

Computational Analysis of Big Data

Week 3

Getting data—scraping and APIs

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

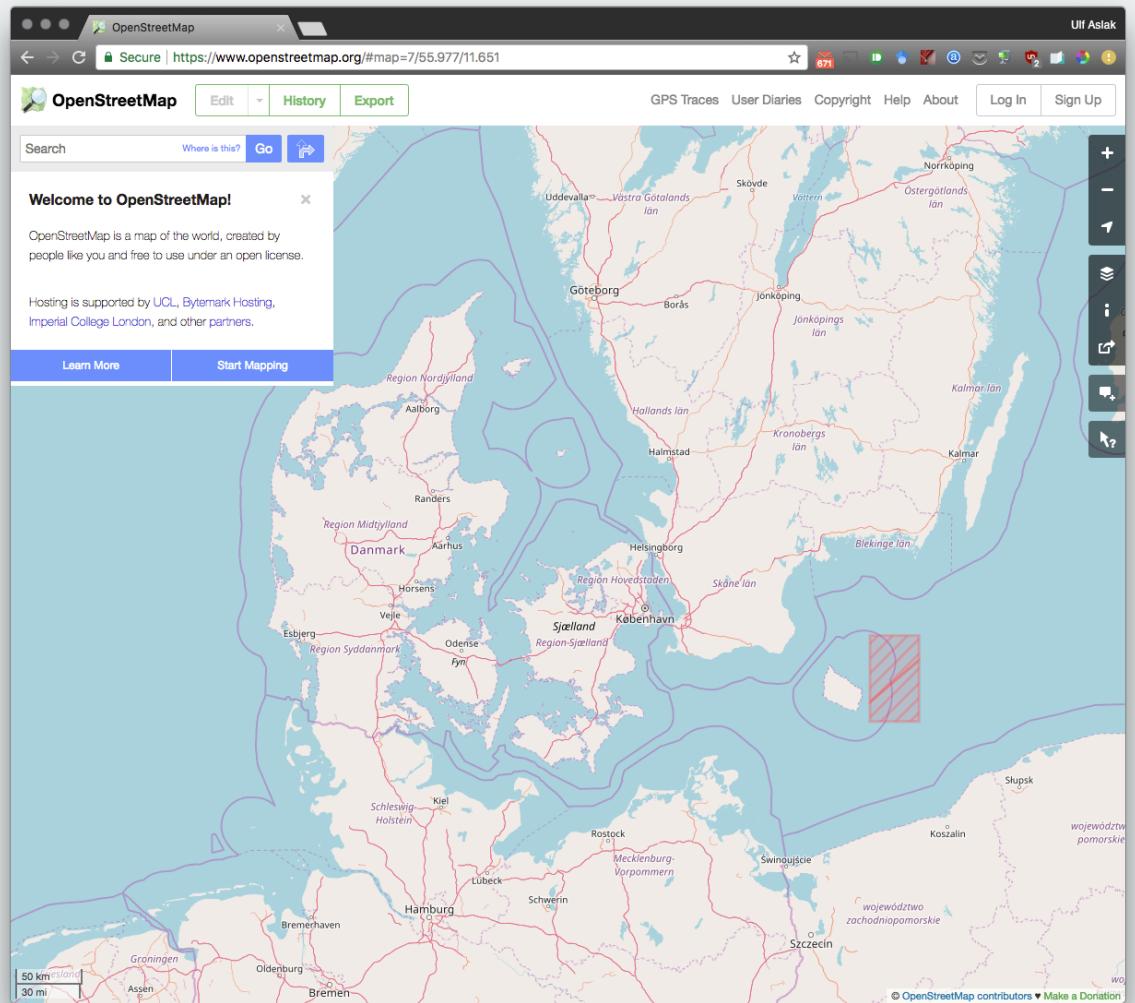
There are three main ways of getting existing data

