MY472 – Week 7: APIs

Pablo Barberá & Akitaka Matsuo

MY 472: Data for Data Scientists

November 13, 2018

Course website: lse-my472.github.io

Course outline

1. Introduction to data
2. The shape of data
3. Cloud computing
4. Basics of HTML and CSS
5. Using data from the internet
6. (Reading week)
7. Working with APIs
8. Creating and managing databases
9. Interacting with online databases
10. Exploratory data analysis
11. Parallel computing

Seminar schedule

7. APIs
   - 3rd marked assignment (in groups)
   - Deadline: November 23
8. SQL
9. Online databases
   - 4th marked assignment (in groups)
   - Deadline: December 7th
10. Exploratory data analysis
11. Course wrap-up
    - 5th marked assignment (individual)
    - Deadline: December 21st

Take-home exam due January 18

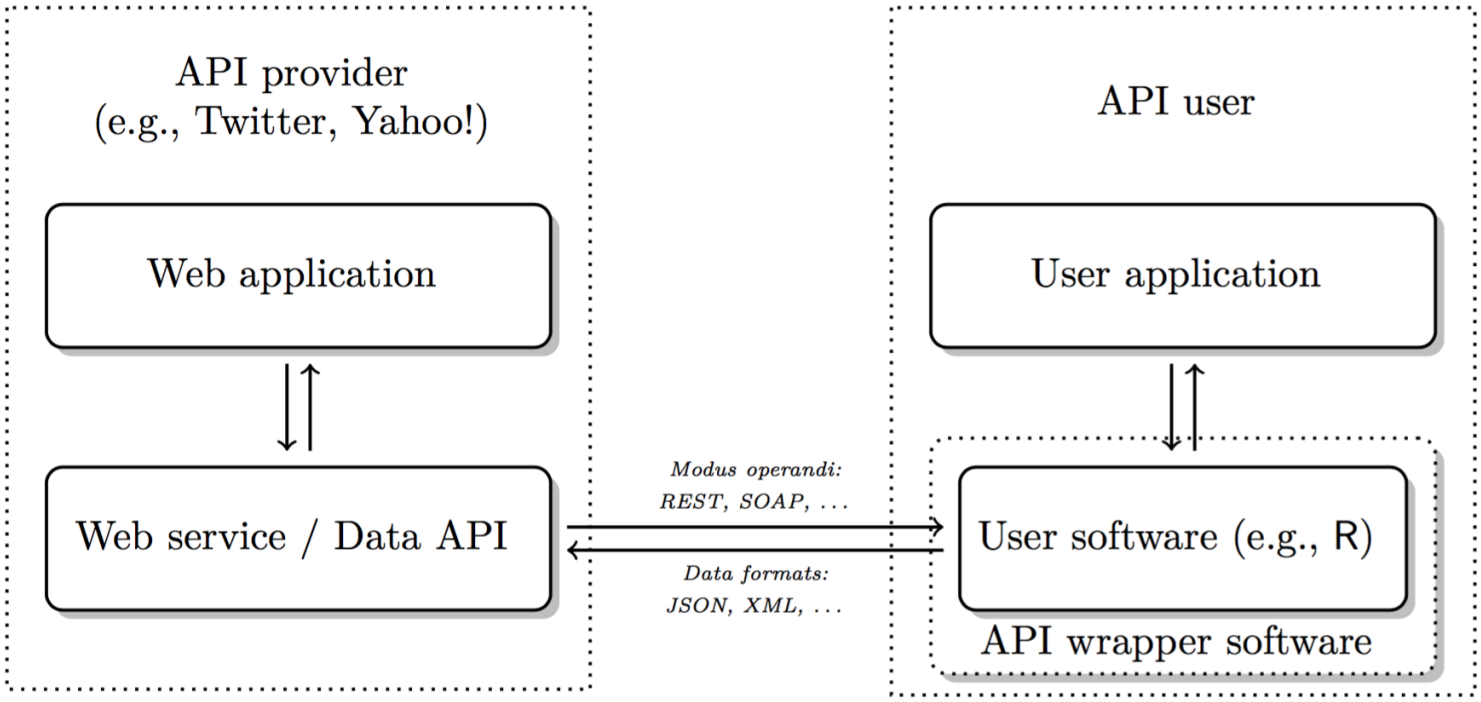
Plan for today

- APIs
  - Definition
  - Types of APIs
  - Constructing an API call
  - Authentication
  - Example: the New York Times API
- Social media data
  - Advantages & disadvantages of social media data
  - What data is available?
  - Twitter APIs

