Outline

- Data APIs and working with JSON
- Data Cleaning with Open Refine
- Data Visualization

API Challenges

Application Programming Interfaces (APIs) give access to data -- but they're generally designed more for application development than research or journalistic purposes.

Examples:

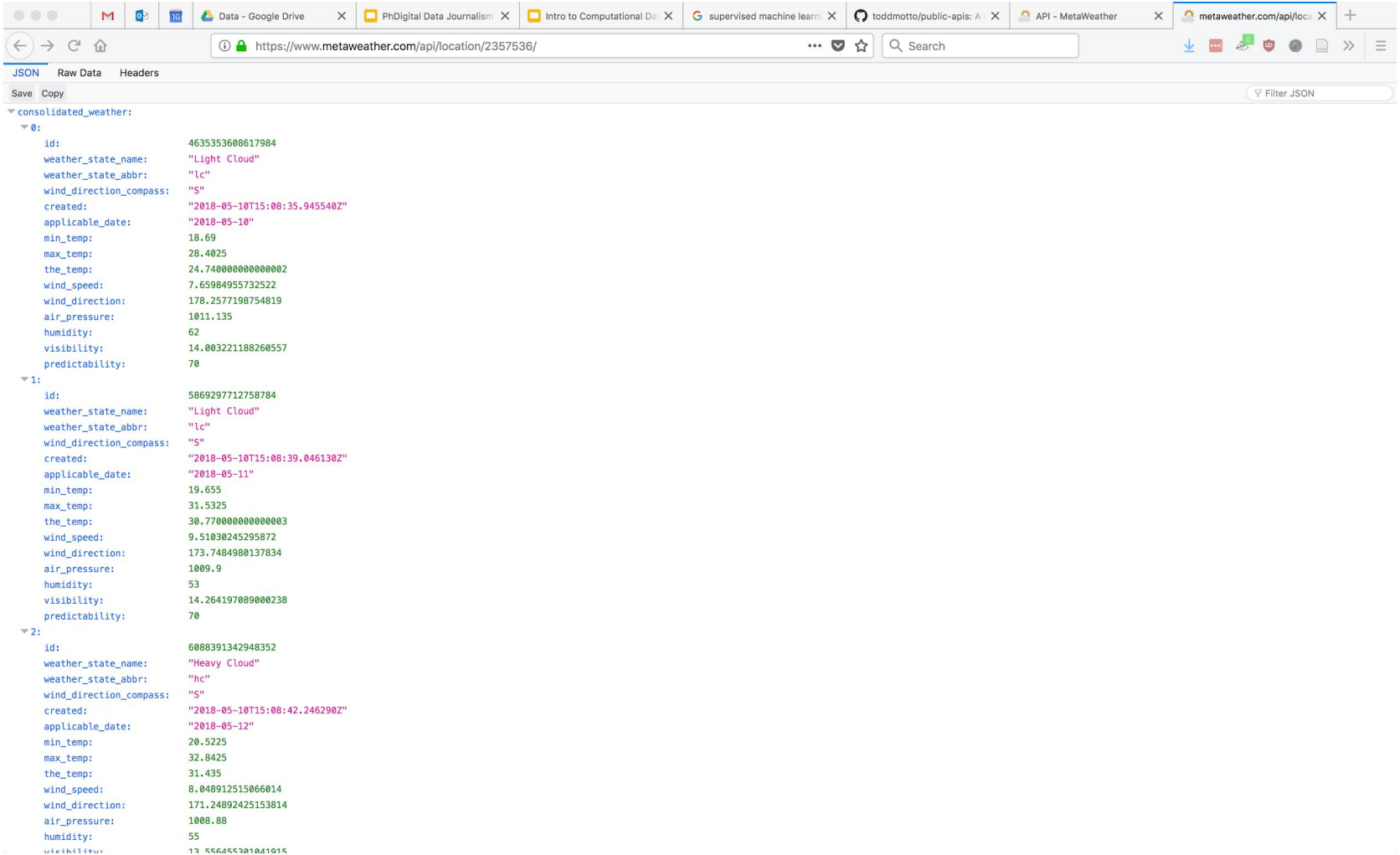
- Pulling someone's Instagram followers for use in your app.
- Pulling the current weather for use in your app.

APIs

APIs typically just return data in JSON format when you access a URL. So,

https://www.metaweather.com/api/location/2357536/

gives me the weather for Austin (which has id 2357536).


