

# Day 11: Data from the Web

ME314: Introduction to Data Science and Big Data Analytics

LSE Methods Summer Programme

15 August 2018

# Day 11 Outline

Social Media Data

Challenges of Social Data

Accessing social media APIs

"Web scraping"

## Social Media Data

# Why social media data?

- ▶ Volume and coverage
- ▶ Twitter: 328 million monthly active users, 530m tweets per day <sup>1</sup>
- ▶ Facebook: 1.32 billion daily active users on average for June 2017, 2 billion monthly active users as of 2017 <sup>2</sup>
- ▶ Real time — new data is available (somewhat) publicly immediately on current events
- ▶ Metadata — geographic location, user device, profile, timestamp and other metadata is accessible.

# Appeal of Social Media data

- ▶ Good case for machine learning and data mining — lots of data, lots of metadata
- ▶ Many-to-many *broadcast* text corpus
- ▶ Social network analysis: a graph of social connections

# Network data structure of social media

- ▶ Broadcast
  - ▶ simplex (e.g. radio, television, smoke signal)
  - ▶ duplex (e.g. round-table meeting, walkie-talkies)
- ▶ Point-to-point: sender specifies receivers
- ▶ Social media allow many of these different forms of communication
- ▶ Twitter in particular is a completely new model of communication
- ▶ Every user is a sensor, receiver, and broadcaster — a distributed sensor network (Crooks et al 2012)

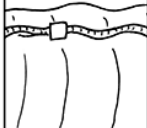
WHEN AN EARTHQUAKE HITS,  
PEOPLE FLOOD THE INTERNET  
WITH POSTS ABOUT IT—SOME  
WITHIN 20 OR 30 SECONDS.

ROBM163 HUGE  
EARTHQUAKE HERE!



DAMAGING SEISMIC  
WAVES TRAVEL AT  
3-5  $\text{km/s}$ . FIBER  
SIGNALS MOVE AT  
 $\sim 200,000 \text{ km/s}$ .

(MINUS NETWORK LAG)



THIS MEANS WHEN THE SEISMIC  
WAVES ARE ABOUT 100 KM OUT,  
THEY BEGIN TO BE OVERTAKEN BY  
THE WAVES OF POSTS ABOUT THEM.



PEOPLE OUTSIDE THIS RADIUS  
MAY GET WORD OF THE QUAKE  
VIA TWITTER, IRC, OR SMS  
BEFORE THE SHAKING HITS.

WHOA!  
EARTHQUAKE!



SADLY, A TWITTERER'S  
FIRST INSTINCT IS NOT  
TO FIND SHELTER.

RT @ROBM163 HUGE  
EARTHQUAKE HERE!



## **Big data, big challenges**



# Big data and bias

The tools we will focus on today will give you the opportunity to dramatically expand the scope of your data collection efforts.

However, size isn't everything.

- ▶ Population bias
  - ▶ Sociodemographics are (strongly) correlated with social media use
- ▶ Self-selection within samples
  - ▶ Partisans are much more likely to post about politics than independents
- ▶ Proprietary algorithms and unknown bias
  - ▶ e.g. Twitter API returns data which is “is not an accurate representation of the overall platform's data”
- ▶ Social media activity does not generalise easily
  - ▶ Twitter is to social scientists what the fruit fly is to biologists – a model organism, but one that generalises poorly

See Ruths and Pfeffer, 2015, “Social media for large studies of behavior”.

# Ethical concerns: informed consent (I)

## Experimental evidence of massive-scale emotional contagion through social networks



Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock

### Significance

We show, via a massive ( $N = 689,003$ ) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.

---

# Ethical concerns: informed consent (I)

## Experimental evidence of massive-scale emotional contagion through social networks



Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock

### Significance

We show, via a massive ( $N = 689,003$ ) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.

---

*“The study was consistent with Facebook’s Data Use Policy, to which all users agree prior to creating an account on Facebook, constituting informed consent for this research.” Kramer et al, PNAS, 2014*

# Ethical concerns: informed consent (I)

## Experimental evidence of massive-scale emotional contagion through social networks



Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock

### Significance

We show, via a massive ( $N = 689,003$ ) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.

