

# Computational Analysis of Big Data

Week 3

## Getting data

## Overview

*The data is already there*

# *The data is already there*

- Company data can give insight into your own **research** questions
- Market data can inform about **trends** that you may want to react upon
- You can **sell it**
- You can **combine multiple sources** and get unique insight that no one else has

# Why do we want *other people's* data?

- Company data can give insight into your own **research** questions
- Market data can inform about **trends** that you may want to react upon
- You can **sell it**
- You can **combine multiple sources** and get unique insight that no one else has

## How Facebook could stop a disease outbreak

January 3, 2018



Credit: National Cancer Institute

Facebook accounts and telephone records can be used to pinpoint the best individuals to vaccinate to stop a disease outbreak in its tracks, researchers said Wednesday.

Such people would be "central" in their social networks, and thus likelier to spread disease-causing germs from one group to another.

Assuming there is an outbreak, and not enough vaccines for every person in the world, immunising these well-connected individuals would remove social "bridges" by which germs can spread, experts wrote in the *Journal of the Royal Society Interface*.

# Why do we want *other people's* data?

- Company data can give insight into your own **research** questions
- Market data can inform about **trends** that you may want to react upon
- You can **sell it**
- You can **combine multiple sources** and get unique insight that no one else has



# Why do we want *other people's* data?

- Company data can give insight into your own **research** questions
- Market data can inform about **trends** that you may want to react upon
- You can **sell it**
- You can **combine multiple sources** and get unique insight that no one else has



# Why do we want *other people's* data?

- Company data can give insight into your own **research** questions
- Market data can inform about **trends** that you may want to react upon
- You can **sell it**
- You can **combine multiple sources** and get unique insight that no one else has

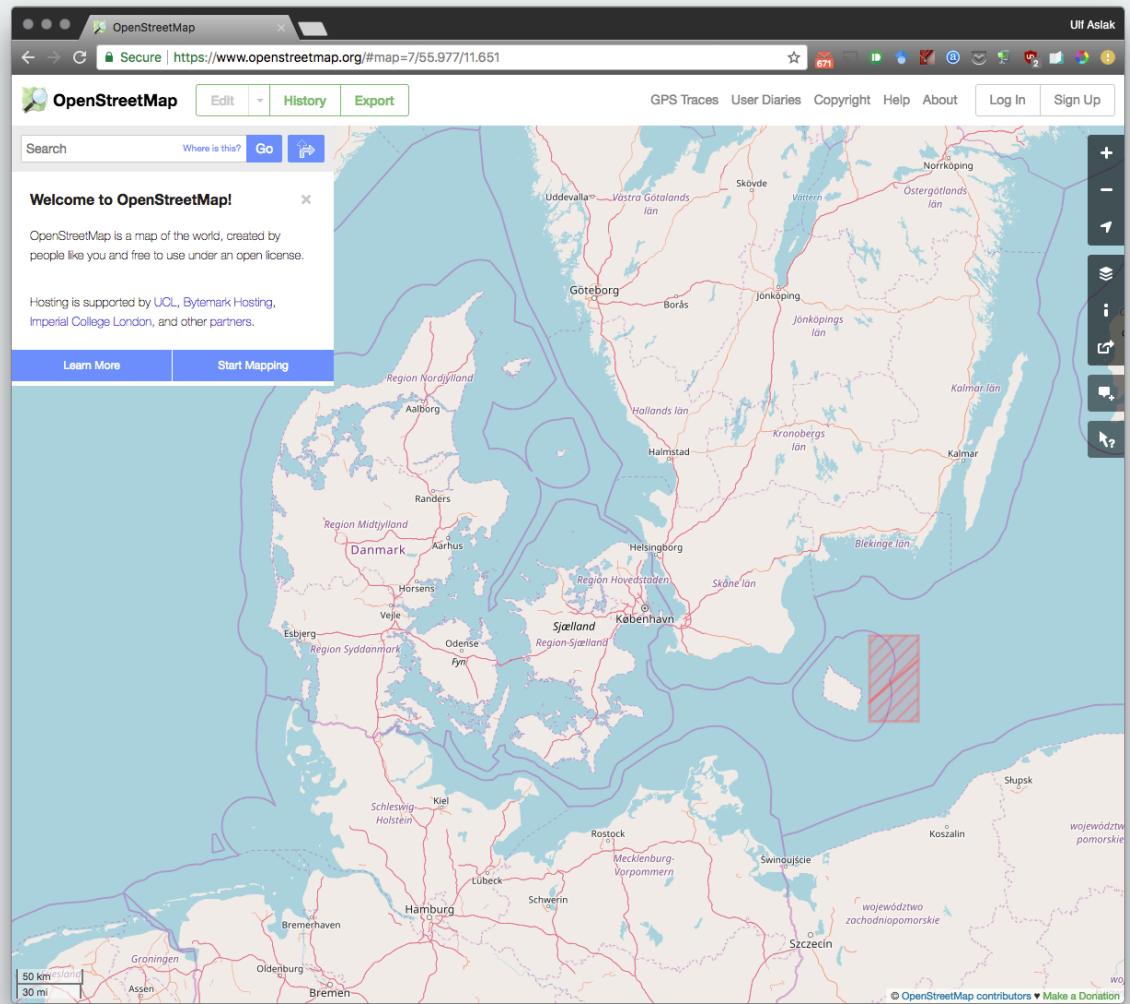
# There are three main ways of getting existing data

- Get **open data** from public institutions, researchers or data sharing sites.
- Request it from someone's **API**. Is very easy, but usually has limits.
- Forcefully take it by **scraping** it from a website

# Open data

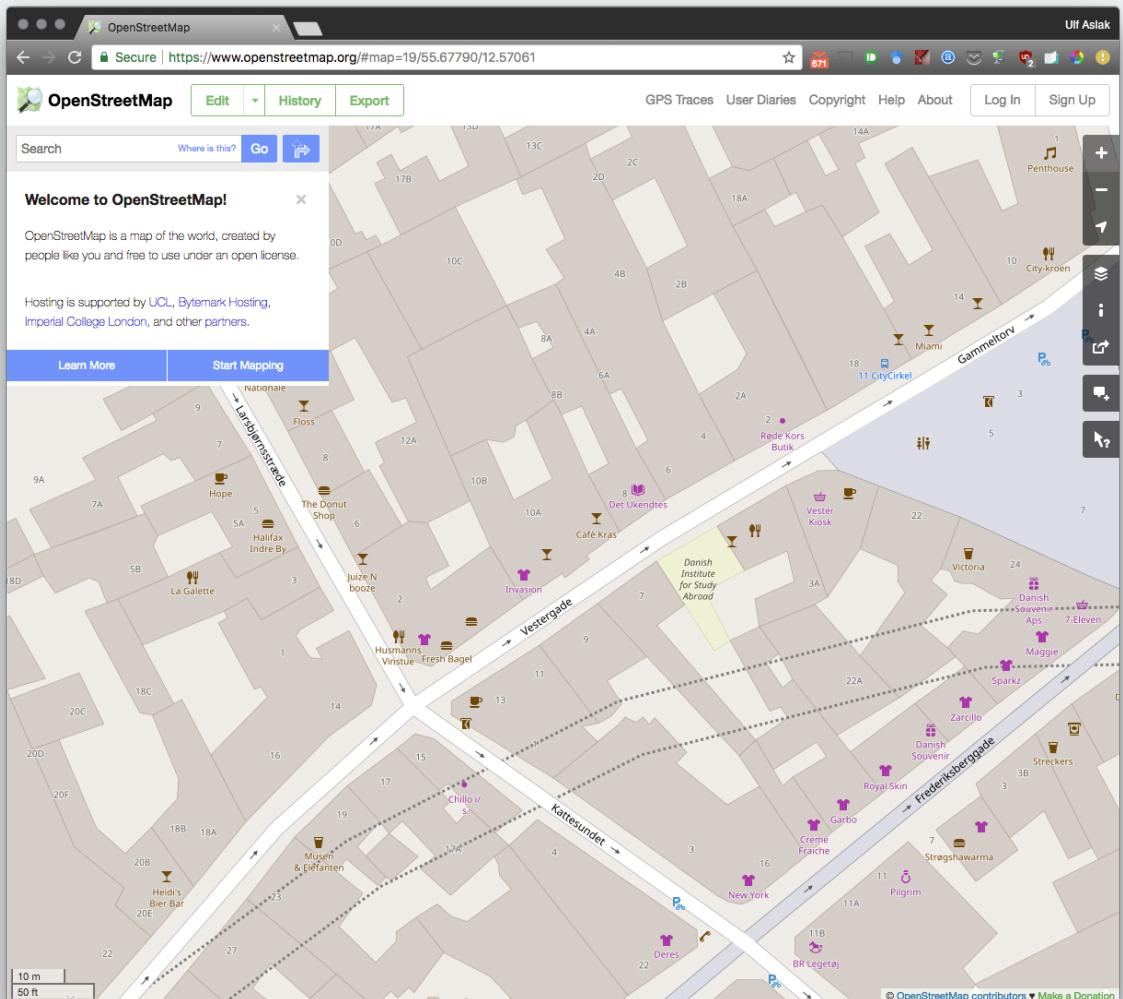
# Open data

- **Geographical data**
- City data
- Political data
- Research data
- Competition datasets
- Transactional data



# Open data

- Geographical data
- City data
- Political data
- Research data
- Competition datasets
- Transactional data



[Learn How to Map in OpenStreetMap](#)

# Open data

- **Geographical data**
- City data
- Political data
- Research data
- Competition datasets
- Transactional data

OpenStreetMap powers map data on thousands of web sites, mobile apps, and hardware devices

OpenStreetMap is built by a community of mappers that contribute and maintain data about roads, trails, cafés, railway stations, and much more, all over the world.