Data - Google Drive
PhDigital Data Journalism
Intro to Computational Da
supervised machine learni
toddmotto/public-apis: A
API - MetaWeather
metaweather.com/api/loca
https://www.metaweather.com/api/location/2357536/
Search
JSON
Raw Data
Headers
Save
Copy
Filter JSON
consolidated_weather:
0:
id: 4635353608617984
weather_state_name: "Light Cloud"
weather_state_abbr: "lc"
wind_direction_compass: "S"
created: "2018-05-10T15:08:35.945540Z"
applicable_date: "2018-05-10"
min_temp: 18.69
max_temp: 28.4025
the_temp: 24.740000000000002
wind_speed: 7.65984955732522
wind_direction: 178.2577198754819
air_pressure: 1011.135
humidity: 62
visibility: 14.003221188260557
predictability: 70
1:
id: 5869297712758784
weather_state_name: "Light Cloud"
weather_state_abbr: "lc"
wind_direction_compass: "S"
created: "2018-05-10T15:08:39.046130Z"
applicable_date: "2018-05-11"
min_temp: 19.655
max_temp: 31.5325
the_temp: 30.770000000000003
wind_speed: 9.51030245295872
wind_direction: 173.7484980137834
air_pressure: 1009.9
humidity: 53
visibility: 14.264197089000238
predictability: 70
2:
id: 6088391342948352
weather_state_name: "Heavy Cloud"
weather_state_abbr: "hc"
wind_direction_compass: "S"
created: "2018-05-10T15:08:42.246290Z"
applicable_date: "2018-05-12"
min_temp: 20.5225
max_temp: 32.8425
the_temp: 31.435
wind_speed: 8.048912515066014
wind_direction: 171.24892425153814
air_pressure: 1008.88
humidity: 55

APIs

https://venmo.com/api/v5/public?limit=1000000

APIs: Challenges

- Finding a useful data API for a data journalism project can be hard.

Mercedes-Benz	Telematics data, remotely access vehicle functions, car configurator, locate service dealers	`apiKey`	Yes	No	Go!
NHTSA	NHTSA Product Information Catalog and Vehicle Listing	No	Yes	Unknown	Go!

Video

API	Description	Auth	HTTPS	CORS	Link
An API of Ice And Fire	Game Of Thrones API	No	Yes	Unknown	Go!
Czech Television	TV programme of Czech TV	No	No	Unknown	Go!
Dailymotion	Dailymotion Developer API	`OAuth`	Yes	Unknown	Go!
Open Movie Database	Movie information	`apiKey`	Yes	Unknown	Go!
Ron Swanson Quotes	Television	No	Yes	Unknown	Go!
SWAPI	Star Wars Information	No	Yes	Unknown	Go!
TMDb	Community-based movie data	`apiKey`	Yes	Unknown	Go!
TVDB	Television data	`apiKey`	Yes	Unknown	Go!
TVMaze	TV Show Data	No	No	Unknown	Go!
Utelly	Check where a tv show or movie is available	`X-Mashape-Key`	Yes	Unknown	Go!
Vimeo	Vimeo Developer API	`OAuth`	Yes	Unknown	Go!
YouTube	Add YouTube functionality to your sites and apps	`OAuth`	Yes	Unknown	Go!

Weather

API	Description	Auth	HTTPS	CORS	Link
Dark Sky	Weather	`apiKey`	Yes	No	Go!

APIs: Challenges

- JSON can be hard to work with in large part because it can contain complex relationships.