---

*“The collection of the data by Facebook may have involved practices that were not fully consistent with the principles of obtaining informed consent and allowing participants to opt out.” Editor-in-Chief, PNAS, 2015*

# Ethical concerns: informed consent (I)

## Experimental evidence of massive-scale emotional contagion through social networks



Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock

### Significance

We show, via a massive ( $N = 689,003$ ) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.

---

Are EULAs (End-User License Agreement) too complex to allow 'informed consent'?

## Ethical concerns: informed consent (II)

### **Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach**

**Whistleblower describes how firm linked to former Trump adviser Steve Bannon compiled user data to target American voters**

## Ethical concerns: informed consent (II)

### Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach

**Whistleblower describes how firm linked to former Trump adviser Steve Bannon compiled user data to target American voters**

- ▶ CA gained access to Facebook data through a partnership with Aleksandre Kogan, a UK academic
- ▶ Kogan presented his data gathering as academic, but agreed to share information with CA
- ▶ Facebook claims that users gave consent to share data with Kogan, but not to the secondary sharing with CA
- ▶ Kogan's app collected profile data from participants' networks using the social graph API

## Ethical concerns: informed consent (II)

### **Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach**

**Whistleblower describes how firm linked to former Trump adviser Steve Bannon compiled user data to target American voters**

Two privacy issues:

1. Lack of consent for data exposure to a third party
2. Lack of consent for harvesting of social graph data



# Big data and neutrality

*“There is nothing about doing data analysis that is neutral. What and how data is collected, how the data is cleaned and stored, what models are constructed, and what questions are asked – all of this is political.” Danah Boyd, NYU*

# Practical concerns

- ▶ Legal issues need to catch up with the technology
- ▶ Large amounts of data
  - ▶ storage problems
  - ▶ analysis problems
- ▶ Language is informal and often non-textual (emoticons, links, images) - and slang, txtspk, emoticons :-)
- ▶ Lots of fake users
- ▶ Commercial interfaces are brittle and opaque
- ▶ A lot of the content is moronic. . .

Example

# Example: Twittdiots



**Michael Matthews**  
@YourBuddyBurns



Follow

I'm tired of this terrorist bullshit fucking w our country. Fuck it, just nuke Czechoslovakia

↩ Reply ↻ Retweet ★ Favorite ... More



**InstrumentalStash**  
@HashHitz



Follow

I Can't believe that pair in the Boston bombing was NOT Towel heads!!! They are Czechoslovakian! Daamn!! FUCK Czechoslovakia!

↩ Reply ↻ Retweet ★ Favorite ... More



**Kaitlynn Schuler**  
@KaitlynnSchuler



Follow

Some Czech mother fucker is about to get LITTTT up. #gethim

↩ Reply ↻ Retweet ★ Favorite ... More



**s\_elliott11**

What did America ever do to the Czech Republic? Where even is the Czech Republic? Have fun with the devil terrorboy

🐦 2 days ago ↩ Reply ↻ Retweet ☆ Favorite



**Jafar El-Shabazz**  
@llcooljaff



Follow

The media fucked up! They was sayin the suspect was a dark skinned male..turned out to be a Czech republican. ???!

↩ Reply ↻ Retweet ★ Favorite ... More

# Example applications

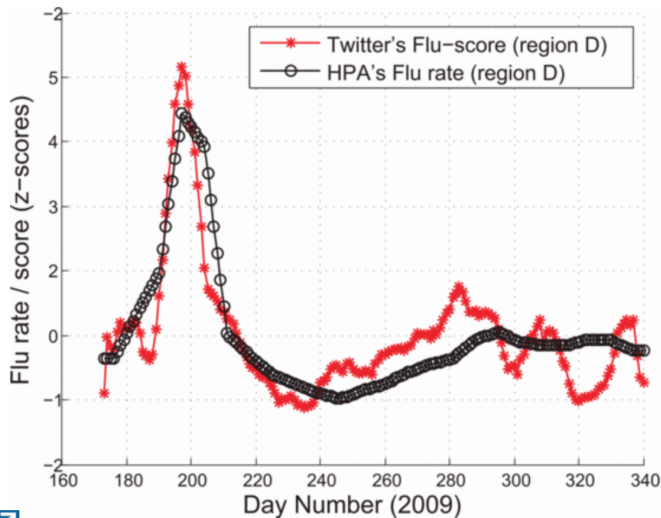
Nevertheless, social media data can provide interesting (and important) insights into human behaviour.