#### Local Knowledge

OpenStreetMap emphasizes local knowledge. Contributors use aerial imagery, GPS devices, and low-tech field maps to verify that OSM is accurate and up to date.

#### Community Driven

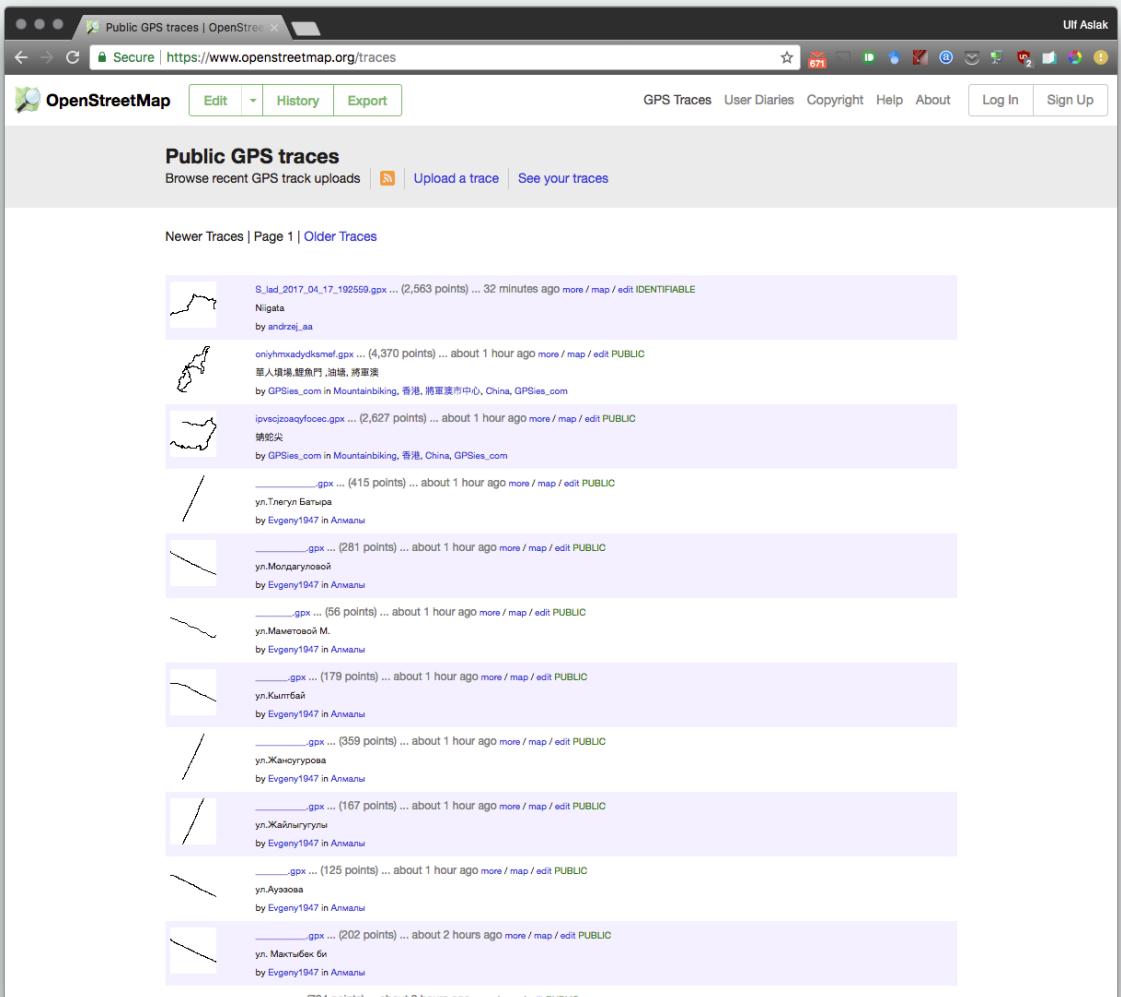
OpenStreetMap's community is diverse, passionate, and growing every day. Our contributors include enthusiast mappers, GIS professionals, engineers running the OSM servers, humanitarians mapping disaster-affected areas, and many more. To learn more about the community, see the [OpenStreetMap Blog](#), [user diaries](#), [community blogs](#), and the [OSM Foundation](#) website.

#### Open Data

OpenStreetMap is *open data*: you are free to use it for any purpose as long as you

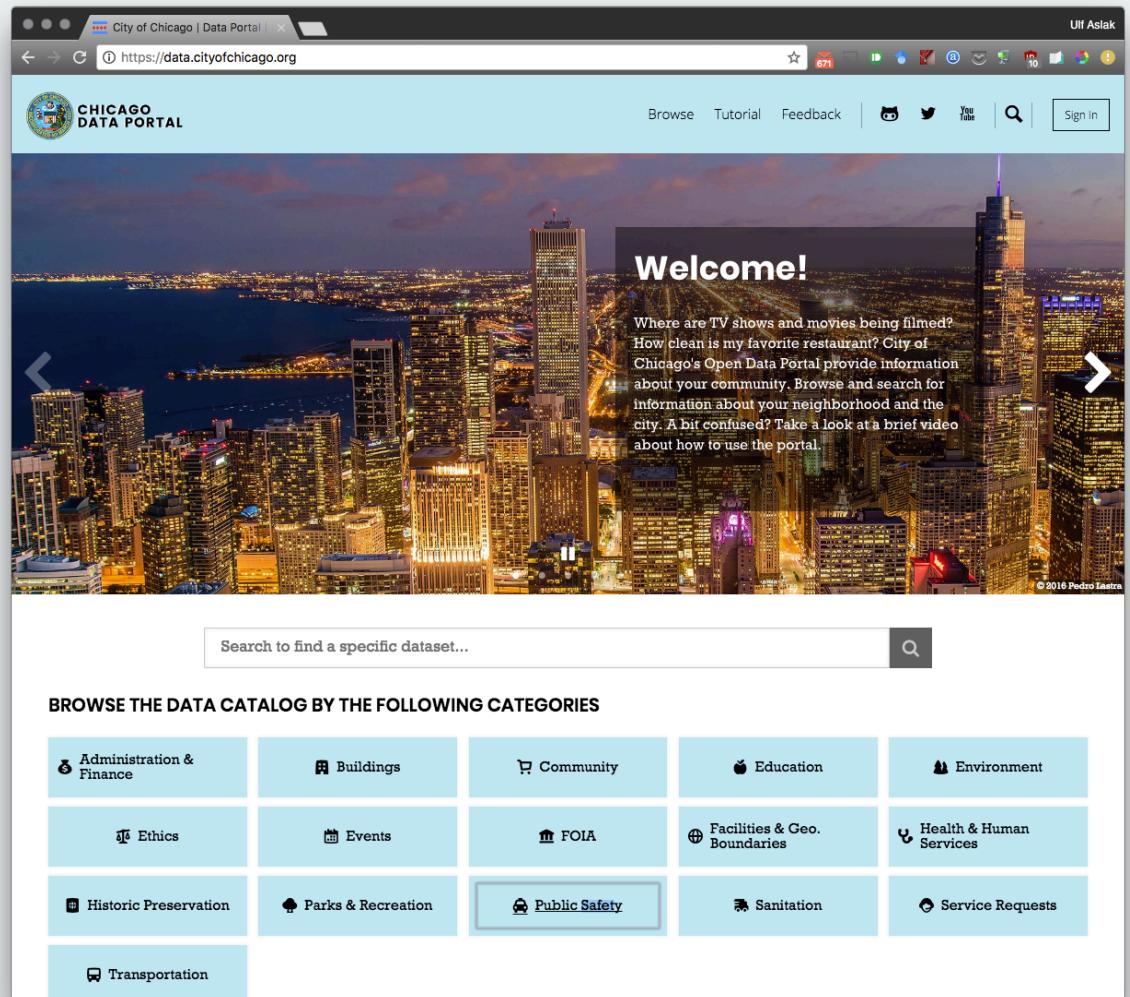
# Open data

- **Geographical data**
- City data
- Political data
- Research data
- Competition datasets
- Transactional data



# Open data

- Geographical data
- **City data**
- Political data
- Research data
- Competition datasets
- Transactional data



# Open data

- Geographical data
- **City data**
- Political data
- Research data
- Competition datasets
- Transactional data

The screenshot shows a web browser displaying the Chicago Data Portal at <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/jzp-q8t2>. The page title is "Crimes - 2001 to present" under the "Public Safety" category. The top navigation bar includes "Browse", "Tutorial", "Feedback", and social media links. A sidebar on the right shows the dataset was updated on January 31, 2018, by the Chicago Police Department.

**Featured Content Using this Data:**

- Crimes - 2001 to present - Dashboard**: Updated January 31, 2018, with 1.01M Views. Description: This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. In order to protect the privacy of crime victims...
- Crimes - 2001 to present - Map**: Updated January 31, 2018, with 47K Views. Description: This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of...
- Crimes - 2018**: Updated January 31, 2018, with 331 Views. Description: This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of...