APIs

API = Application Programming Interface; a set of structured http requests that return data in a lightweight format.

HTTP = Hypertext Transfer Protocol; how browsers and e-mail clients communicate with servers.



Source: Munzert et al, 2014, Figure 9.8

APIs

Types of APIs:

1. RESTful APIs: queries for static information at current moment (e.g. user profiles, posts, etc.)
2. Streaming APIs: changes in users' data in real time (e.g. new tweets, weather alerts...)

APIs generally have extensive documentation:

- Written for developers, so must be understandable for humans
- What to look for: endpoints and parameters.

Most APIs are rate-limited:

- Restrictions on number of API calls by user/IP address and period of time.
- Commercial APIs may impose a monthly fee

Connecting with an API

Constructing a REST API call:

- Baseline URL endpoint:
  `https://maps.googleapis.com/maps/api/geocode/json`
- Parameters: `?address=london`
- Authentication token (optional): `&key=XXXXX`

From R, use `httr` package to make `GET` request:

```
library(httr)
r <- GET(
"https://maps.googleapis.com/maps/api/geocode/json",
query=list(address="london", key="XXXXX"))
```

If request was successful, returned code will be 200, where `4xx` indicates client errors and `5xx` indicates server errors.
If you need to attach data, use `POST` request.

```
{
   "results" : [
      {
         "address_components" : [
            {
               "long_name" : "London",
               "short_name" : "London",
               "types" : [ "locality", "political" ]
            },
            {
               "long_name" : "London",
               "short_name" : "London",
               "types" : [ "postal_town" ]
            }
         ],
         "formatted_address" : "London, UK",
         "geometry" : {
            "bounds" : {
               "northeast" : {
                  "lat" : 51.6723432,
                  "lng" : 0.148271
               },
               "southwest" : {
                  "lat" : 51.38494009999999,
                  "lng" : -0.3514683
               }
            },
            "location" : {
               "lat" : 51.5073509,
               "lng" : -0.1277583
            },
...
}
```

```
{
...
            "location_type" : "APPROXIMATE",
            "viewport" : {
               "northeast" : {
                  "lat" : 51.6723432,
                  "lng" : 0.148271
               },
               "southwest" : {
                  "lat" : 51.38494009999999,
                  "lng" : -0.3514683
               }
            }
         },
         "place_id" : "ChIJdd4hrwug2EcRmSrV3Vo6llI",
         "types" : [ "locality", "political" ]
      }
   ],
   "status" : "OK"
}
```

JSON

Response is often in JSON format (Javascript Object Notation).

- Type: `content(r, "text")`
- Data stored in key-value pairs. Why? Lightweight, more flexible than traditional table format.
- Curly brackets embrace objets; square brackets enclose arrays (vectors)
- Use `fromJSON` function from `jsonlite` package to read JSON data into R
- But many packages have their own specific functions to read data in JSON format; `content(r, "parsed")`

Authentication

- Many APIs require an access key or token
- An alternative, open standard is called OAuth
- Connections without sharing username or password, only temporary tokens that can be refreshed
- `httr` package in R implements most cases (examples)

R packages

Before starting a new project, worth checking if there's already an R package for that API. Where to look?

- CRAN Web Technologies Task View (but only packages released in CRAN)
- GitHub (including unreleased packages and most recent versions of packages)
- rOpenSci Consortium

Also see this great list of APIs in case you need inspiration.

Why APIs?