- ▶ Offers measures of actual behaviors, as compared with self-reports of behaviors
  - ▶ Particularly important for studying phenomena where people either deliberately or accidentally misreport their behaviour
  - ▶ i.e. reporting of social ties, anti-social behaviour, etc
- ▶ Offers instantaneous information about potentially important trends
  - ▶ i.e. employment (Toole et al, 2015); public opinion (Beauchamp, 2016); health (Ginsberg et al, 2009)
- ▶ Offers the opportunity to study the mechanics of social systems
  - ▶ How do individuals interact? How do they form social ties? How does segregation occur?

# Example application (I)

- ▶ Tracking disease through google search terms and social media (Lampos et al 2010)
  - ▶ Locate tweets in urban centres
  - ▶ Uses a Porter stemmer and stopwords
  - ▶ Uses regression to learn which words are associated with flu outbreaks: 97 'markers' (features)
  - ▶ Use this association to observe current outbreaks

## Example application (I)



## Example application (II)

- ▶ Can social media content predict market outcomes? (Brown et al, 2017, Economic Inquiry)
  - ▶ Combine 13.8 million tweets about football matches with prices from an online betting exchange
  - ▶ Predict premier league match outcomes as a function of a) betting prices and b) average twitter sentiment towards each team
  - ▶ “We find that positive tone does predict match outcomes in a way not fully captured by betting prices.”

# Other examples

- ▶ Predicting election outcomes or polls
- ▶ Sentiment: particularly for financial or corporate interests
- ▶ Government security/intelligence
- ▶ Social network analysis: a graph of social connections
- ▶ Nulty et al (2015) study of EP 2014



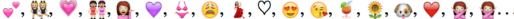

# 'Fixing' the biased twitter sample

## Barbera, 2017, Working Paper

- ▶ Using twitter in studies of social behaviour is difficult because we lack information about the sociodemographic characteristics of twitter users
- ▶ Researchers cannot estimate 'survey' weights to recover representativeness of their samples as we do with traditional surveys
- ▶ (Additional problem: many interesting questions require demographic information!)
- ▶ Solution: match a large (250,000) sample of twitter users to voter registration records, which provide information on age, gender, race, party identification, and - indirectly - house value
- ▶ Train a classifier to learn the text features most associated with these demographics
- ▶ Predict demographics for many other users

# 'Fixing' the biased twitter sample

Table 4: Top predictive features (emoji, words, accounts) most associated with each category.

Female	 love, women, hair, girl, husband, mom, omg, cute, excited, <3, girls, yay, happy, hubby, boyfriend, :, can't, baby, wine, thank, heart, nails... @TheEllenShow, @khloekardashian, @MileyCyrus, @Starbucks, @jtimberlake, @VictoriasSecret, @WomensHealthMag, @channingtatum...
Male	 bro, man, wife, good, causewereguys, gay, great, dude, f*ck, nice, game, iphone, ni**a, church, time, #gay, girlfriend, bruh, sportscenter... @SportsCenter, @danieltoosh, @MensHealthMag, @AdamScheffer, @ConanOBrien, @KingJames, @katyperry, @ActuallyNPH...

# 'Fixing' the biased twitter sample

Age: 18-25

👉, 🧑, 😊, 😊, 😊, 😊, 😊, 😊, 😊, 🖱️, 🧑, 😊, 😊, 🎓, 😊, 😊, 📖, 🗨️...

class, college, semester, life, (:, sportscenter, campus, best, literally, like, haha, just, :d, finals, classes, okay, professor, exam, studying...

@SportsCenter, @wizkhalifa, @MileyCyrus, @danieltosh, @instagram, @EmWatson, @KevinHart4real, @UberFacts, @vine...