**About this Dataset**

Updated	Metadata
<b>January 31, 2018</b>	Time Period: 2001 to present, minus the most recent seven days
Data Last Updated: January 31, 2018	Frequency: Data are updated daily.
Metadata Last Updated: September 27, 2017	Data Owner: Police
Date Created: September 30, 2011	
Views: 1.01M	Topics: Public Safety
Downloads: 47K	

# Open data

- Geographical data
- **City data**
- Political data
- Research data
- Competition datasets
- Transactional data

The screenshot shows the homepage of the Copenhagen Data website. At the top, there is a navigation bar with links for 'Dataset', 'organizations', 'groups', and 'About'. A search bar with the placeholder 'Søg datasæt...' and a magnifying glass icon is also present. The main content area displays a list of 269 data sets found, sorted by relevance. The first dataset listed is 'Air pollution', which is described as being measured with automatic instruments every hour. It includes a link to a CSV file. Below this, there are sections for 'Tables and benches' (described as a retired master register) and 'Base map' (a basic map of Copenhagen). On the left side, there is a sidebar with a tree view of organizations and groups. Under 'organizations', categories like 'The City of Copenhagen' (128), 'Health Care' (45), and 'ITS' (21) are listed. Under 'groups', 'Geography' (56) and 'Transport and infras ... (27)' are shown. There are also buttons for 'dwg', 'dGN', and 'ZIP'.

# Open data

- Geographical data
- City data
- Political data
- Research data
- Competition datasets
- Transactional data

The screenshot shows a search results page from CONGRESS.GOV. The search query is "homeland security", with a subtitle "medicare". The results are filtered by "Congressional Record". There are 1-100 of 577,488 results, displayed in 100 per page view, sorted by Issue Date - Newest to Old. The results are categorized under "CONGRESSIONAL RECORD" and include:

- 1. Daily Digest - Next Meeting of the SENATE + Next Meeting of the HOUSE OF REPRESENTATIVES + Other End Matter**  
Issue and Section: January 30, 2018 - Daily Digest (Vol. 164, No. 20)  
Page: D105 (PDF)
- 2. Daily Digest - COMMITTEE MEETINGS FOR 2018-02-02**  
Issue and Section: January 30, 2018 - Daily Digest (Vol. 164, No. 20)  
Page: D105 (PDF)
- 3. Daily Digest - NEW PUBLIC LAWS**  
Issue and Section: January 30, 2018 - Daily Digest (Vol. 164, No. 20)  
Page: D105 (PDF)
- 4. Daily Digest - House Committee Meetings**  
Issue and Section: January 30, 2018 - Daily Digest (Vol. 164, No. 20)  
Page: D103 (PDF)
- 5. Daily Digest - House of Representatives**  
Issue and Section: January 30, 2018 - Daily Digest (Vol. 164, No. 20)  
Page: D102 (PDF)
- 6. Daily Digest - Senate Committee Meetings**  
Issue and Section: January 30, 2018 - Daily Digest (Vol. 164, No. 20)  
Page: D101 (PDF)

Below the results, there's a link to "Digest - Highlights + Senate".

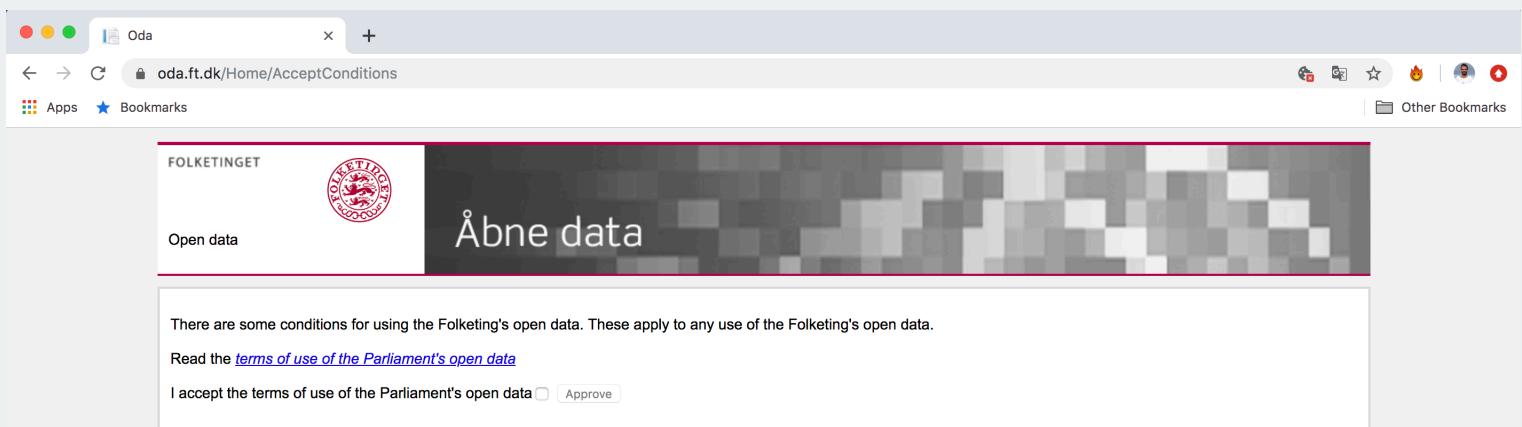
# Open data

- Geographical data
- City data
- Political data
- Research data
- Competition datasets
- Transactional data

The screenshot shows a web browser displaying the CONGRESS.GOV website at <https://www.congress.gov/about/data>. The page title is "Bill Status Bulk Data". The left sidebar has a "Related" section with links like "Explore a Bill", "About Accounts", "Creating and Using Email Alerts", etc. The main content area discusses the availability of Bill Status data from GPO's FDsys bulk data repository. It provides instructions for importing data into spreadsheets and databases, and links to "Linking to Congress.gov" and "How to embed a Congress.gov search box on your website". Below this, there's information about Congressional bulk data from the House and Senate, including links to "Legislative Documents in XML at the United States House of Representatives" and "XML Sources Available on Senate.gov". A "Bill Status Bulk Data" section is also present. At the bottom, there's a "CONGRESS.GOV" footer with links to Site Content, Help, Resources, House Links, and Senate Links.

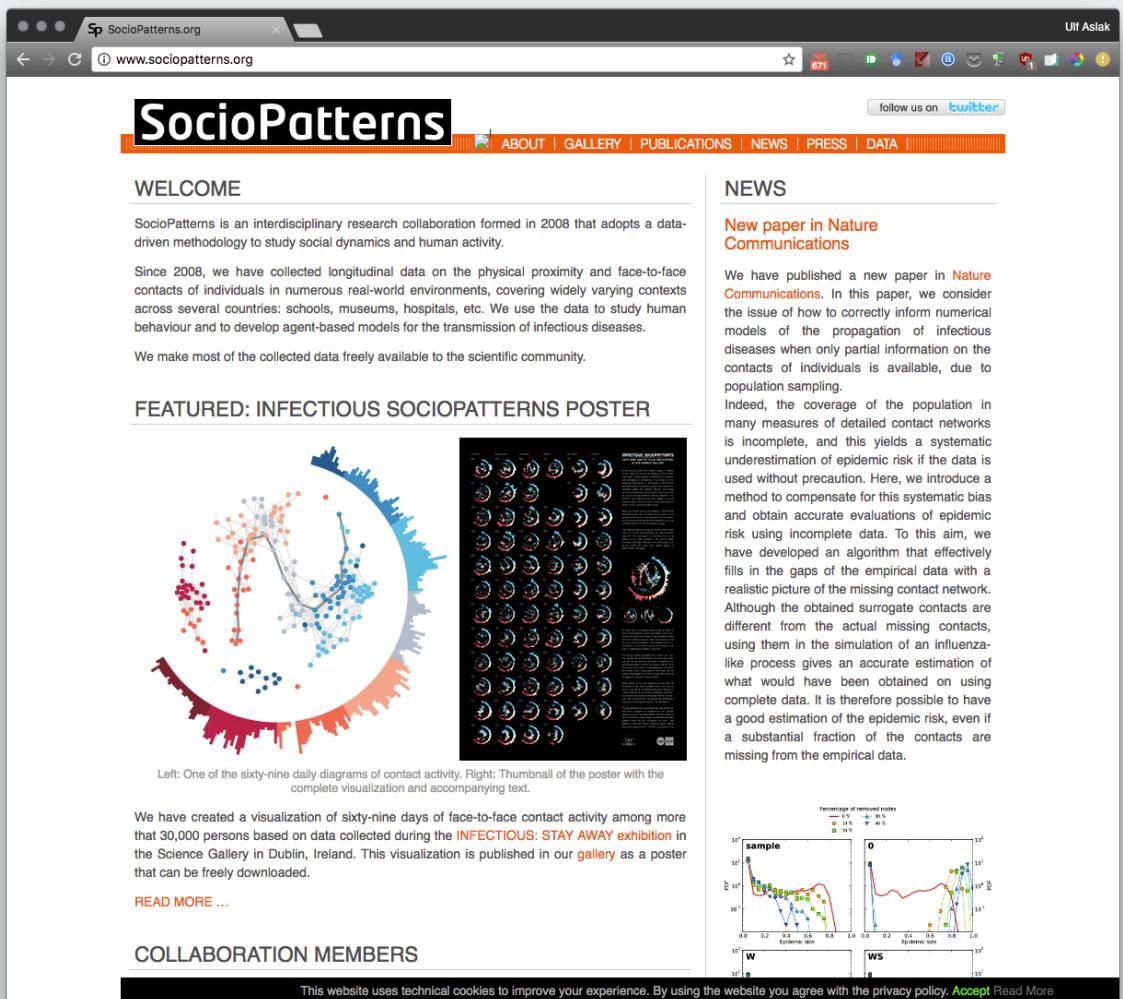
# Open data

- Geographical data
- City data
- Political data
- Research data
- Competition datasets
- Transactional data