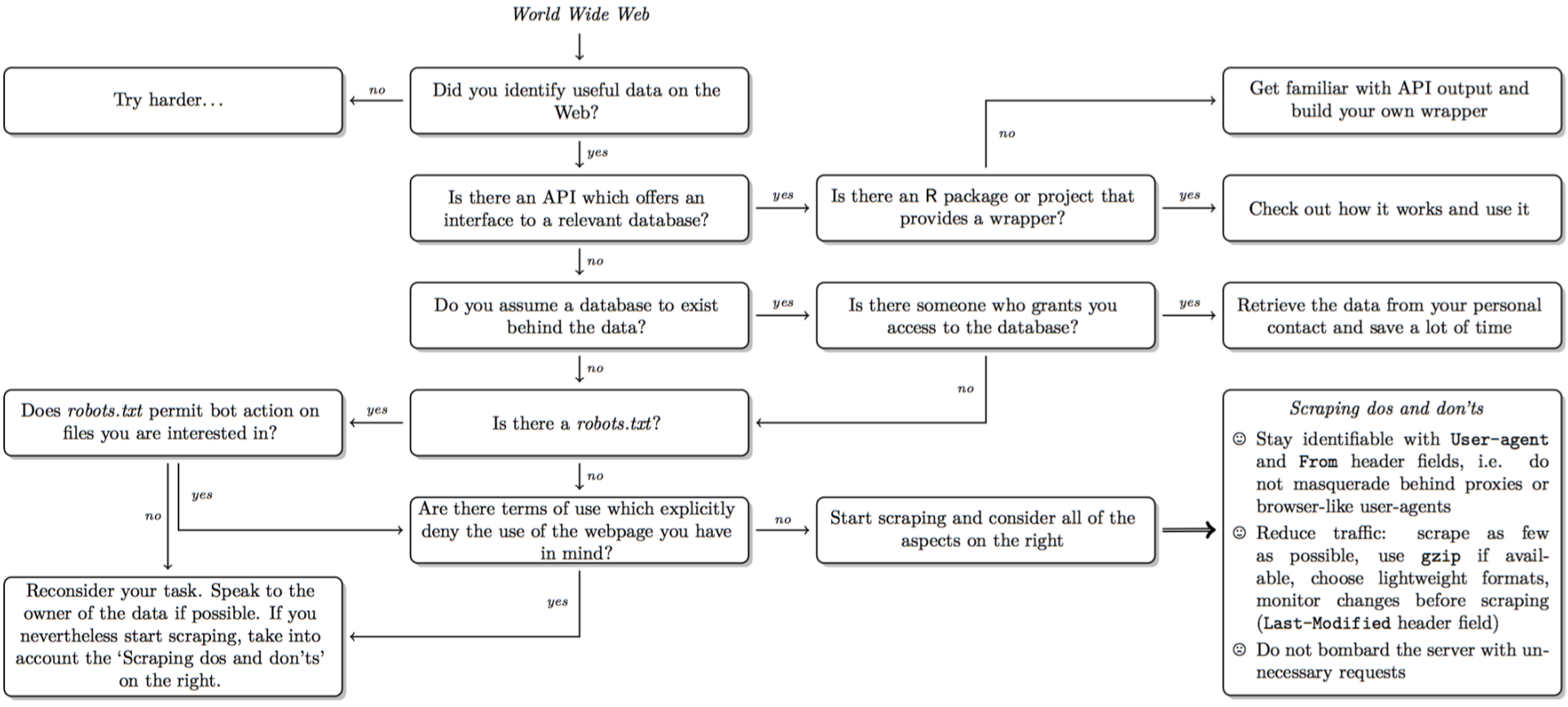
Advantages:

- 'Pure' data collection: avoid malformed HTML, no legal issues, clear data structures, more trust in data collection...
- Standardized data access procedures: transparency, replicability
- Robustness: benefits from 'wisdom of the crowds'

Disadvantages

- They're not too common (yet!)
- Dependency on API providers
- Rate limits

Decisions, decisions...