- 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

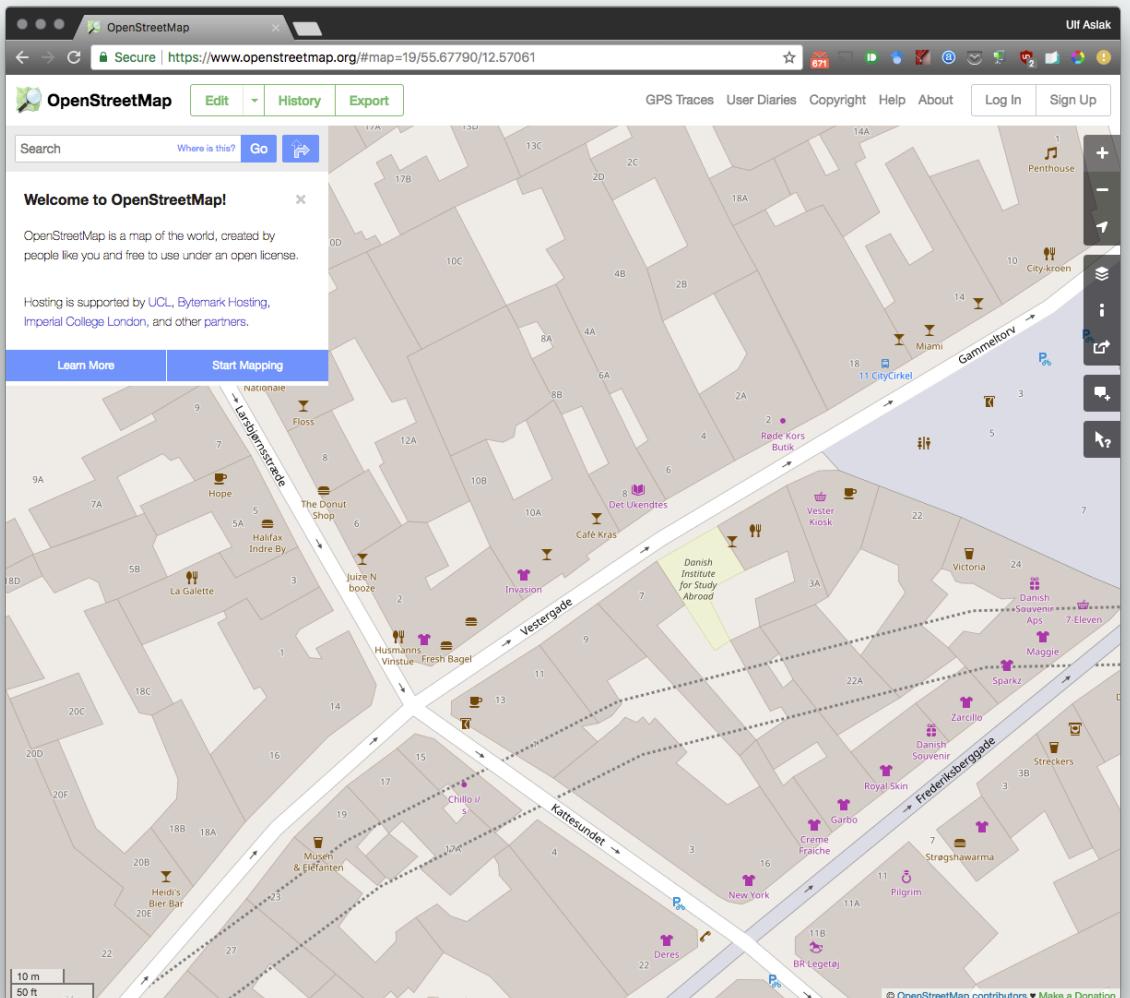
Open data

- **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

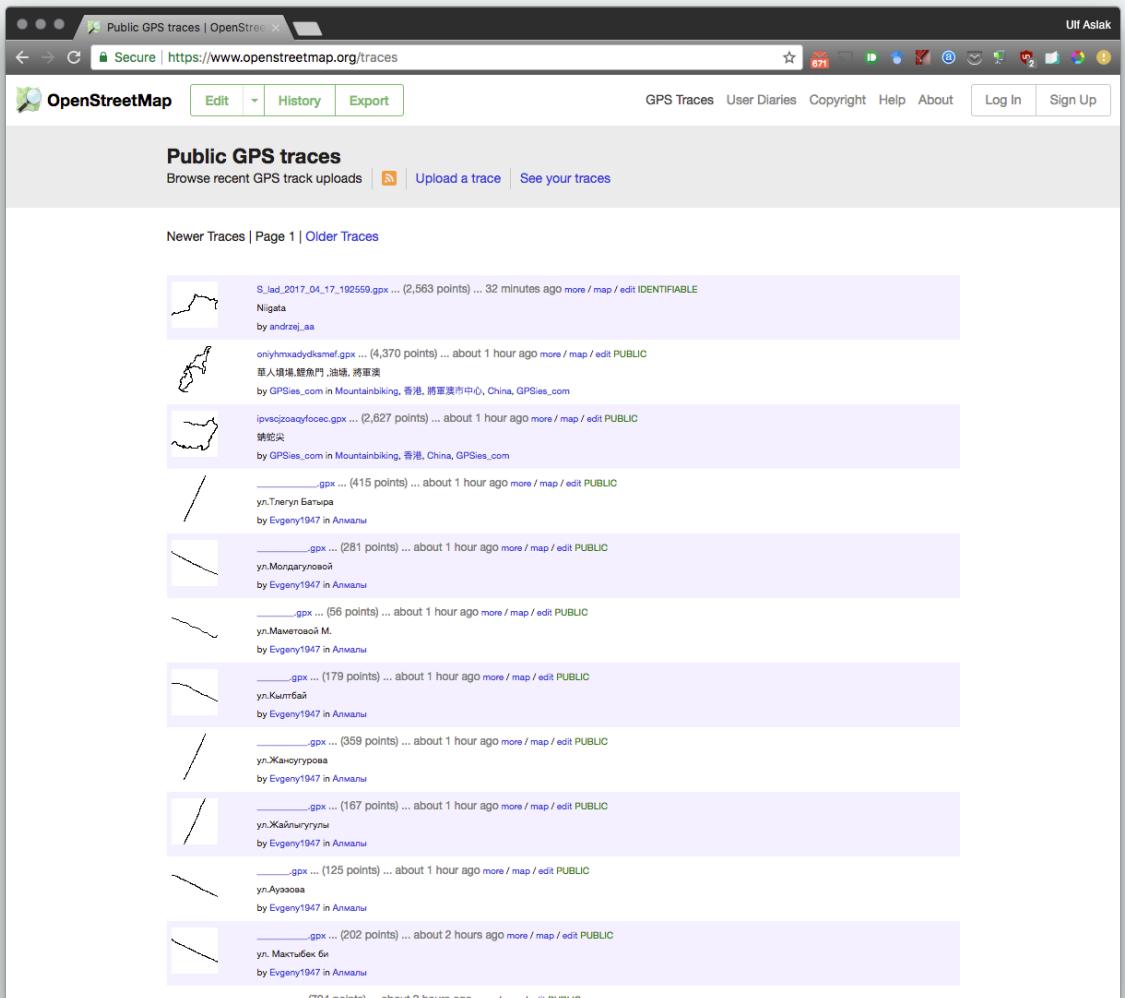
OpenStreetMap's community is diverse, passionate, and growing every day. Our contributors include enthusiast mappers, GIS professionals, engineers running the OSM servers, humanitarians mapping disaster-affected areas, and many more. To learn more about the community, see the [OpenStreetMap Blog](#),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