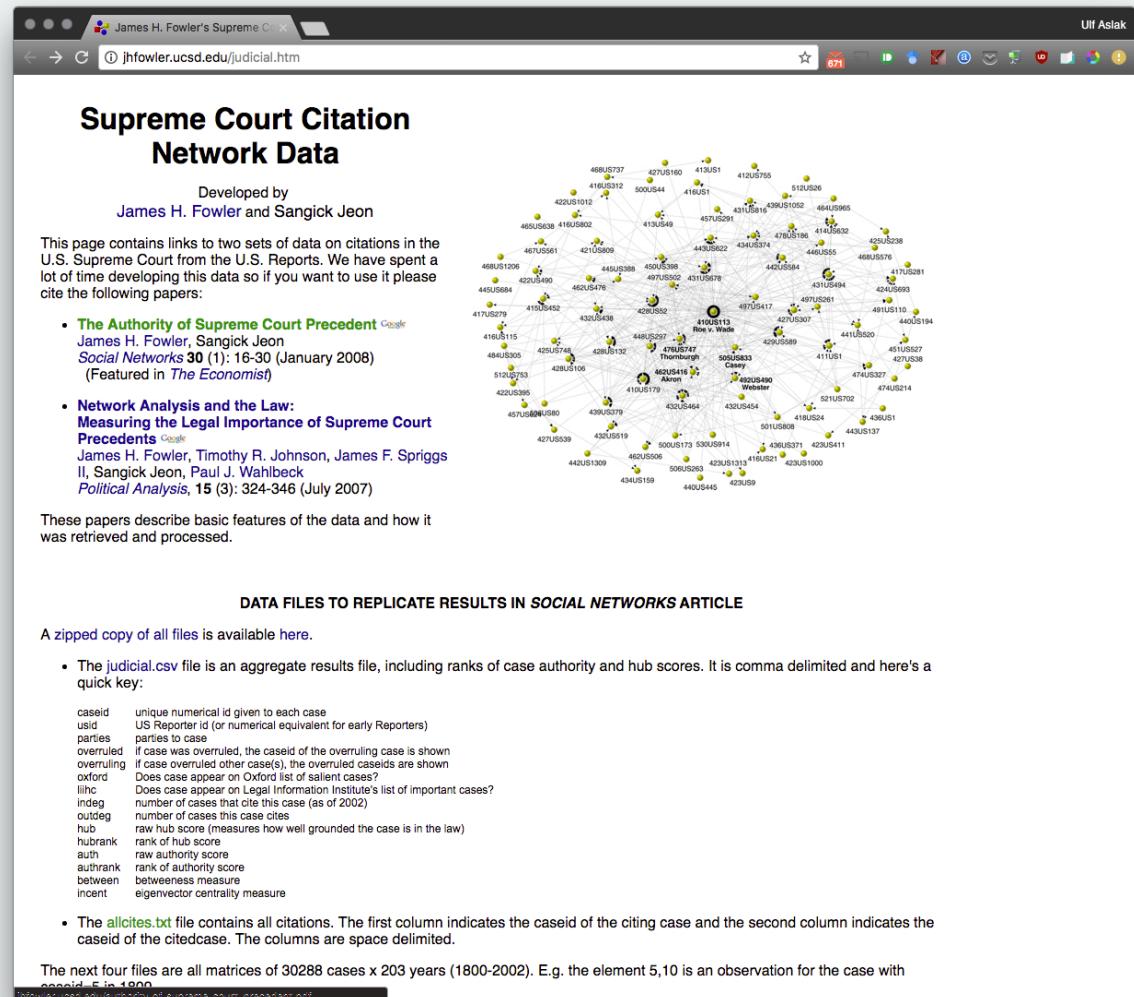
# Open data

- Geographical data
- City data
- Political data
- **Research data**
- Competition datasets
- Transactional data



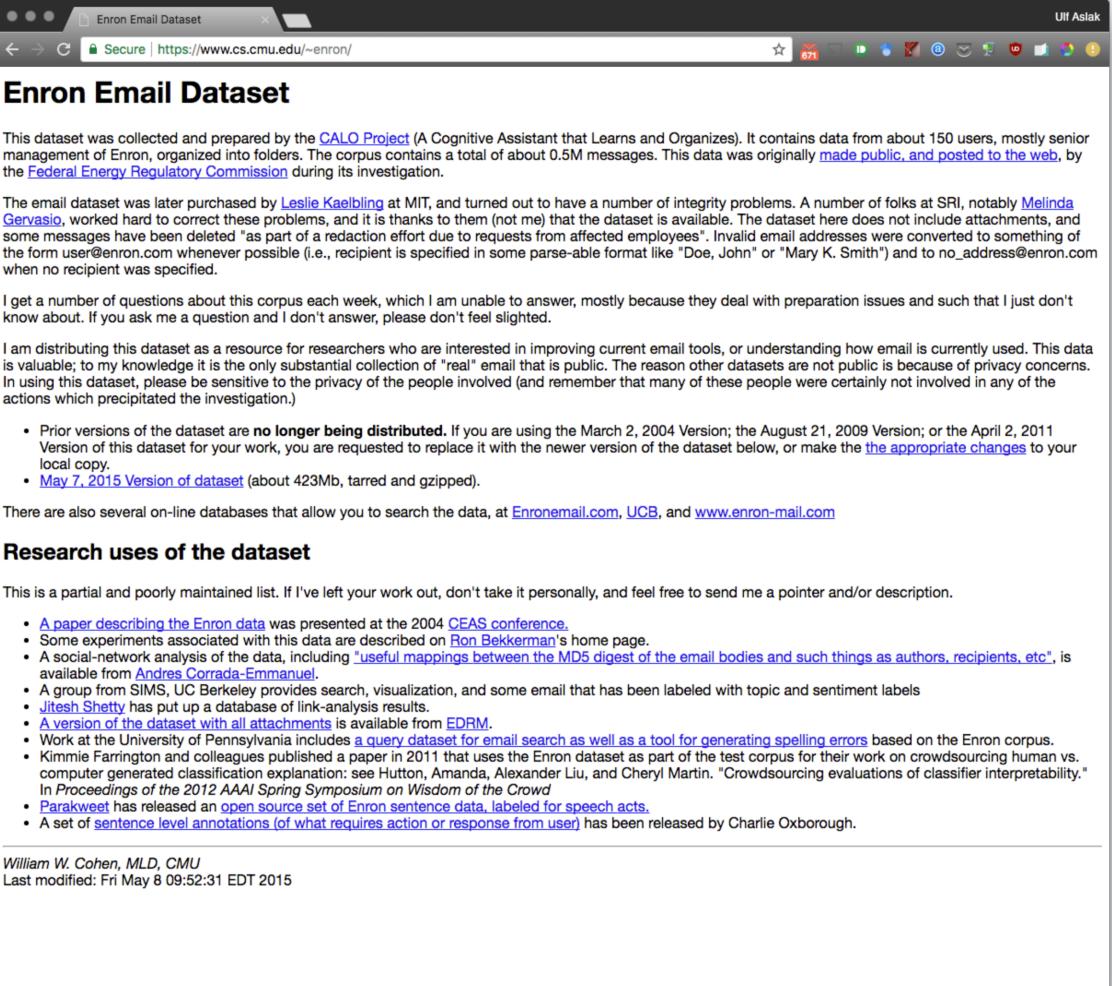
# Open data

- Geographical data
- City data
- Political data
- Research data
- Competition datasets
- Transactional data



# Open data

- Geographical data
- City data
- Political data
- Research data
- Competition datasets
- Transactional data

A screenshot of a web browser displaying the "Enron Email Dataset" page. The page title is "Enron Email Dataset". The content discusses the dataset's history, mentioning the CALO Project, Melinda Gervasio, and the Federal Energy Regulatory Commission. It notes integrity issues and the availability of different versions. A section on "Research uses of the dataset" lists various academic and practical applications, including a paper at the 2004 CEAS conference, social-network analysis, and work at the University of Pennsylvania. A footer provides authorship information for William W. Cohen.

This dataset was collected and prepared by the [CALO Project](#) (A Cognitive Assistant that Learns and Organizes). It contains data from about 150 users, mostly senior management of Enron, organized into folders. The corpus contains a total of about 0.5M messages. This data was originally [made public, and posted to the web](#), by the [Federal Energy Regulatory Commission](#) during its investigation.

The email dataset was later purchased by [Leslie Kaelbling](#) at MIT, and turned out to have a number of integrity problems. A number of folks at SRI, notably [Melinda Gervasio](#), worked hard to correct these problems, and it is thanks to them (not me) that the dataset is available. The dataset here does not include attachments, and some messages have been deleted "as part of a redaction effort due to requests from affected employees". Invalid email addresses were converted to something of the form user@enron.com whenever possible (i.e., recipient is specified in some parseable format like "Doe, John" or "Mary K. Smith") and to no\_address@enron.com when no recipient was specified.

I get a number of questions about this corpus each week, which I am unable to answer, mostly because they deal with preparation issues and such that I just don't know about. If you ask me a question and I don't answer, please don't feel slighted.