Example: the New York Times API

see `01-nytimes-api.Rmd`

Social media data

Social media data

What are the main advantages of using social media data to study human behavior?

1. Unobtrusive data collection at scale, e.g. in study of networks, censorship
2. Homogeneity in data format across actors, countries, and over time, e.g. in study of political rhetoric
3. Temporal and spatial data granularity, e.g. in study of geographic segregation
4. Increasing representativeness of social media users, e.g. in study of political elites

Social media research

Two different approaches in the growing field of social media research:

1. Social media as a new source of data
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
   - Affordable field experiments
2. How social media affects social behavior
   - Collective action and social movements
   - Political campaigns
   - Social capital and interpersonal communication
   - Political attitudes and behavior

Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature

By Joshua A. Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan

WILLIAM + FLORA Hewlett Foundation

Social media data and social science: challenges

1. Big data, big bias?
2. Spam and bots
3. The privacy paradox
4. Ethical concerns

1. Big data, big bias?

SOCIAL SCIENCES

Social media for large studies of behavior

Large-scale studies of human behavior in social media need to be held to higher methodological standards

By **Derek Ruths**[1]* *and* **Jürgen Pfeffer**[2]

On 3 November 1948, the day after Harry Truman won the United States presidential elections, the *Chicago Tribune* published one of the most famous erroneous headlines in newspaper history: "Dewey Defeats Truman" (*1*, *2*). The headline was informed by telephone surveys, which had inadver-

different social media platforms (*8*). For instance, Instagram is "especially appealing to adults aged 18 to 29, African-American, Latinos, women, urban residents" (*9*) whereas Pinterest is dominated by females, aged 25 to 34, with an average annual household income of $100,000 (*10*). These sampling biases are rarely corrected for (if even acknowledged).

Proprietary algorithms for public data. Platform-specific sampling problems, for example, the highest-volume source of pub-

The rise of "embedded researc
searchers who have special rela
with providers that give them ele
cess to platform-specific data, al
and resources) is creating a divic
media research community. Such
ers, for example, can see a platfor
workings and make accommodat
may not be able to reveal their c
or the data used to generate their f

Ruths and Pfeffer, 2015, "Social media for large studies of behavior", *Science*

Big data, big bias?

Sources of bias (Ruths and Pfeffer, 2015; Lazer et al, 2017)

- Population bias
  - Sociodemographic characteristics are correlated with presence on social media
- Self-selection within samples
  - Partisans more likely to post about politics (Barberá & Rivero, 2014)
- Proprietary algorithms for public data
  - Twitter API does not always return 100% of publicly available tweets (Morstatter et al, 2014)
- Human behavior and online platform design
  - e.g. *Google Flu* (Lazer et al, 2014)

1. Big data, big bias?

Reducing biases and flaws in social media data

DATA COLLECTION

- 1. Quantifies platform-specific biases (platform design, user base, platform-specific behavior, platform storage policies)
- 2. Quantifies biases of available data (access constraints, platform-side filtering)
- 3. Quantifies proxy population biases/mismatches

METHODS

- 4. Applies filters/corrects for nonhuman accounts in data
- 5. Accounts for platform and proxy population biases
  a. Corrects for platform-specific and proxy population biases
  OR
  b. Tests robustness of findings
- 6. Accounts for platform-specific algorithms
  a. Shows results for more than one platform
  OR
  b. Shows results for time-separated data sets from the same platform
- 7. For new methods: compares results to existing methods on the same data
- 8. For new social phenomena or methods or classifiers: reports performance on two or more distinct data sets (one of which was not used during classifier development or design)

Issues in evaluating data from social media. Large-scale social media studies of human behavior should i address issues listed and discussed herein (further discussion in supplementary materials).

Ruths and Pfeffer, 2015, "Social media for large studies of behavior", *Science*

2. Spam and bots

"Follow your coordinators. We need to start tweeting, all at the same time, using the hashtag #ItsTimeForMexico. . . and don't forget to retweet tweets from the candidate's account..."