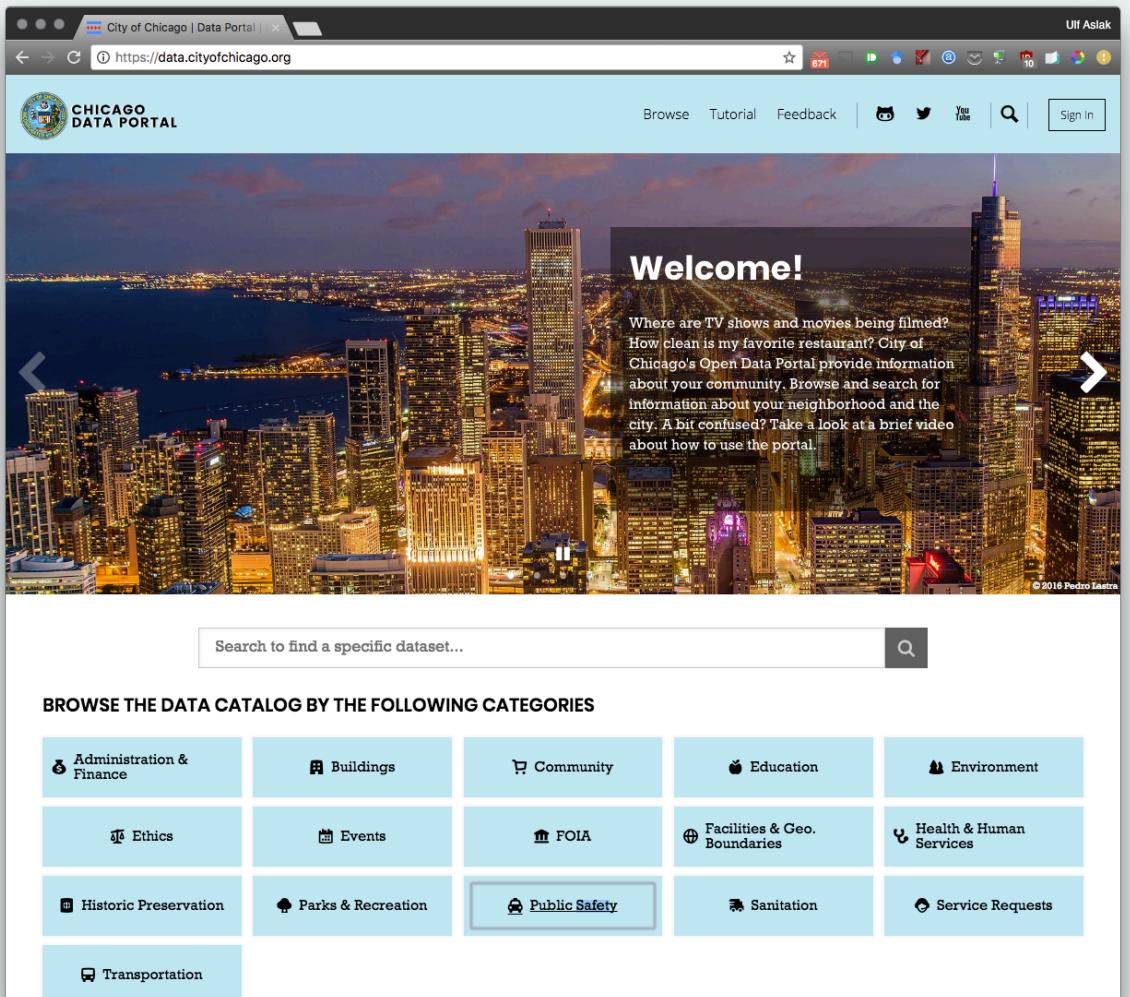
Open data

- **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 the Chicago Data Portal interface for the 'Crimes - 2001 to present' dataset. At the top, there's a navigation bar with links for 'Browse', 'Tutorial', 'Feedback', and social media icons. On the right, there's a 'Sign in' button.

The main content area has a title 'Crimes - 2001 to present' and a 'Public Safety' tag. Below the title is a detailed description of the dataset, mentioning it reflects reported incidents of crime from 2001 to present, minus the most recent seven days. It notes the data is extracted from the Chicago Police Department's CLEAR system. A 'More' link is present at the end of the description.

On the right side of the main content area, there's a sidebar with information about the last update (January 31, 2018) and the data provider (Chicago Police Department). There are also links for 'View Data', 'Visualize', 'Export', and 'API'.

Below the main description, there's a section titled 'Featured Content Using this Data' which lists three related datasets: 'Crimes - 2001 to present - Dashboard', 'Crimes - 2001 to present - Map', and 'Crimes - 2018'. Each featured item has a thumbnail, a date (January 31, 2018), and a view count.

At the bottom of the page, there's a section titled 'About this Dataset' with tabs for 'Updated' (set to January 31, 2018), 'Metadata', and 'Topics'. The 'Updated' tab shows details like 'Data Last Updated' (January 31, 2018), 'Metadata Last Updated' (September 27, 2017), and 'Date Created' (September 30, 2011). The 'Metadata' tab includes fields for 'Time Period' (2001 to present, minus the most recent seven days), 'Frequency' (Data are updated daily), and 'Data Owner' (Police). The 'Topics' tab lists 'Category' (Public Safety).

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's a navigation bar with links for 'Dataset' (which is highlighted), 'organizations', 'groups', and 'About'. There's also a search bar with the placeholder 'Søg datasæt...' and a magnifying glass icon.

The main content area has a heading '269 data sets found' and a sorting dropdown set to 'Relevance'. Below this, there are several sections:

- Air pollution**: A brief description stating data is measured with automatic instruments and updated every hour, listing substances like NO2, NOx, CO, O3, and PM10.
- Tables and benches**: A note indicating it's retired and replaced by the Terrain Equipment data set, mentioning the Danish Technological and Environmental Administration.
- Base map**: A basic map of the City of Copenhagen with download options for 'dwg', 'dGN', and 'ZIP' formats.

On the left side, there's a sidebar with a tree view of categories:

- organizations**:
 - The City of Copenhagen (128)
 - Health Care (45)
 - ITS (21)
 - Technology and Environment (16)
 - Employment and ... (12)
 - Statistics (8)
 - Environmental accounting (7)
 - Climate (7)
 - Area and Urban Renewal (4)
 - Culture and Leisure ... (4)
- Show more Organizations**
- groups**:
 - Geography (56)
 - Transport and infras ... (27)

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", "medicare". The results are filtered by "Congressional Record". There are 1-100 of 577,488 results, 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)

Other filters visible on the left include "Legislation", "Members", "Senate Communications", "House Communications", "Nominations", "Treaty Documents", "Committee Reports", "Members", "Check all", and "Limit Your Search".