I am distributing this dataset as a resource for researchers who are interested in improving current email tools, or understanding how email is currently used. This data is valuable; to my knowledge it is the only substantial collection of "real" email that is public. The reason other datasets are not public is because of privacy concerns. In using this dataset, please be sensitive to the privacy of the people involved (and remember that many of these people were certainly not involved in any of the actions which precipitated the investigation.)

- Prior versions of the dataset are [no longer being distributed](#). If you are using the March 2, 2004 Version; the August 21, 2009 Version; or the April 2, 2011 Version of this dataset for your work, you are requested to replace it with the newer version of the dataset below, or make the [the appropriate changes](#) to your local copy.
- [May 7, 2015 Version of dataset](#) (about 423Mb, tarred and gzipped).

There are also several on-line databases that allow you to search the data, at [Enronemail.com](#), [UCB](#), and [www.enron-mail.com](#)

**Research uses of the dataset**

This is a partial and poorly maintained list. If I've left your work out, don't take it personally, and feel free to send me a pointer and/or description.

- [A paper describing the Enron data](#) was presented at the 2004 [CEAS conference](#).
- Some experiments associated with this data are described on [Ron Bekkerman's home page](#).
- A social-network analysis of the data, including "[useful mappings between the MD5 digest of the email bodies and such things as authors, recipients, etc](#)", is available from [Andres Corrada-Emmanuel](#).
- A group from SIMS, UC Berkeley provides search, visualization, and some email that has been labeled with topic and sentiment labels
- [Jitesh Shetty](#) has put up a database of link-analysis results.
- [A version of the dataset with all attachments](#) is available from [EDRM](#).
- Work at the University of Pennsylvania includes [a query dataset for email search as well as a tool for generating spelling errors](#) based on the Enron corpus.
- Kimmie Farrington and colleagues published a paper in 2011 that uses the Enron dataset as part of the test corpus for their work on crowdsourcing human vs. computer generated classification explanation: see Hutton, Amanda, Alexander Liu, and Cheryl Martin. "Crowdsourcing evaluations of classifier interpretability." In *Proceedings of the 2012 AAAI Spring Symposium on Wisdom of the Crowd*
- [Parakweet](#) has released an open source set of Enron sentence data, labeled for speech acts.
- A set of [sentence level annotations](#) (of what requires action or response from user) has been released by Charlie Oxborough.

William W. Cohen, MLD, CMU  
Last modified: Fri May 8 09:52:31 EDT 2015

# Open data

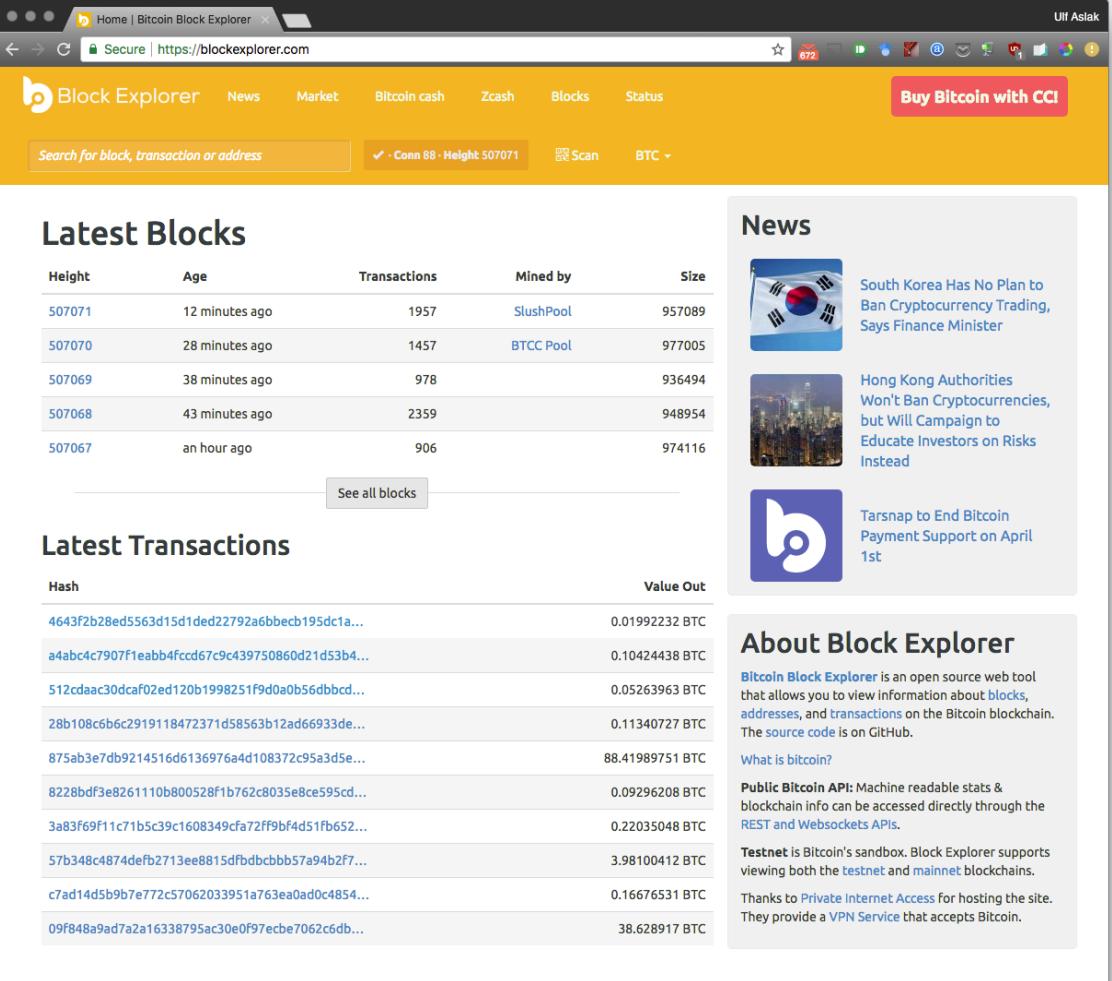
- Geographical data
- City data
- Political data
- Research data
- **Competition datasets**
- Transactional data

The screenshot shows a web browser displaying the Kaggle datasets page at <https://www.kaggle.com/datasets>. The page is titled "Datasets | Kaggle". At the top, there are tabs for "Public", "Your Datasets", and "Favorites". Below the tabs, a search bar says "Sort by Hotness". The main content area displays a list of 10,348 datasets. Each dataset entry includes a thumbnail, the number of rows (e.g., 84, 15, 31, 28, 201, 241, 196, 224), the dataset name, a brief description, the creator, and the last update time. To the right of each entry, there are tags, file formats (e.g., CSV, BigQuery), sizes, and download counts.

Dataset ID	Name	Description	Creator	Last Update	Tags	File Formats	Size	Downloads
84	<b>Chocolate Bar Ratings</b>	Expert ratings of over 1,700 chocolate bars	Rachael Tatman	updated 6 months ago	critical theory, food and drink	CSV, CCO	125 KB	52
15	<b>Hacker News</b>	All posts from Y Combinator's social news website from 2006 to late 2017	Hacker News	updated 2 months ago	journalism, information techn..., internet, big query	BigQuery, CCO	14 GB	3
31	<b>Historical Air Quality</b>	Air Quality Data Collected at Outdoor Monitors Across the US	US Environmental Protection Agency	updated 2 months ago	pollution	BigQuery, CCO	323 GB	2
28	<b>GitHub Repos</b>	Code and comments from 2.8 million repos	Github	updated 2 months ago	programming lang..., programming, software engineer...	BigQuery, Other	3 TB	9
201	<b>TED Talks</b>	Data about TED Talks on the TED.com website until September 21st, 2017	Rounak Banik	updated 4 months ago	Clothing	CSV, CC4	34 MB	23
241	<b>Fashion MNIST</b>	An MNIST-like dataset of 70,000 28x28 labeled fashion images	Zalando Research	updated 2 months ago	multiclass classifi..., object identification	Other	69 MB	5
196	<b>(MBTI) Myers-Briggs Personality Type Dataset</b>	Includes a large number of people's MBTI type and content written by them	Mitchell J	updated 4 months ago	personality, demographics, linguistics, + 2 more...	CSV, CC0	60 MB	11
224	<b>Zillow Economics Data</b>	Turning on the lights in housing research.	Zillow	updated 7 days ago	housing, business, demographics, economics	CSV, Other	511 MB	22

# Open data

- Geographical data
- City data
- Political data
- Research data
- Competition datasets
- Transactional data



The screenshot shows the homepage of the Bitcoin Block Explorer. At the top, there's a search bar with the placeholder "Search for block, transaction or address" and a status message "✓ - Conn 88 - Height 507071". Below the search bar are navigation links for "Block Explorer", "News", "Market", "Bitcoin cash", "Zcash", "Blocks", and "Status". A pink button on the right says "Buy Bitcoin with CCI".