Age: 26-40

🧑, 🧑, 🧑, 📷, 💪, 😊, ✈️, 😊, 🍷, 😊, 🚲, 😊, 🍷, 🍷, 🍷...

excited, work, amazing, bar, awesome, wedding, #tbt, pretty, #nofilter, ppl, bday, time, lil, #love, yay, #latergram, office, game, tonight, boo, super...

@danieltosh, @ConanOBrien, @jtimberlake, @StephenAtHome, @chelseahandler, @KimKardashian, @instagram, @NPR, @britneyspears...

Age:  $\geq 40$

🍰, 😊, 🏁, 😊, 🐾, 📱, 🏀, 💖, 🌹, 🌟, 🙏, 🌟, 🎸, 🎸, 🏀...

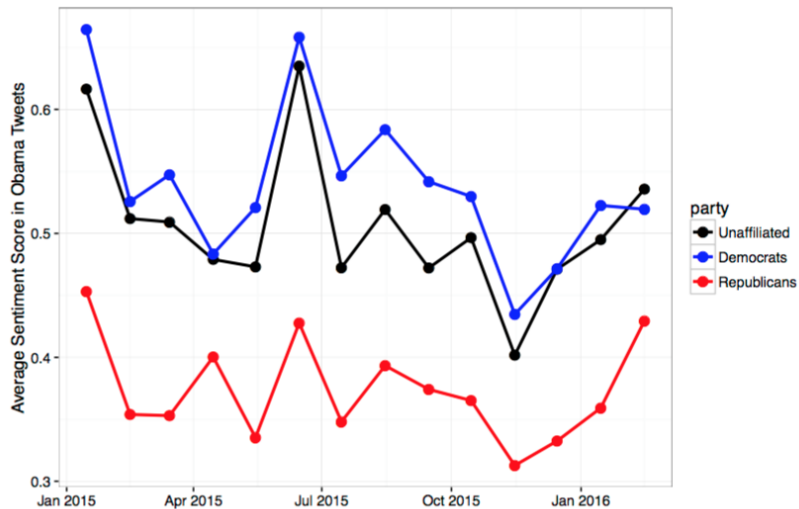
great, daughter, son, nice, r, good, ok, kids, congratulations, obama, hi, nbcthevoice, wow, happy, hope, beautiful, sorry, rock, grandson, amen...

@jimmyfallon, @cnnbrk, @YouTube, @Pink, @TheEllenShow, @NBCTheVoice, @SteveMartinToGo, @Oprah, @sethmeyers, @FoxNews...

# 'Fixing' the biased twitter sample

Democrat	 philly, barackobama, la, sf, pittsburgh, women, nytimes, philadelphia, smh, president, gop, black, hillaryclinton, gay, republicans ... @BarackObama, @rihanna, @maddow, @billclinton, @khloekardashian, @billmaher, @Oprah, @KevinHart4real, @algore, @MichelleObama ...
Republican	 foxnews, #tcot, church, christmas, oklahoma, florida, obama, great, realdonaldtrump, golf, beach, megynkelly, tula, byu, seanhannity ... @FoxNews, @danieltosh, @TimTebow, @MittRomney, @taylorswift13, @jimmyfallon, @RyanSeacrest, @Starbucks, @JimGaffigan ...
Unaffiliated	 ohio, arkansas, columbus, cleveland, cincinnati, utah, toledo, cavs, #wps, browns, ar, akron, hogs, bengals, kent, dayton, #cbj, reds ... @instagram, @SportsCenter, @KingJames, @vine, @AnnaKendrick47, @wizkhalifa, @WhatTheFFacts, @galifianakis, @ActuallyNPH ...