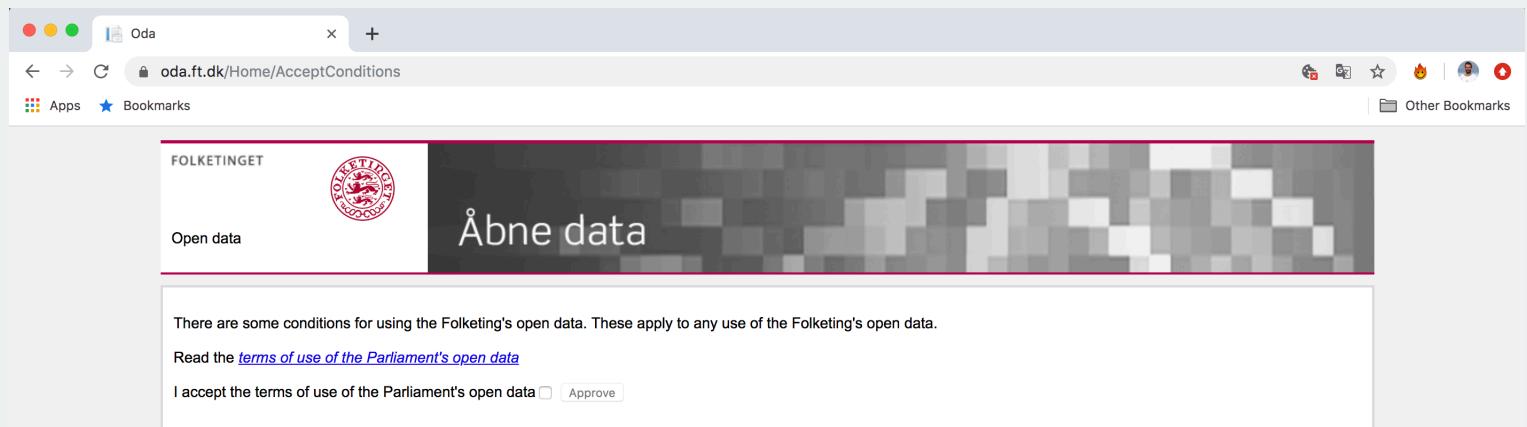
Open data

- 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 "About..." section with links to various congressional resources like Congress.gov, Committees, and Legislation. The main content area is titled "Bill Status Bulk Data" and explains that data is available from GPO's FDsys bulk data repository. It includes links for XML data imports and usage guidance. A "Related" sidebar on the right lists links for exploring bills, accounts, email alerts, and frequently asked questions. The footer contains site navigation links for Site Content, Help, Resources, House Links, and Senate Links.

