314498","indices":[3,17]}],"urls":[]},"metadata":{"iso_language_code":"en","result_type":"recent"},"source":"\u003ca href=\"http:\/\/twitter.com\" rel=\"nofollow\"\u003eTwitter Web Client\u003c\/a\u003e","in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":904320034247503873,"id_str":"904320034247503873","name":"Dollarsbad","screen_name":"dollarsbad","location":"","description":"#f4f #followback #fallowyou #fallowme #crypto #bch #bts #eth #etc #ltc #ico #bitcoin #blockchain #dollarfail #trader \n#\u0432\u0437\u0430\u0438\u043c\u043d\u044b\u0439\u0444\u0430\u043b\u043b\u043e\u0432\u0438\u043d\u0433 #\u043a\u0440\u0438\u043f\u0442\u0430 #bounty #mining","url":null,"entities":{"description":{"urls":[]}},"protected":false,"followers_count":7249,"friends_count":6426,"listed_count":10,"created_at":"Sun Sep 03 12:27:52 +0000 2017","favourites_count":279,"utc_offset":null,"time_zone":null,"geo_enabled":false,"verified":false,"statuses_count":270,"lang":"ru","contributors_enabled":false,"is_translator":false,"is_translation_enabled":false,"profile_background_color":"F5F8FA","profile_background_image_url":null,"profile_background_image_url_https":null,"profile_background_tile":false,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/904322292146204672\/LAmUupJE_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/904322292146204672\/LAmUupJE_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/904320034247503873\/1504441996","profile_link_color":"1DA1F2","profile_sidebar_border_color":"C0DEED","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"has_extended_profile":true,"default_profile":true,"default_profile_image":false,"following":false,"follow_request_sent":false,"notifications":false,"translator_type":"none"},"geo":null,"coordinates":null,"place":null,"contributors":null,"retweeted_status":{"created_at":"Sun Jan 07 16:55:23 +0000 2018","id":950048228585754626,"id_str":"950048228585754626","text":"It's important to understand why you want to be involved with crypto. It isn't all about getting rich quick -- It\u2026 https:\/\/t.co\/w4L9wim7Ye","truncated":true,"entities":{"hashtags":[],"symbols":[],"user_mentions":[],"urls":[{"url":"https:\/\/t.co\/w4L9wim7Ye","expanded_url":"https:\/\/twitter.com\/i\/web\/status\/950048228585754626","display_url":"twitter.com\/i\/web\/status\/9\u2026","indices":[116,139]}]},"metadata":{"iso_language_code":"en","result_type":"recent"},"source":"\u003ca href=\"http:\/\/twitter.com\" rel=\"nofollow\"\u003eTwitter Web Client\u003c\/a\u003e","in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":905959920222314498,"id_str":"905959920222314498","name":"Beaxy","screen_name":"BeaxyExchange","location":"Worldwide","description":"All-In-One Cryptocurrency Exchange","url":"https:\/\/t.co\/KvsZZ4nPNh","entities":{"url":{"urls":[{"url":"https:\/\/t.co\/KvsZZ4nPNh","expanded_url":"https:\/\/beaxy.com","display_url":"beaxy.com","indices":[0,23]}]},"description":{"urls":[]}},"protected":false,"followers_count":16842,"friends_count":61,"listed_count":38,"created_at":"Fri Sep 08 01:04:11 +0000 2017","favourites_count":2,"utc_offset":-18000,"time_zone":"Central Time (US & Canada)","geo_enabled":false,"verified":false,"statuses_count":71,"lang":"en","contributors_enabled":false,"is_translator":false,"is_translation_enabled":false,"profile_background_color":"000000","profile_background_image_url":"http:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_tile":false,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/922641316797808640\/tvVaNoli_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/922641316797808640\/tvVaNoli_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/905959920222314498\/1520222882","profile_link_color":"00ACAC","profile_sidebar_border_color":"000000","profile_sidebar_fill_color":"000000","profile_text_color":"000000","profile_use_background_image":false,"has_extended_profile":false,"default_profile":false,"default_profile_image":false,"following":false,"follow_request_sent":false,"notifications":false,"translator_type":"none"},"geo":null,"coordinates":null,"place":null,"contributors":null,"is_quote_status":false,"retweet_count":175,"favorite_count":183,"favorited":false,"retweeted":false,"possibly_sensitive":false,"lang":"en"},"is_quote_status":false,"retweet_count":175,"favorite_count":0,"favorited":false,"retweeted":false,"lang":"en"}],"search_metadata":{"completed_in":0.045,"max_id":980168945314484225,"max_id_str":"980168945314484225","next_results":"?max_id=980168945314484224&q=bitcoin&lang=en&count=1&include_entities=1","query":"bitcoin","refresh_url":"?since_id=980168945314484225&q=bitcoin&lang=en&include_entities=1","count":1,"since_id":0,"since_id_str":"0"}}

Thank you!

Slides will be up at
kokiljaidka.wordpress.com