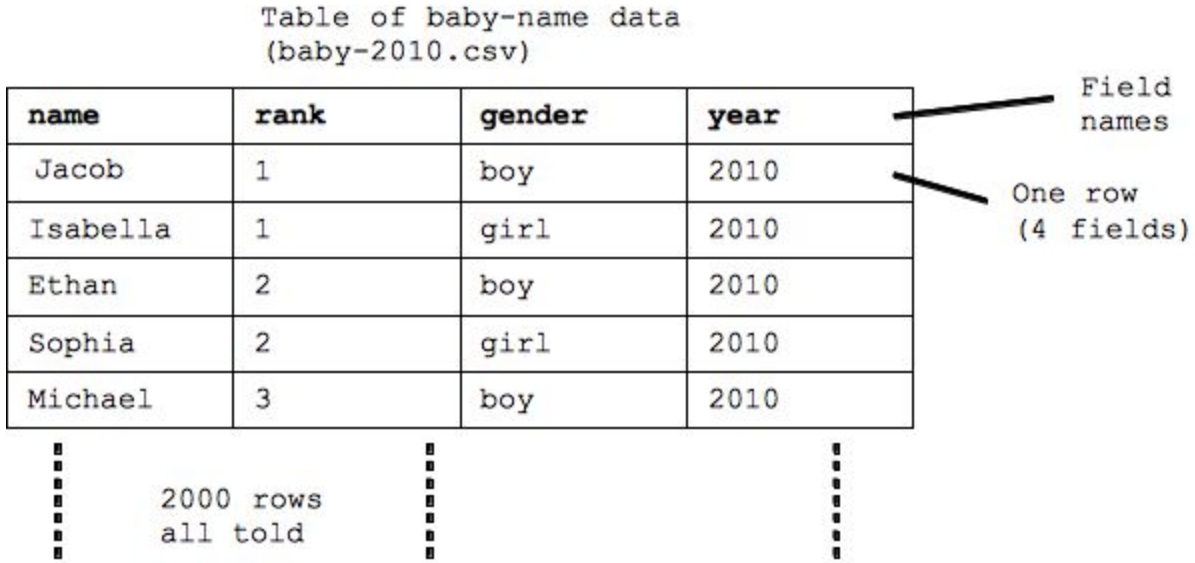


Table of baby-name data
(baby-2010.csv)
name
rank
gender
year
Jacob
1
boy
2010
Isabella
1
girl
2010
Ethan
2
boy
2010
Sophia
2
girl
2010
Michael
3
boy
2010
Field names
One row (4 fields)
2000 rows all told

```
{
    "products": [{
        "productid": "98429",
        "title": "Out Of My Own Book",
        "artist": "Dejan Milicevic",
        "releasedate": "2013-07-12",
        "catno": "plax100-8",
        "genre": "",
        "ean": "4250252410119",
        "upc": "881969712414",
        "cover": "http:\/\/images.deeep.net\/product\/98429.600.jpg",
        "tracks": [{
            "isrc": "DEAZ31310604",
            "artist": "Dejan Milicevic",
            "title": "Out Of My Own Book",
            "mixversion": "",
            "length": "00:06:32",
            "genre": "Dance > House > Deep",
            "snippet": "http:\/\/www.dcmc.info\/snippets\/566940",
            "player": "http:\/\/www.dcmc.info\/snippets\/566940"
        }, {
            "isrc": "DEAZ31310605",
            "artist": "Dejan Milicevic",
            "title": "Undefined Bank Of The 303",
            "mixversion": "",
            "length": "00:06:44",
            "genre": "Dance > House > Deep",
            "snippet": "http:\/\/www.dcmc.info\/snippets\/566941",
            "player": "http:\/\/www.dcmc.info\/snippets\/566941"
        }, {
            "isrc": "DEAZ31310606",
            "artist": "Dejan Milicevic",
            "title": "Virtual Librarians",
            "mixversion": ""
```

APIs: Challenges

- JSON can be hard to work with in large part because it can contain complex relationships.

  https://api.fda.gov/drug/event.json?search=patient.reaction.reactionmeddrapt:%22fatigue%22&limit=1

APIs: Challenges

- JSON can be hard to work with in large part because it can contain complex relationships.

  Convert JSON to table: https://konklone.io/json/

APIs: Challenges

- Lots of social media APIs return text or images, which are very difficult to work with using the techniques we tend to teach data journalists.
- Many of the more interesting APIs require access keys or don’t work using the URL formats we looked at earlier.

API Activity

- Find an interesting API from https://github.com/toddmotto/public-apis
- Figure out how to access data and examine what you get.

Cleaning Data

Messy data might result from:

- Human error
- Inconsistent terminology
- Inconvenient categories
- Data spread over multiple files

Cleaning Data

Cleaning data often takes around 80% of the total project time.

Cleaning Data

OpenRefine is a free and useful tool that’s also easy to teach with.

Activity: Clean some Texas Department of Criminal Justice data.