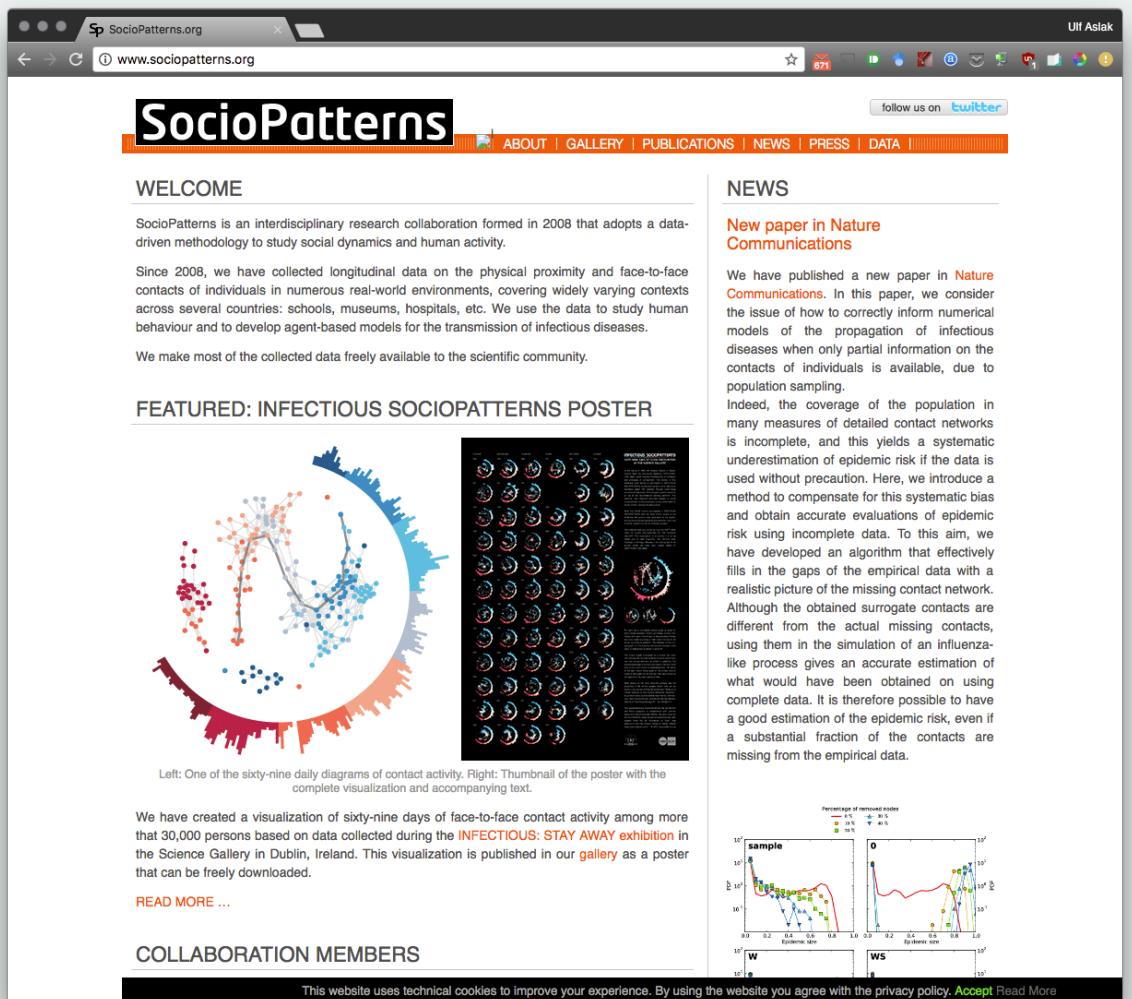
Open data

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



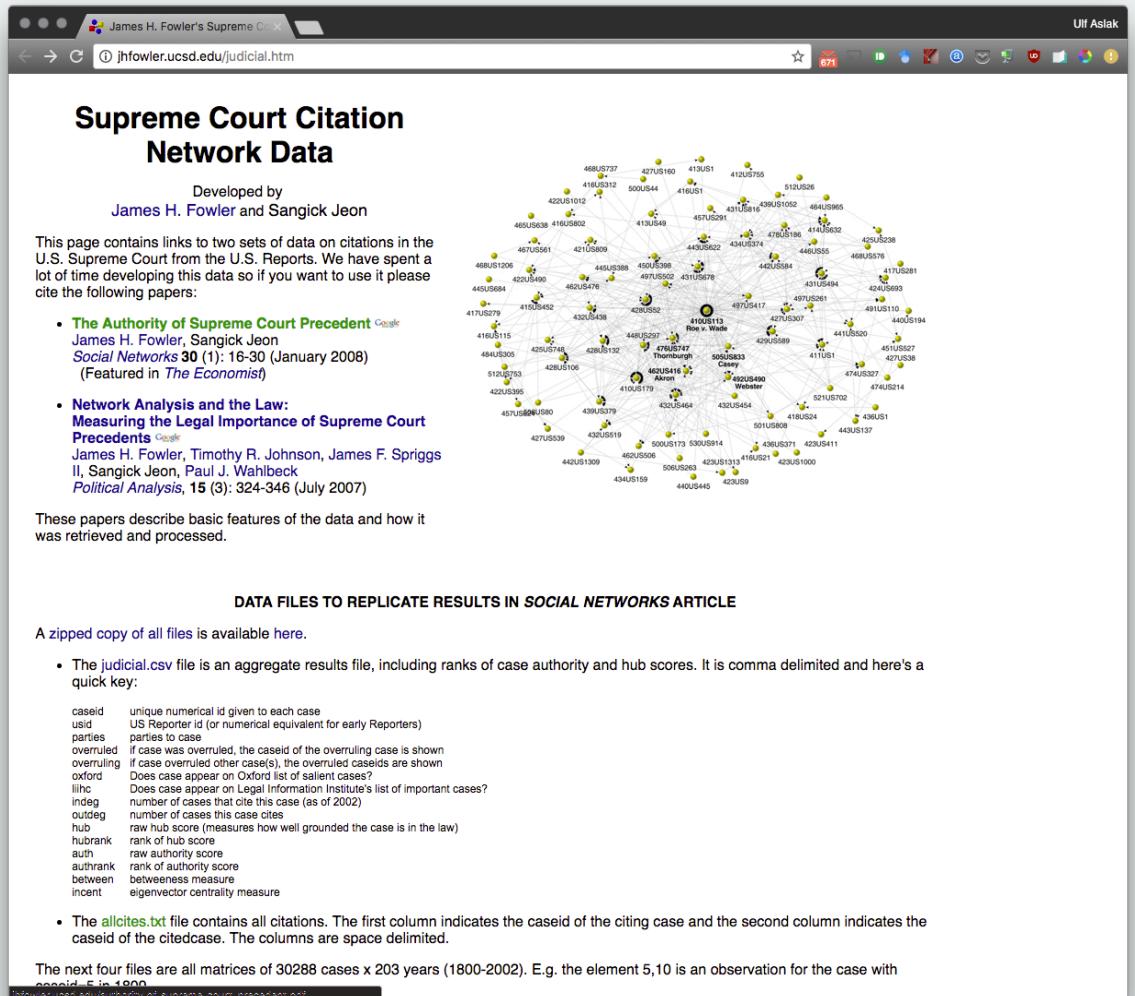
Open data

- 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

Description lawyers data Lazega

This data set comes from a network study of corporate law partnership that was carried out in a Northeastern US corporate law firm, referred to as SG&R, 1988-1991 in New England. It includes (among others) measurements of networks among the 71 attorneys (partners and associates) of this firm, i.e. their strong-coworker network, advice network, friendship network, and indirect control networks. Various members' attributes are also part of the dataset, including seniority, formal status, office in which they work, gender, lawschool attended, individual performance measurements (hours worked, fees brought in), attitudes concerning various management policy options, etc.

This dataset was used to identify social processes such as bounded solidarity, lateral control, quality control, knowledge sharing, balancing powers, regulation, etc. among peers.

The ethnography, organizational and network analyses of this case are available in Lazega (2001).

[Download the data set.](#)

Background:

- What do corporate lawyers do? Litigation and corporate work.
- Division of work and interdependencies.
- Three offices, no departments, built-in pressures to grow, intake and assignment rules.
- Partners and associates: hierarchy, up or out rule, billing targets.
- Partnership agreement (sharing benefits equally, 90% exclusion rule, governance structure, elusive committee system) and incompleteness of the contracts.
- Informal, unwritten rules (ex: no moonlighting, no investment in buildings, no nepotism, no borrowing to pay partners, etc.).
- Huge incentives to behave opportunistically ; thus the dataset is appropriate for the study of social processes that make cooperation among rival partners possible.

Sociometric name generators used to elicit coworkers, advice, and 'friendship' ties at SG&R:

"Here is the list of all the members of your Firm."

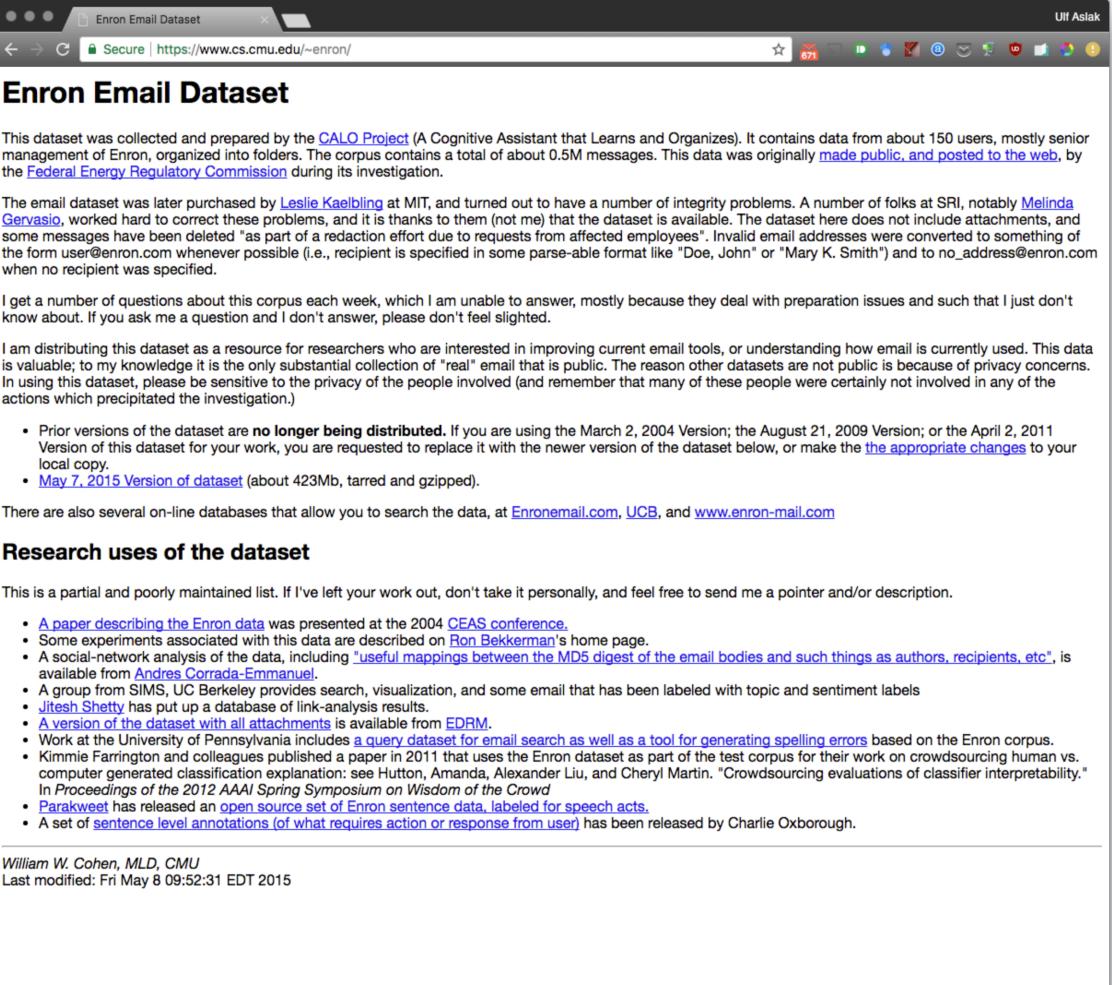
Strong coworkers network:
 "Because most firms like yours are also organized very informally, it is difficult to get a clear idea of how the members really work together. Think back over the past year, consider all the lawyers in your Firm. Would you go through this list and check the names of those with whom you have worked with. [By "worked with" I mean that you have spent time together on at least one case, that you have been assigned to the same case, that they read or used your work product or that you have read or used their work product; this includes professional work done within the Firm like Bar association work, administration, etc.]"