# Validation



## Social Media Data access

# How can we access this data?

- ▶ API: Application Programming Interface — a way for two pieces of software to talk to each other
- ▶ Twitter, facebook, google — all expose public web services
- ▶ Your software can receive (and also send) data automatically through these services
- ▶ Data is sent by http — the same way your browser does it
- ▶ Most services have helping code (known as a wrapper) to construct http requests
- ▶ both the wrapper and the service itself are called APIs
- ▶ http service also sometimes known as REST (REpresentational State Transfer)

# HyperText Transfer Protocol

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

## Why are we interested in HTTP?

facebook

YAHOO!

twitter

myspace.com  
a place for friends

**Because nearly everything a typical user does on the Internet uses HTTP**

CNN.com

@mail.ru



Google  
Earth

Gmail  
by Google



# Anatomy of a http request

```
https://api.twitter.com/1.1/search/tweets.json?  
q=Nick+Clegg%21&since_id=24012619984051000&max_id=25012619984051
```

Nick Clegg! becomes Nick+Clegg%21

- ▶ Parameters to the API are encoded in the URL
- ▶ you must encode requests — spaces and non ASCII characters are replaced

# cURL and wget

- ▶ It's not usually necessary to construct these kind of requests yourself
- ▶ R, Python, and other programming languages have libraries to make it easier
- ▶ Usually you will need cURL installed to access an API, wget for downloading a website
- ▶ The documentation for the API will describe the parameters that are available.

# Available social media APIs

- ▶ Wikipedia: mediawiki
- ▶ Google
  - ▶ google plus
  - ▶ blogger
- ▶ reddit
- ▶ foursquare
- ▶ facebook
- ▶ twitter: REST, Streaming, firehose, commercial

Note: both Twitter and Facebook have increased the registration hurdles required for accessing their APIs recently.

# The twitter APIs: Search

- ▶ This is the most comprehensive API
- ▶ Returns a sample of historical data from the last 7 days.
- ▶ Stateless: you send a command and receive a result.
- ▶ http GET requests return information
- ▶ http POST requests upload or alter information (e.g. twitterbots)
- ▶ The manual: <https://developer.twitter.com/en/docs/tweets/search/overview>
- ▶ R package : twitterR

# The twitter APIs: Streaming

- ▶ Connect to the twitter server and collect tweets as they fly by.
- ▶ The manual: <https://developer.twitter.com/en/docs/tweets/filter-realtime/overview>
- ▶ R package: streamR

# Authentication

- ▶ Username and Password
- ▶ OAuth (ROauth): share a key without sharing a username and password
- ▶ IP address limitations
- ▶ Rate limitations
- ▶ Per-user and per-application

# The Output: JSON and XML

- ▶ XML: eXtensible Markup Language: encodes documents in a form that is both human-readable and machine readable
- ▶ JSON : JavaScript Object Notation
- ▶ If you have a choice, you probably want JSON
- ▶ JSON uses key:value pairs, XML uses trees
- ▶ JSON is easily read into a programming language
- ▶ RJSONIO and xml2 are the relevant R packages

## And finally... the data.

- ▶ Full of spam, bots, unicode, and gibberish
- ▶ Lots of retweets (approximately one-third retweets, replies, tweets)
- ▶ Only 1% show location — some methods exist to infer location
- ▶ All aspects of metadata and reply/retweet structure are available
- ▶ All aspects of network structure: followers and 'friends', profile information



# Twitterbots

- ▶ API also allows actions such as posting tweets (POST)
- ▶ Examples:
  - ▶ @netflix\_bot posts new content using netflix api
  - ▶ @eqbot posts earthquake warnings
  - ▶ @pentametrone posts pairs of tweets in rhyming couplets <sup>3</sup>

# Twitterbots



**Big Ben**

@big\_ben\_clock



Follow

**BONG BONG BONG BONG BONG BONG BONG BONG  
BONG BONG**

10:00 AM - 10 Oct 2014



73



63

# Twitterbots in research

- ▶ Munger, 2017, Political Behaviour
- ▶ Research question: Does social sanctioning reduce racist online harassment?
- ▶ Design:
  - ▶ Randomly assign a sample of racist Twitter users to a treatment and control group
  - ▶ Treatment group: direct 'bots' representing in-group and out-group members to sanction users for their use of racist terms
  - ▶ Control group: leave users alone
  - ▶ Measure whether treatment group reduce their use of racist language in subsequent weeks