https://bit.ly/2k5okRn

Visualization Resources

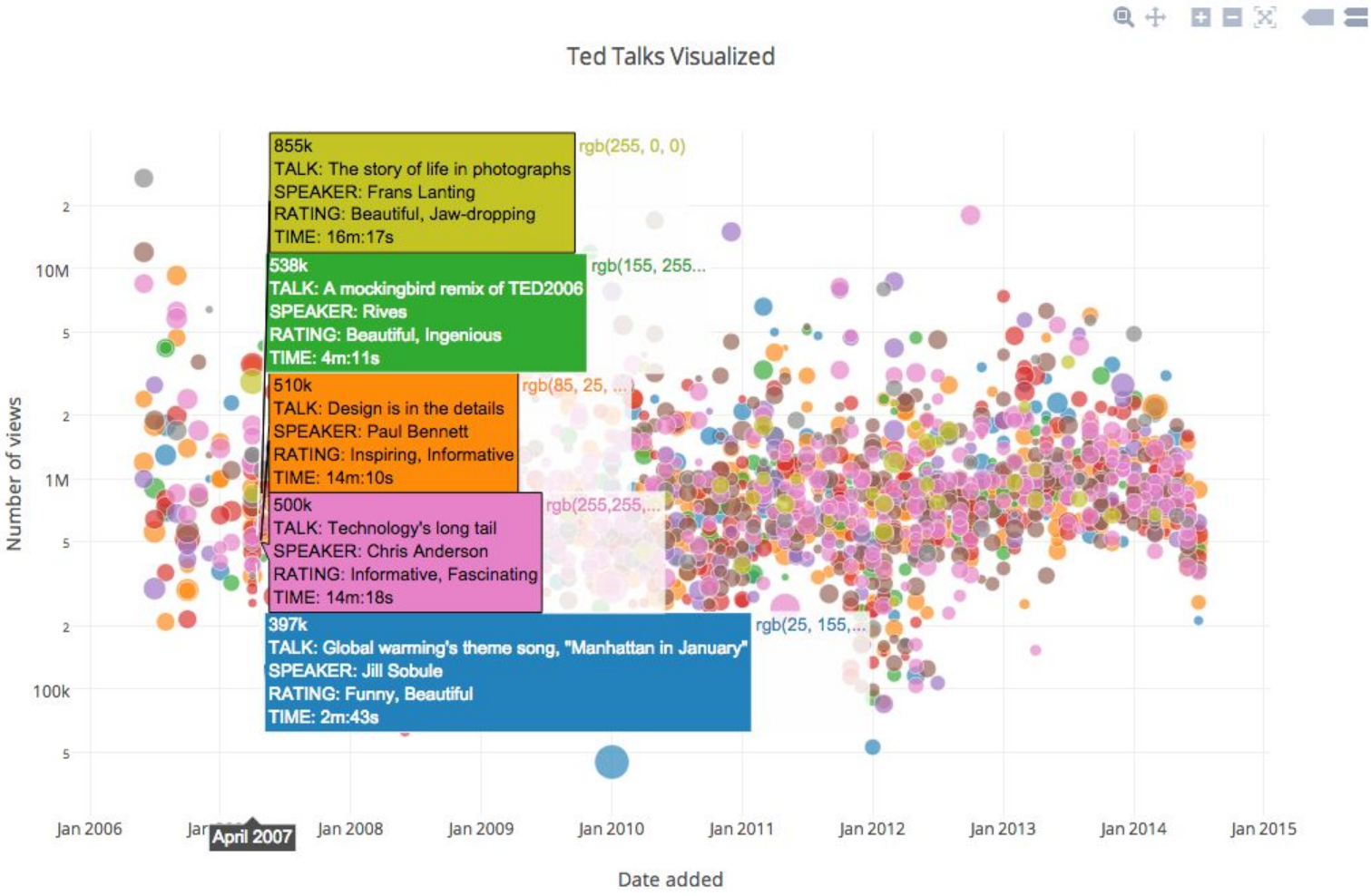
Tableau

- Free for students and labs (but then pretty expensive)
- Very easy to use and produces high quality, interactive products
- Used in a variety of professional settings

Visualization Resources

JavaScript libraries (e.g., Google charts, Plotly)