Basic advice network:
 "Think back over the past year, consider all the lawyers in your Firm. To whom did you go for basic professional advice? For instance, you want to make sure that you are handling a case right, making a proper decision, and you want to consult someone whose professional opinions are in general of great value to you. By advice I do not mean simply technical advice."

'Friendship' network:
 "Would you go through this list, and check the names of those you socialize with outside work. You know their family, they know yours, for instance. I do not mean all the people you are simply on a friendly level with, or people you happen to meet

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 URL is https://www.cs.cmu.edu/~enron/. The page title is "Enron Email Dataset". The content discusses the dataset's history, mentioning the CALO Project, the Federal Energy Regulatory Commission, Leslie Kaelbling, and Melinda Gervasio. It also notes the purchase by SRI and the work to correct integrity problems. The page includes a note from William W. Cohen, MLD, CMU, dated May 8, 2015. A section titled "Research uses of the dataset" lists various academic and research applications of the dataset.

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 [Enronmail.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 the Kaggle Datasets interface. At the top, there are tabs for 'Public', 'Your Datasets', and 'Favorites'. Below that, a search bar and filters for 'Sort by' (set to 'Hotness'), 'Sizes', 'File types', 'Licenses', and 'Tags' are visible. The main area displays a list of 10,348 datasets. Each dataset entry includes a thumbnail, a title, a brief description, the number of rows (e.g., 84, 15, 31, 28, 201, 241, 196, 224), the last updated time, and a detailed view icon. To the right of each dataset, there is a column of tags and file formats. For example, the 'Chocolate Bar Ratings' dataset is tagged with 'critical theory', 'food and drink', and has CSV and CC0 file formats. The 'Hacker News' dataset is tagged with 'journalism', 'information techn...', 'internet', and 'big query', with BigQuery and CC0 formats. The 'Historical Air Quality' dataset is tagged with 'pollution', 'BigQuery', and 'CC0', with BigQuery and CC0 formats. The 'GitHub Repos' dataset is tagged with 'programming lang...', 'programming', 'software engineer...', and 'BigQuery', with BigQuery and Other formats. The 'TED Talks' dataset is tagged with 'clothing', 'multiclass classifi...', 'object identification', and 'CSV', with CSV and CC4 formats. The 'Fashion MNIST' dataset is tagged with 'clothing', 'multiclass classifi...', 'object identification', and 'Other', with Other and CC4 formats. The '(MBTI) Myers-Briggs Personality Type Dataset' is tagged with 'personality', 'demographics', 'linguistics', and 'CSV', with CSV and CC0 formats. The 'Zillow Economics Data' dataset is tagged with 'housing', 'business', 'demographics', and 'economics', with CSV and Other formats.

Dataset Title	Description	Last Updated	Tags	File Formats	Rows
Chocolate Bar Ratings	Expert ratings of over 1,700 chocolate bars	Rachael Tatman updated 6 months ago	critical theory, food and drink	CSV, CC0	84
Hacker News	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, CC0	15
Historical Air Quality	Air Quality Data Collected at Outdoor Monitors Across the US	US Environmental Protection Agency updated 2 months ago	pollution	BigQuery, CC0	31
GitHub Repos	Code and comments from 2.8 million repos	Github updated 2 months ago	programming lang..., programming, software engineer...	BigQuery, Other	28
TED Talks	Data about TED Talks on the TED.com website until September 21st, 2017	Rounak Banik updated 4 months ago	CSV, CC4	201	
Fashion MNIST	An MNIST-like dataset of 70,000 28x28 labeled fashion images	Zalando Research updated 2 months ago	clothing, multiclass classifi..., object identification	Other, CC4	241
(MBTI) Myers-Briggs Personality Type Dataset	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	196
Zillow Economics Data	Turning on the lights in housing research.	Zillow updated 7 days ago	housing, business, demographics, economics	CSV, Other	224

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". Below it, a message says "✓ - Conn 88 · Height 507071". There are tabs for "Scan" and "BTC". 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
4643f2b28ed5563d15d1ded22792a6bbecb195dc1a...	0.01992232 BTC
a4abc4c7907f1eabb4fccd67c9c439750860d21d53b4...	0.10424438 BTC
512cdac30dcaf02ed120b1998251f9d0a0b56dbbcd...	0.05263963 BTC
28b108c6b6c2919118472371d58563b12ad66933de...	0.11340727 BTC
875ab3e7db9214516d6136976a4d108372c95a3d5e...	88.41989751 BTC
8228bdf3e8261110b800528f1b762c8035e8ce595cd...	0.09296208 BTC
3a83f69f11c71b5c39c1608349cfa72ff9bf4d51fb652...	0.22035048 BTC
57b348c4874defb2713ee8815dfbdbcb57a94b2f7...	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?](#)