Unidentified PRI campaign manager

minutes before the May 8, 2012 Mexican Presidential debate

2. Spam and bots

Ferrara et al, 2016, *Communications of the ACM*

3. The privacy paradox

Online data present a paradox in the protection of privacy: Data are at once too revealing in terms of privacy protection, yet also not revealing enough in terms of providing the demographic background information needed by social scientists.

Golder & Macy, *Digital footprints, 2014*

4. Ethical concerns

1. Shifting notion of *informed consent*

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]

[a]Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and Departments of [b]Communication and [c]Information Science, Cornell University, Ithaca, NY 14853

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

Emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. Emotional contagion is well established in laboratory experiments, with people transferring positive and negative emotions to others. Data from a large real-world social network, collected over a 20-y period suggests that longer-lasting moods (e.g., depression, happiness) can be transferred through networks [Fowler JH, Christakis NA (2008) *BMJ* 337:a2338], although the results are controversial. In an experiment with people who use Facebook, we test whether emotional contagion occurs

demonstrated that (*i*) emotional contagion occurs via text-based computer-mediated communication (7); (*ii*) contagion of psychological and physiological qualities has been suggested based on correlational data for social networks generally (7, 8); and (*iii*) people's emotional expressions on Facebook predict friends' emotional expressions, even days later (7) (although some shared experiences may in fact last several days). To date, however, there is no experimental evidence that emotions or moods are contagious in the absence of direct interaction between experiencer and target.

On Facebook, people frequently express emotions, which are

2. Most personal data can be de-anonymized

Ethics and Information Technology

December 2010, Volume 12, Issue 4, pp 313–325

"But the data is already public": on the ethics of research in Facebook

Authors Authors and affiliations

Twitter data

Twitter APIs

Two different methods to collect Twitter data:

1. REST API:
   - Queries for specific information about users and tweets
   - Search recent tweets
   - Examples: user profile, list of followers and friends, tweets generated by a given user ("timeline"), users lists, etc.
   - R library: tweetscores (also twitteR, rtweet)
2. Streaming API:
   - Connect to the "stream" of tweets as they are being published
   - Three streaming APIs:
     - 2.1 Filter stream: tweets filtered by keywords
     - 2.2 Geo stream: tweets filtered by location
     - 2.3 Sample stream: 1% random sample of tweets
   - R library: streamR

Important limitation: tweets can only be downloaded in real time (exception: user timelines, $\sim$ 3,200 most recent tweets are available)

Anatomy of a tweet

Anatomy of a tweet

Tweets are stored in JSON format:


```
{ "created_at": "Wed Nov 07 04:16:18 +0000 2012",
  "id": 266031293945503744,
  "text": "Four more years. http://t.co/bAJE6Vom",
  "source": "web",
  "user": {
    "id": 813286,
    "name": "Barack Obama",
    "screen_name": "BarackObama",
    "location": "Washington, DC",
    "description": "This account is run by Organizing for Action staff.
        Tweets from the President are signed -bo.",
    "url": "http://t.co/8aJ56Jcemr",
    "protected": false,
    "followers_count": 54873124,
    "friends_count": 654580,
    "listed_count": 202495,
    "created_at": "Mon Mar 05 22:08:25 +0000 2007",
    "time_zone": "Eastern Time (US & Canada)",
    "statuses_count": 10687,
    "lang": "en" },
  "coordinates": null,
  "retweet_count": 756411,
  "favorite_count": 288867,
  "lang": "en"
}
```

Streaming API

- Recommended method to collect tweets
- Potential issues:
  - Filter streams have same rate limit as spritzer: when volume reaches 1% of all tweets, it will return random sample
  - Good to restart stream connections regularly.
- My workflow:
  - Amazon EC2, cloud computing
  - Cron jobs to restart R scripts every hour.
  - Save tweets in .json files, one per day.