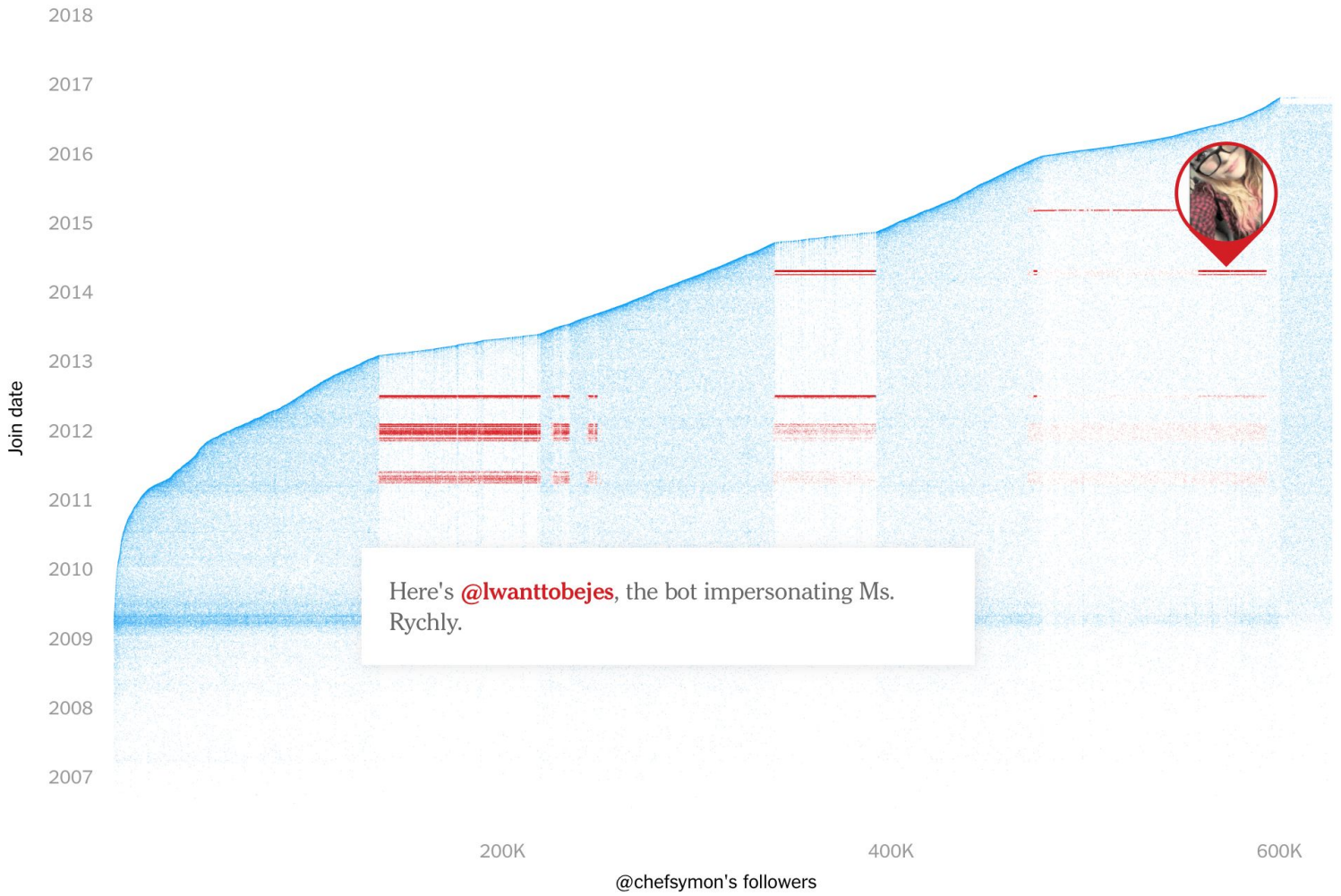
- Require minimal coding
- Used in professional settings
- Often limited in style and options


Ted Talks Visualized
855k
TALK: The story of life in photographs
SPEAKER: Frans Lanting
RATING: Beautiful, Jaw-dropping
TIME: 16m:17s
rgb(255, 0, 0)
538k
TALK: A mockingbird remix of TED2006
SPEAKER: Rives
RATING: Beautiful, Ingenious
TIME: 4m:11s
rgb(155, 255...
510k
TALK: Design is in the details
SPEAKER: Paul Bennett
RATING: Inspiring, Informative
TIME: 14m:10s
rgb(85, 25, ...)
500k
TALK: Technology's long tail
SPEAKER: Chris Anderson
RATING: Informative, Fascinating
TIME: 14m:18s
rgb(255,255,...
397k
TALK: Global warming's theme song, "Manhattan in January"
SPEAKER: Jill Sobule
RATING: Funny, Beautiful
TIME: 2m:43s
rgb(25, 155,...
Number of views
2
10M
5
2
1M
5
2
100k
5
Jan 2006
Jan
April 2007
Jan 2008
Jan 2009
Jan 2010
Jan 2011
Jan 2012
Jan 2013
Jan 2014
Jan 2015
Date added

Visualization Resources

JavaScript libraries (D3)

- Require advanced, idiosyncratic coding
- Allows many more choices


2018
2017
2016
2015
2014
2013
2012
2011
2010
2009
2008
2007
Join date
200K
400K
600K
@chefsymon's followers
Here's @lwanttobejes, the bot impersonating Ms. Rychly.

Visual Literacy

Communication students come in with very little experience interpreting graphs or other forms of visual evidence.


Stockholm
Singapore
Tokyo
London
Shanghai
Los Angeles
Madrid
Johan...
Seoul
Beijing
Jakarta
São Pa...
Berlin
Mexic...
Istanbul
Rome
Montreal
Cairo
Bangkok
Barcelona
Hong Kong
Cape Town
Rio de Jan...
Moscow
Casab...
Delhi
Dubrovnik
Praha
Mumbai
New York City
Paris

Visual Literacy

- Nathan Yau, *Data Points*.
- *Data + Design*. https://infoactive.co/data-design