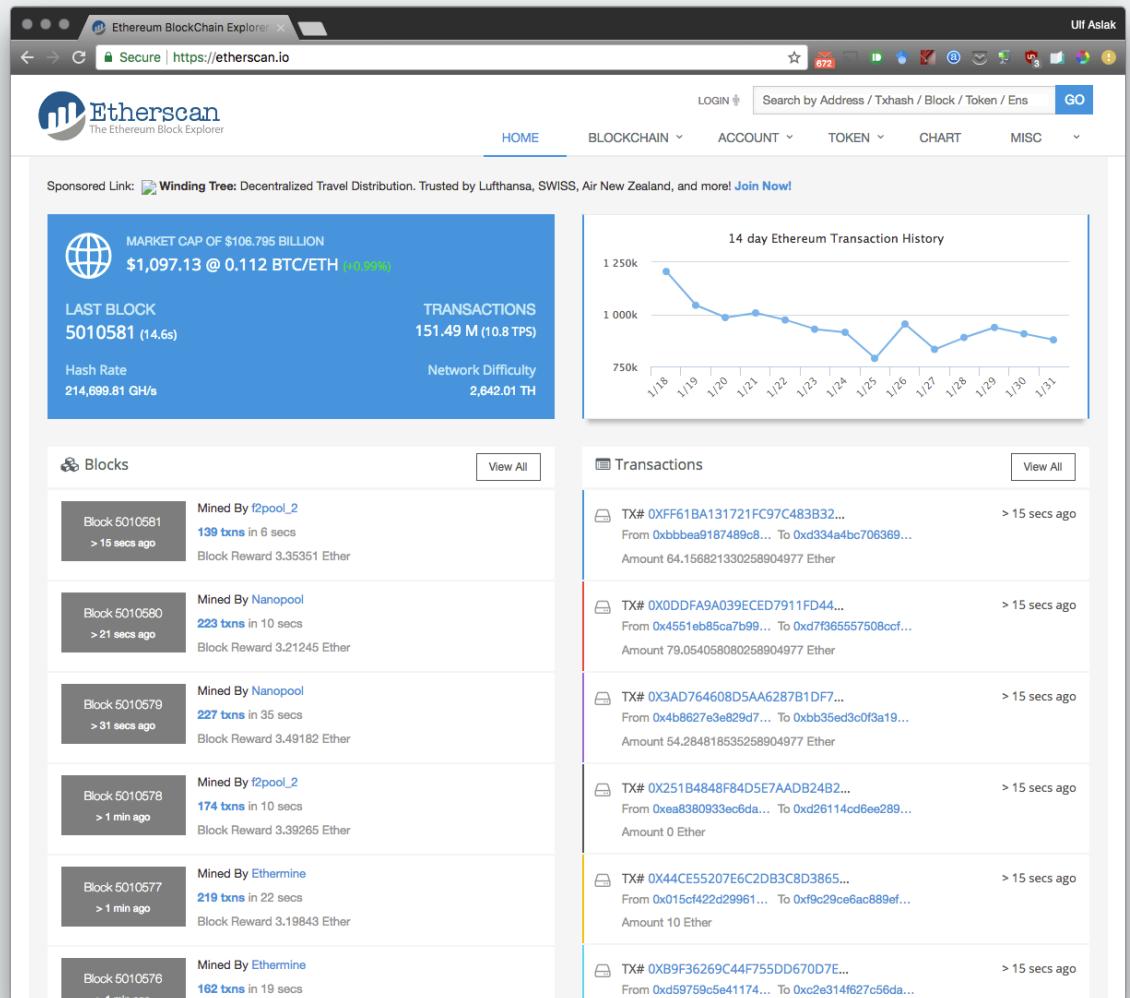
[Public Bitcoin API:](#) Machine readable stats & blockchain info can be accessed directly through the [REST](#) and [Websockets APIs](#).

[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



APIs

APIs

What is an API?

APIs

Batman - Wikipedia

Secure | <https://en.wikipedia.org/wiki/Batman>

Ulf Aslak

Article Talk Read View source View history Search Wikipedia

Batman

From Wikipedia, the free encyclopedia

This article is about the fictional character. For other uses, see [Batman \(disambiguation\)](#).

Batman is a fictional superhero appearing in American comic books published by DC Comics. The character was created by artist Bob Kane and writer Bill Finger,^{[4][5]} and first appeared in *Detective Comics* #27 (1939). Originally named the "Bat-Man", the character is also referred to by such epithets as the Caped Crusader, the Dark Knight, and the World's Greatest Detective.^[6]

Batman's secret identity is **Bruce Wayne**, a wealthy American playboy, philanthropist, and owner of **Wayne Enterprises**. After witnessing the murder of his parents Dr. Thomas Wayne and Martha Wayne as a child, he swore vengeance against criminals, an oath tempered by a sense of justice. Bruce Wayne trains himself physically and intellectually and crafts a **bat-inspired** persona to fight crime.^[7]

Batman operates in the fictional **Gotham City** with assistance from various supporting characters, including his butler **Alfred**, police commissioner **Gordon**, and vigilante allies such as **Robin**. Unlike most superheroes, Batman does not possess any **superpowers**; rather, he relies on his genius intellect, physical prowess, martial arts abilities, detective skills, science and technology, vast wealth, intimidation, and indomitable will. A large assortment of villains make up Batman's **rogues gallery**, including his archenemy, the **Joker**.

The character became popular soon after his introduction in 1939 and gained his own comic book title, **Batman**, the following year. As the decades went on, differing interpretations of the character emerged. The late 1960s **Batman** television series used a **camp aesthetic**, which continued to be associated with the character for years after the show ended. Various creators worked to return the character to his dark roots, culminating in 1986 with *The Dark Knight Returns* by Frank Miller. The success of Warner Bros.' live-action **Batman** feature films have helped maintain the character's prominence in mainstream culture.^[8]

An American cultural icon, Batman has garnered enormous popularity and is among the most identifiable comic book characters. Batman has been licensed and adapted into a variety of media, from radio to television and film, and appears on various merchandise sold around the world, such as toys and video games. The character has also intrigued psychiatrists, with many trying to understand his psyche. In 2015, *FanSided* ranked Batman as number one on their list of "50 Greatest Super Heroes In Comic Book History".^[9] Kevin Conroy, Bruce Greenwood, Peter Weller, Anthony Ruivivar, Jason O'Mara, and Will Arnett, among others, have provided the character's voice for animated adaptations. Batman has been depicted in both film and television by Lewis Wilson, Robert Lowery, Adam West, Michael Keaton, Val Kilmer, George Clooney, Christian Bale, and Ben Affleck.