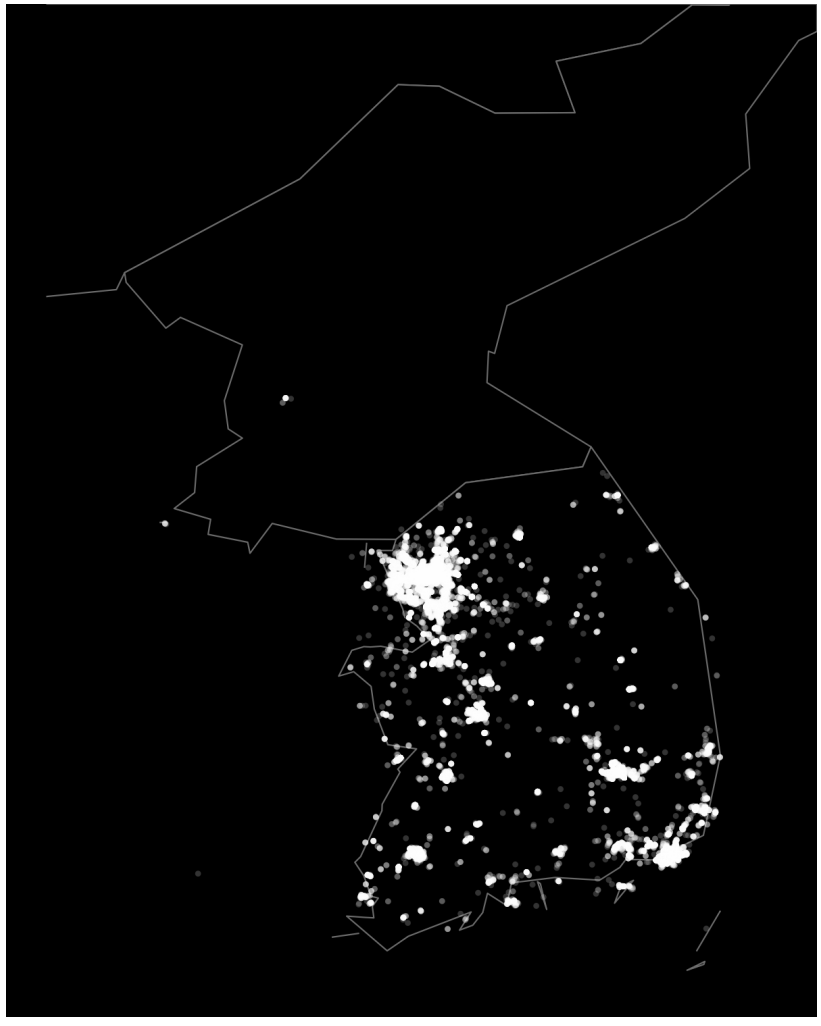
Sampling bias?

Morstatter et al, 2013, *ICWSM*, "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose":

- 1% random sample from Streaming API is not truly random
- Less popular hashtags, users, topics... less likely to be sampled
- But for keyword-based samples, bias is not as important

González-Bailón et al, 2014, *Social Networks*, "Assessing the bias in samples of large online networks":

- Small samples collected by filtering with a subset of relevant hashtags can be biased
- Central, most active users are more likely to be sampled
- Data collected via search (REST) API more biased than those collected with Streaming API



Tweets from Korea: 40k tweets collected in 2014 (left)
Korean peninsula at night, 2003 (right). Source: NASA.

Who is tweeting from North Korea?

Twitter user: @uriminzok_engl

Facebook data

Collecting Facebook data

Facebook used to allow access to public pages' data through the Graph API:

1. Posts on public pages and groups
2. Likes, reactions, comments, replies...

Currently not available.

Aggregate-level statistics available through the FB Marketing API. See the code by Connor Gilroy (UW)

Access to other (anonymized) data used in published studies requires permission from Facebook or from users.

Social Science One as a new model for academic partnerships with Facebook.

Example: Twitter API

see `02-twitter-streaming-api.Rmd`

see `03-twitter-rest-api.Rmd`