**Latest Blocks**

Height	Age	Transactions	Mined by	Size
507071	12 minutes ago	1957	SlushPool	957089
507070	28 minutes ago	1457	BTCC Pool	977005
507069	38 minutes ago	978		936494
507068	43 minutes ago	2359		948954
507067	an hour ago	906		974116

[See all blocks](#)

**Latest Transactions**

Hash	Value Out
4643fb28ed5563d15d1ded22792a6bbebc195dc1...	0.01992232 BTC
a4abc4c7907f1eabb4fccd67c9c439750860d21d53b4...	0.10424438 BTC
512cdac30dcaf02ed120b1998251f9d0a0b56dbbcd...	0.05263963 BTC
28b108c6b6c2919118472371d58563b12ad66933de...	0.11340727 BTC
875ab3e7db9214516d6136976a4d108372c95a3d5e...	88.41989751 BTC
8228bd3e8261110b800528f1b762c8035e8ce595cd...	0.09296208 BTC
3a83f69f11c71b5c39c1608349cfa72ff9bf4d51fb652...	0.22035048 BTC
57b348c4874defb2713ee8815dfbdbbbb57a94b2f7...	3.98100412 BTC
c7ad14d5b9b7e772c57062033951a763ea0ad0c4854...	0.16676531 BTC
09f848a9ad7a2a16338795ac30e0f97ecbe7062c6db...	38.628917 BTC

**News**

-  South Korea Has No Plan to Ban Cryptocurrency Trading, Says Finance Minister
-  Hong Kong Authorities Won't Ban Cryptocurrencies, but Will Campaign to Educate Investors on Risks Instead
-  Tarsnap to End Bitcoin Payment Support on April 1st

**About Block Explorer**

Bitcoin Block Explorer is an open source web tool that allows you to view information about [blocks](#), [addresses](#), and [transactions](#) on the Bitcoin blockchain. The [source code](#) is on GitHub.

[What is bitcoin?](#)

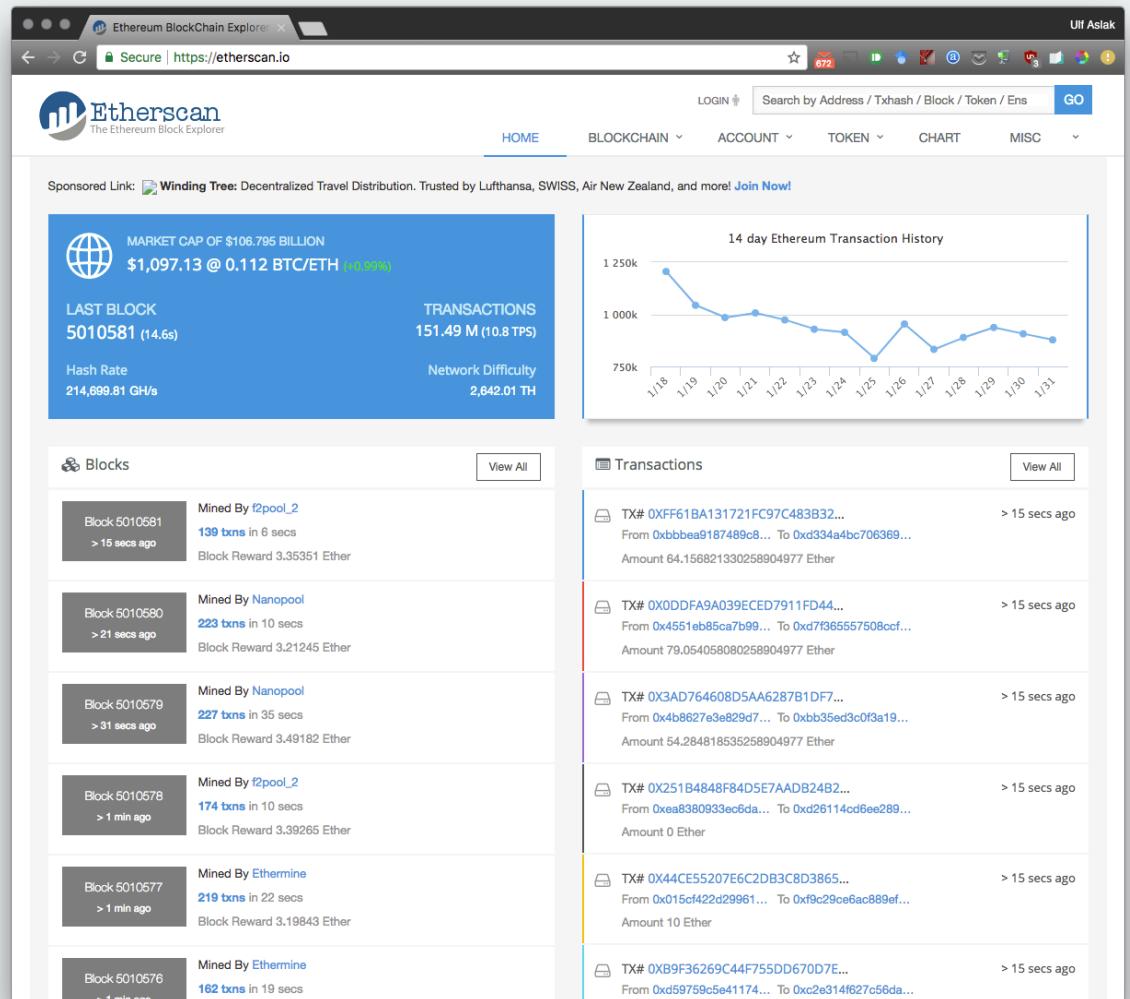
[Public Bitcoin API:](#) Machine readable stats & blockchain info can be accessed directly through the [REST](#) and [Websockets APIs](#).

[Testnet](#) is Bitcoin's sandbox. Block Explorer supports viewing both the [testnet](#) and [mainnet](#) blockchains.

Thanks to [Private Internet Access](#) for hosting the site. They provide a [VPN Service](#) that accepts Bitcoin.

# Open data

- Geographical data
- City data
- Political data
- Research data
- Competition datasets
- Transactional data



Overview ● ●

Open data ●

APIs ●

Scraping ○

# APIs

# APIs

## What is an API?

# APIs

The screenshot shows the Wikipedia article for "Batman". The main content area describes the character's origin, powers, and popularity. To the right, there is a large image of Batman in action, with the caption "Art by Tony Daniel". Below the image is a "Publication information" section listing the publisher (DC Comics), first appearance (Detective Comics #27), and creators (Bob Kane, Bill Finger). Further down is an "In-story information" section detailing alter egos, team affiliations, and partnerships.

Publication information	
Publisher	DC Comics
First appearance	Detective Comics #27 (cover date May 1939 / release date March 1939)
Created by	Bob Kane Bill Finger <sup>[1]</sup>

In-story information	
Alter ego	Bruce Wayne
Team affiliations	Batman Family Justice League Outsiders Batmen of All Nations Batman Incorporated
Partnerships	Robin (various) James Gordon Superman Wonder Woman Batgirl (various)

# APIs

```
1 import requests as rq
2
3 query = "https://en.wikipedia.org/w/api.php?format=json&action=query&titles=Batman&prop=revisions&rvprop=content"
4 response = rq.get(query)
5 result = response.json()
```

# APIs

```
1 import requests as rq
2
3 query = "https://en.wikipedia.org/w/api.php?format=json&action=query&titles=Batman&prop=revisions&rvprop=content"
4 response = rq.get(query)
5 result = response.json()
```

The diagram shows a code snippet for making a request to the Wikipedia API. The URL is defined on line 3: `query = "https://en.wikipedia.org/w/api.php?format=json&action=query&titles=Batman&prop=revisions&rvprop=content"`. Two arrows point from labels to specific parts of this URL:

- An arrow from the blue box labeled "API address" points to the prefix `https://en.wikipedia.org/w/api.php?`.
- An arrow from the red box labeled "Query parameters" points to the query parameters `format=json&action=query&titles=Batman&prop=revisions&rvprop=content`.

API address      Query parameters

# APIs

```
1 import requests as rq
2
3 query = "https://en.wikipedia.org/w/api.php?format=json&action=query&titles=Batman&prop=revisions&rvprop=content"
4 response = rq.get(query)
5 result = response.json()
```

# APIs

```
1 import requests as rq
2
3 query = "https://en.wikipedia.org/w/api.php?format=json&action=query&titles=Batman&prop=revisions&rvprop=content"
4 response = rq.get(query)
5 result = response.json()
```

# APIs

```
1 import requests as rq
2
3 query = "https://en.wikipedia.org/w/api.php?format=json&action=query&titles=Batman&prop=revisions&rvprop=content"
4 response = rq.get(query)
5 result = response.json()
```

# APIs

```
1 import requests as rq
2
3 query = "https://en.wikipedia.org/w/api.php?format=json&action=query&titles=Batman&prop=revisions&rvprop=content"
4 response = rq.get(query)
5 result = response.json()
```

```
1 import json  
2  
3 print json.dumps(result, indent=4)
```

Last executed 2018-02-01 10:28:27 in 10ms

```
{  
  "batchcomplete": "",  
  "query": {  
    "pages": {  
      "4335": {  
        "ns": 0,  
        "pageid": 4335,  
        "revisions": [  
          {  
            "*": "{{About|the fictional character}}\n{{pp-semi-indef}}\n{{pp-move-indef}}\n{{Use mdy da  
tes|date=October 2015}}\n{{Infobox comics character\n|image = Batman Detective Comics Vol 2 1.png<!-- Do NO  
T change this image without consensus from the Talk Page-->\n|imagesize =\n|converted = y\n|caption  
= Art by [[Tony Daniel]]\n|alt = Batman descends upon Gotham City\n|publisher = [[DC Comics]]\n|debut  
= ''[[Detective Comics]]'' #27<br />(cover date May 1939 /<br>release date March 1939)<!-- \"Debut\"  
indicates the first appearance of a character, not a change to the character's backstory. -->\n|creators =  
{{plainlist|\n*[[Bob Kane]]\n*[[Bill Finger]][[Ron Goulart|Goulart, Ron]], 'Comic Book Encyclopedia' ([[HarperCollins|Harper Entertainment]], New York, 2004) {{ISBN|978-0-06-053816-3}}</ref>\n}}\n|alter_ego = Bruce Wayne<!-- Do not enter a middle name. He has been depicted with too many different middle names to enter a specific one here. Also, there is no past or current, dead, or alive in fiction from a real world perspective; the infobox should cover the Batman known to the public consciousness and not a current comic book storyline. -->\n|alliances = {{plainlist|\n*[[List of Batman supporting characters|Batman Family|Batman Family]]\n*[[Justice League]]\n*[[Outs
```

# (web) Scraping

# Scraping

The screenshot shows the Rotten Tomatoes homepage. At the top, there's a navigation bar with links for MOVIES & DVDs, TV, NEWS, and TICKETS & SHOWTIMES. Below the navigation is a banner featuring several movie posters and headlines: "35 Haunted House Movies Ranked Best to Worst by Tomatometer", "Black Lightning Debuts as Winter's Best-Reviewed Show (So Far)", and "Hostiles Is the 61st Best-Reviewed Western Movie".