## Twitterbots in research



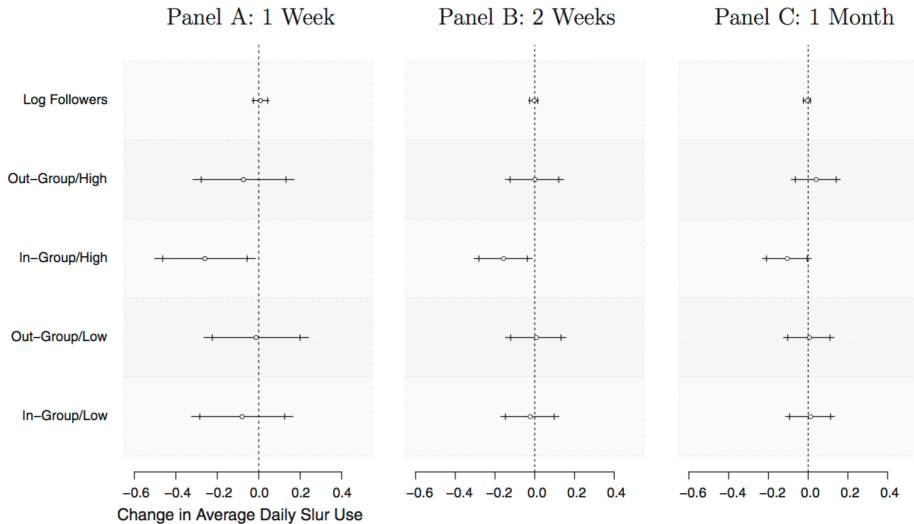
**Rasheed** [REDACTED]

@Rasheed [REDACTED]

@ [REDACTED] Hey man, just remember that there are real people who are hurt when you harass them with that kind of language

# Twitterbots in research

## Results



# Twitter uses: Exploiting the meta-data (non-textual)

- ▶ location
- ▶ time
- ▶ username
- ▶ user descriptions
- ▶ networks of followers
- ▶ retweets of followers and texts

# Connecting through R

R packages

- ▶ Twitter: twitteR for REST, streamR for Streaming
- ▶ Facebook: Rfacebook

Python: tweepy and facebook-sdk

other open-source tools exist

Integration with quanteda is fairly straightforward

## Other social media access packages

- ▶ `tumblrR` R interface to the Tumblr web API
- ▶ `instaR` R interface to Instagram API
- ▶ `Rlinkedin` R interface to LinkedIn API
- ▶ `RedditExtractorR` R interface for Reddit API



## Demonstration

## **Web scraping**

**(How to get visible content directly from web pages)**

# Scraping text from the web

- ▶ web crawlers/spider download sites by traversing links
- ▶ Python - scraPy, BeautifulSoup
- ▶ R - Rvest
- ▶ Chrome web plugins, import.io
- ▶ cUrl, wget, or other tools available ('httrack')
- ▶ Problems: rate limiting, ethical issues

## Demonstration

# Make scraping unnecessary!

- ▶ Organizations and governments should be aware of need for open, machine-readable data
- ▶ data.gov.uk, data.gov
- ▶ Data should be available in human and machine format!
- ▶ Make the raw data available in as many formats as possible.
- ▶ Consider machine readability at time of data collection
- ▶ Provide an Application Programming Interface (API)

# Summing up

- ▶ Social media data provides important new opportunities for gaining insights into human behaviour
- ▶ Misuse of social media data, particularly, is a real and serious problem
- ▶ The ethical implications of this type of data are still unclear
- ▶ Practically, accessing (some) big data sources is simple through APIs
- ▶ Webscraping can be a useful (though annoying) tool for collecting other data available online



# Commercial interfaces are brittle and opaque

This was an example in last year's slides:

```
## Code for API examples from Social Media class , 2017  
library(Rfacebook)
```

```
## Scraping most recent 200 posts from Trump FB page  
trump <- getPage("DonaldTrump", token = token, n = 100)
```

```
> Error in callAPI(url = url, token = token, api = api)
```

Why doesn't it work?

## Facebook shuts off access to user data for hundreds of thousands of apps

2

*All app makers who did not submit to the company's review process by its August 1st deadline are being cut off*

By [Nick Statt](#) | [@nickstatt](#) | Jul 31, 2018, 6:35pm EDT

Back