On the left side, there are three main sections: "MOVIES OPENING THIS WEEK", "TOP BOX OFFICE", and "COMING SOON TO THEATERS".

- MOVIES OPENING THIS WEEK:**

Rating	Title	Release Date
No Score Yet	Winchester	FEB 2
87%	A Fantastic Woman (Una muje...)	FEB 2
60%	Bilal: A New Breed of Hero	FEB 2
72%	Before We Vanish (Sanpo suru ...)	FEB 2
50%	Braven	FEB 2

[View All >](#)
- TOP BOX OFFICE:**

Rating	Title	Gross
45%	Maze Runner: The Death Cure	\$24.2M
76%	Jumanji: Welcome to the Jungle	\$16.2M
72%	Hostiles	\$10.2M
54%	The Greatest Showman	\$9.7M
88%	The Post	\$9.2M
54%	12 Strong	\$8.8M
41%	Den of Thieves	\$8.7M
92%	The Shape of Water	\$6M
100%	Paddington 2	\$5.8M
80%	Padmaavat	\$4.5M

[View All >](#)
- COMING SOON TO THEATERS:**

Rating	Title	Release Date
No Score Yet	Fifty Shades Freed	FEB 9

On the right side, there are two main sections: "NEW TV TONIGHT" and "MOST POPULAR TV ON RT".

- NEW TV TONIGHT:**

Rating	Title
29%	A.P. Bio
100%	The Good Place
100%	How to Get Away With Murder
100%	Supernatural
92%	Grey's Anatomy
91%	Scandal
89%	Arrow
88%	Will & Grace
75%	Young Sheldon
45%	S.W.A.T.
No Score Yet	Chicago Fire

[View All >](#)
- MOST POPULAR TV ON RT:**

Rating	Title
100%	Black Lightning
100%	Counterpart
97%	The End of the F***ing World
64%	Altered Carbon
85%	American Crime Story
No Score Yet	Babylon Berlin

[View All >](#)

At the bottom right, there are social media sharing icons for Facebook, Twitter, Pinterest, Google+, and LinkedIn.

# Scraping

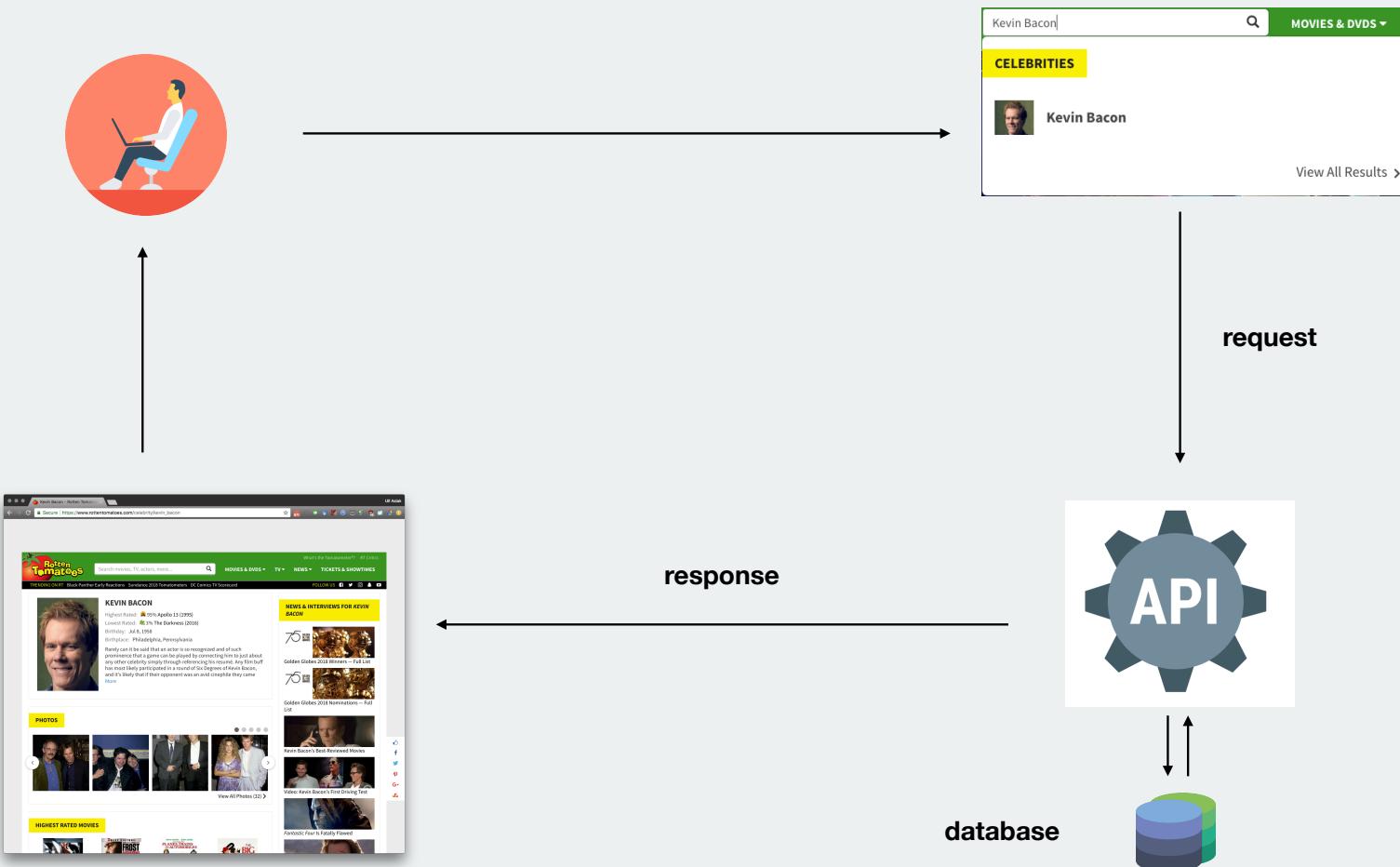
The screenshot shows the Rotten Tomatoes homepage. A search bar at the top contains the name "Kevin Bacon". Below the search bar, there's a "CELEBRITIES" section featuring a thumbnail of Kevin Bacon. A tooltip-like overlay appears over his name, containing the text "View All Results >". The main content area includes sections for "MOVIES OPENING THIS WEEK", "TOP BOX OFFICE", "COMING SOON TO THEATERS", "NEW TV TONIGHT", "MOST POPULAR TV ON RT", and "TOP DVD & STREAMING MOVIES". Each section lists various titles with their release dates or air dates and Rotten Tomatoes scores.

Section	Title	Score	Release/Air Date
MOVIES OPENING THIS WEEK	Winchester	No Score Yet	FEB 2
	A Fantastic Woman (Una muje...)	87%	FEB 2
	Bilal: A New Breed of Hero	60%	FEB 2
	Before We Vanish (Sanpo suru ...)	72%	FEB 2
	Braven	50%	FEB 2
View All >			
TOP BOX OFFICE	Maze Runner: The Death Cure	45%	\$24.2M
	Jumanji: Welcome to the Jungle	76%	\$16.2M
	Hostiles	72%	\$10.2M
	The Greatest Showman	54%	\$9.7M
	The Post	88%	\$9.2M
	12 Strong	54%	\$8.8M
	Den of Thieves	41%	\$8.7M
	The Shape of Water	92%	\$6M
	Paddington 2	100%	\$5.8M
	Padmaavat	80%	\$4.5M
View All >			
COMING SOON TO THEATERS	Fifty Shades Freed	No Score Yet	FEB 9
	View All >		
TOP DVD & STREAMING MOVIES			

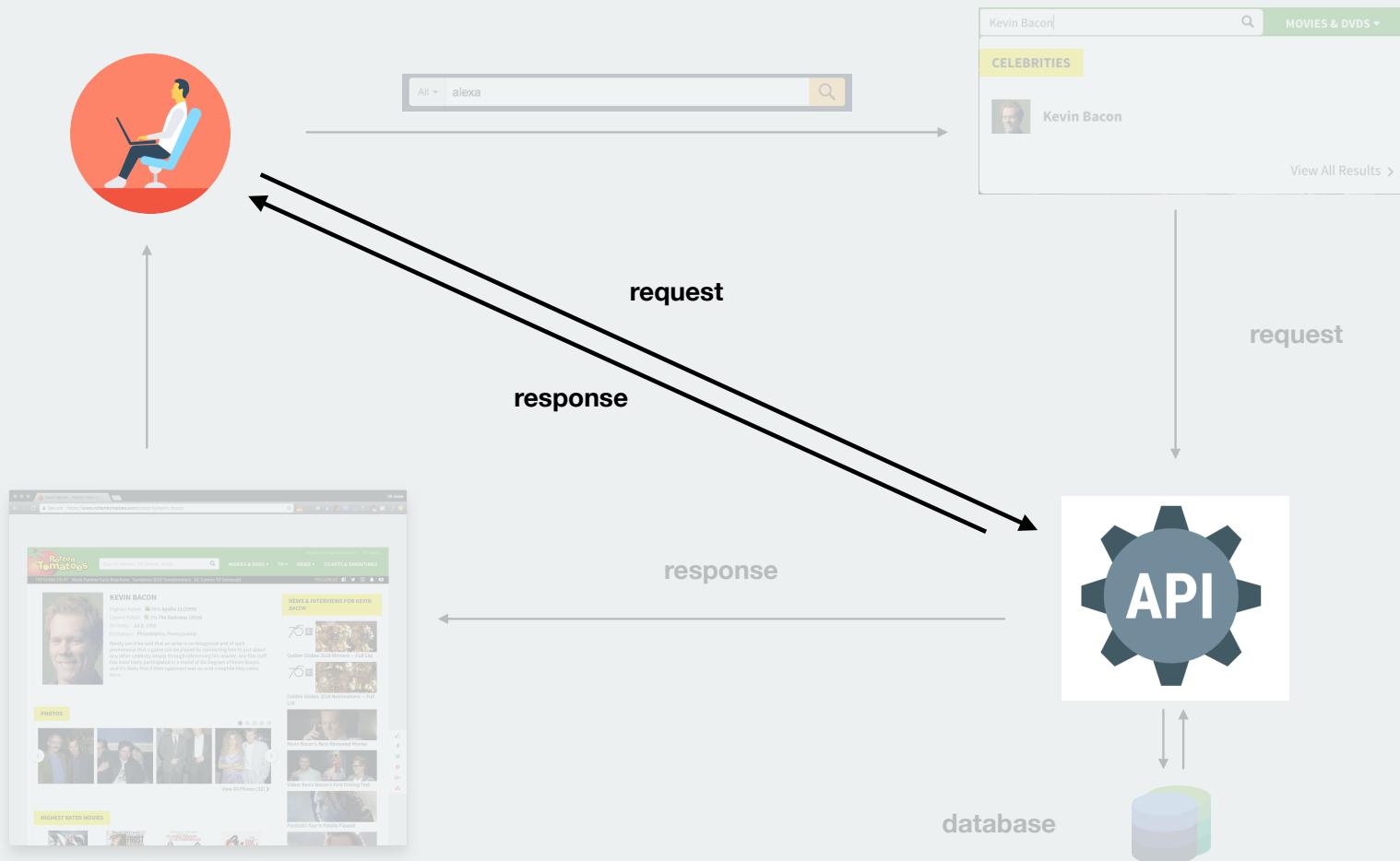
# Scraping

The screenshot shows the Rotten Tomatoes website for Kevin Bacon. At the top, there's a navigation bar with links for MOVIES & DVDs, TV, NEWS, and TICKETS & SHOWTIMES. Below the navigation is a search bar and a trending section. The main content area features a large photo of Kevin Bacon and his bio. It includes information about his highest and lowest rated movies, birthdate, and birthplace. A "More" link is present. Below this is a "PHOTOS" section with a grid of images and a "View All Photos (32)" link. To the right, there's a sidebar titled "NEWS & INTERVIEWS FOR KEVIN BACON" with links to Golden Globes 2018 winners and nominations, as well as a video of Kevin Bacon's first driving test. The bottom of the page shows a "HIGHEST RATED MOVIES" section with small movie posters.

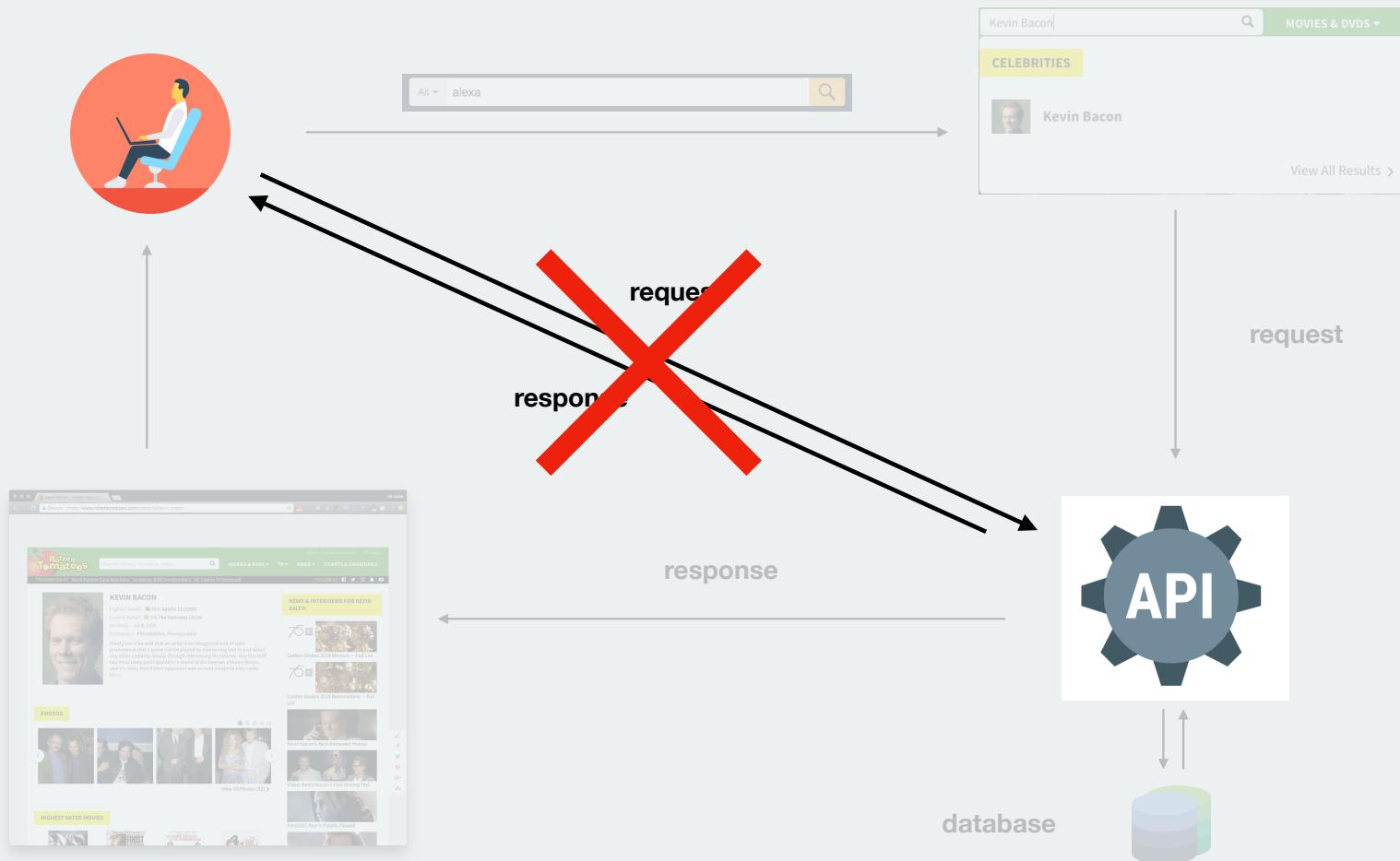
# Scraping



# Scraping



# Scraping



# Scraping

The screenshot shows the Rotten Tomatoes website for Kevin Bacon. At the top, there's a navigation bar with links for MOVIES & DVDs, TV, NEWS, and TICKETS & SHOWTIMES. Below the navigation is a search bar and a trending section. The main content area features a large photo of Kevin Bacon and his bio. It includes information about his highest and lowest rated movies, birthdate, and birthplace. A "More" link is present. Below this is a "PHOTOS" section with a grid of images and a "View All Photos (32)" link. To the right, there's a sidebar titled "NEWS & INTERVIEWS FOR KEVIN BACON" with links to Golden Globes 2018 winners and nominations, as well as a video of Kevin Bacon's first driving test. The bottom of the page shows a "HIGHEST RATED MOVIES" section with small movie posters.

# Scraping

The screenshot shows the Rotten Tomatoes website for Kevin Bacon. The main content includes his photo, basic stats (highest rated movie, lowest rated movie), and a bio mentioning his birthday and birthplace. Below this is a 'PHOTOS' section with a grid of images and a 'HIGHEST RATED MOVIES' section with movie thumbnails. A context menu is open over the page, listing options like Back, Forward, Reload, Save As..., Print..., Cast..., Translate to English, and several social sharing links for services like Reading List, Block element, Email page link..., Pushbullet, Save To Pocket, and goo.gl. The menu also includes View Page Source and Inspect.

# Scraping

Always check GitHub:  
probably someone has already written the  
scrapper you need

# Recap

- Get **open data** from public institutions, researchers or data sharing sites.
- Request it from someone's **API**. Is very easy, but usually has limits.
- “Forcefully” take it by **scraping** it from a website

# Get today's exercise from the GitHub repo!

Assignment 1 will come out at the end of this class  
I will send an announcement about it :)