Contents [hide]

- 1 Publication history
 - 1.1 Creation
 - 1.2 Golden Age
 - 1.3 Silver and Bronze Age
 - 1.4 Modern Age

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()
```

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 dates|date=October 2015}}\n{{Infobox comic character|image = Batman Detective Comics Vol 2 1.png<!--Do NOT change this image without consensus from the Talk Page-->|image_size = \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]]<ref>[[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]]\n*[[Justice League]]\n*[[Outsiders|Outsiders]]}}\n}}\n}}\n}}\n}}\n}}\n}}\n}}\n}}\n}}
```

(web) Scraping

Scraping

The screenshot shows the Rotten Tomatoes homepage with the following sections:

- TRENDING ON RT:** Black Panther Early Reactions, Sundance 2018 Tomatometer, DC Comics TV Scorecard.
- GALLERY:** 35 Haunted House Movies Ranked Best to Worst by Tomatometer.
- SCORECARD:** Black Lightning Debuts as Winter's Best-Reviewed Show (So Far).
- COUNTDOWN:** Hostiles Is the 61st Best-Reviewed Western Movie.
- 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
- 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 >](#)
- TOP DVD & STREAMING MOVIES:**

Scraping

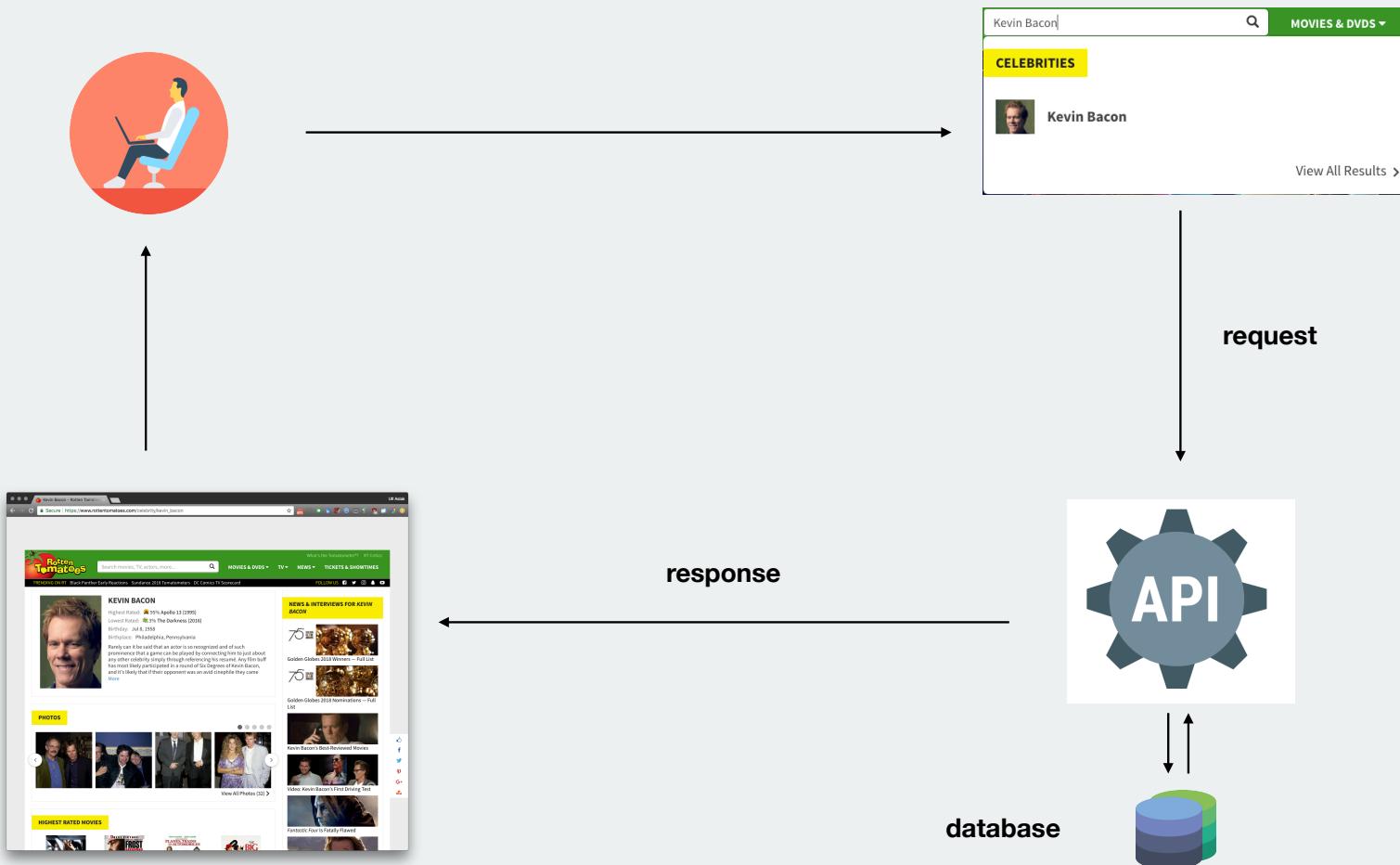
The screenshot shows the Rotten Tomatoes homepage. A search bar at the top contains the name "Kevin Bacon". Below the search bar, there's a "CELEBRITIES" section featuring a thumbnail of Kevin Bacon and a "View All Results >" link. The main content area includes sections for "MOVIES OPENING THIS WEEK", "TOP BOX OFFICE", "COMING SOON TO THEATERS", "NEW TV TONIGHT", "MOST POPULAR TV ON RT", and "TOP DVD & STREAMING MOVIES". Each section lists movie or TV show titles along with their Rotten Tomatoes scores and other details like release dates or box office earnings.

Section	Title	Score	Details
MOVIES OPENING THIS WEEK	Winchester	29%	A.P. Bio
	A Fantastic Woman (Una muje...)	100%	The Good Place
	Bilal: A New Breed of Hero	100%	How to Get Away With Murder
	Before We Vanish (Sanpo suru ...)	100%	Supernatural
	Braven	92%	Grey's Anatomy
TOP BOX OFFICE	Braven	91%	Scandal
	Maze Runner: The Death Cure	89%	Arrow
	Jumanji: Welcome to the Jungle	88%	Will & Grace
	Hostiles	75%	Young Sheldon
	The Greatest Showman	45%	S.W.A.T.
COMING SOON TO THEATERS	The Post	45%	Chicago Fire
	12 Strong	72%	View All >
	Den of Thieves	72%	
	The Shape of Water	72%	
	Paddington 2	72%	
TOP DVD & STREAMING MOVIES	Padmaavat	54%	View All >
	Fifty Shades Freed	54%	
	No Score Yet	41%	
	No Score Yet	100%	
	No Score Yet	100%	

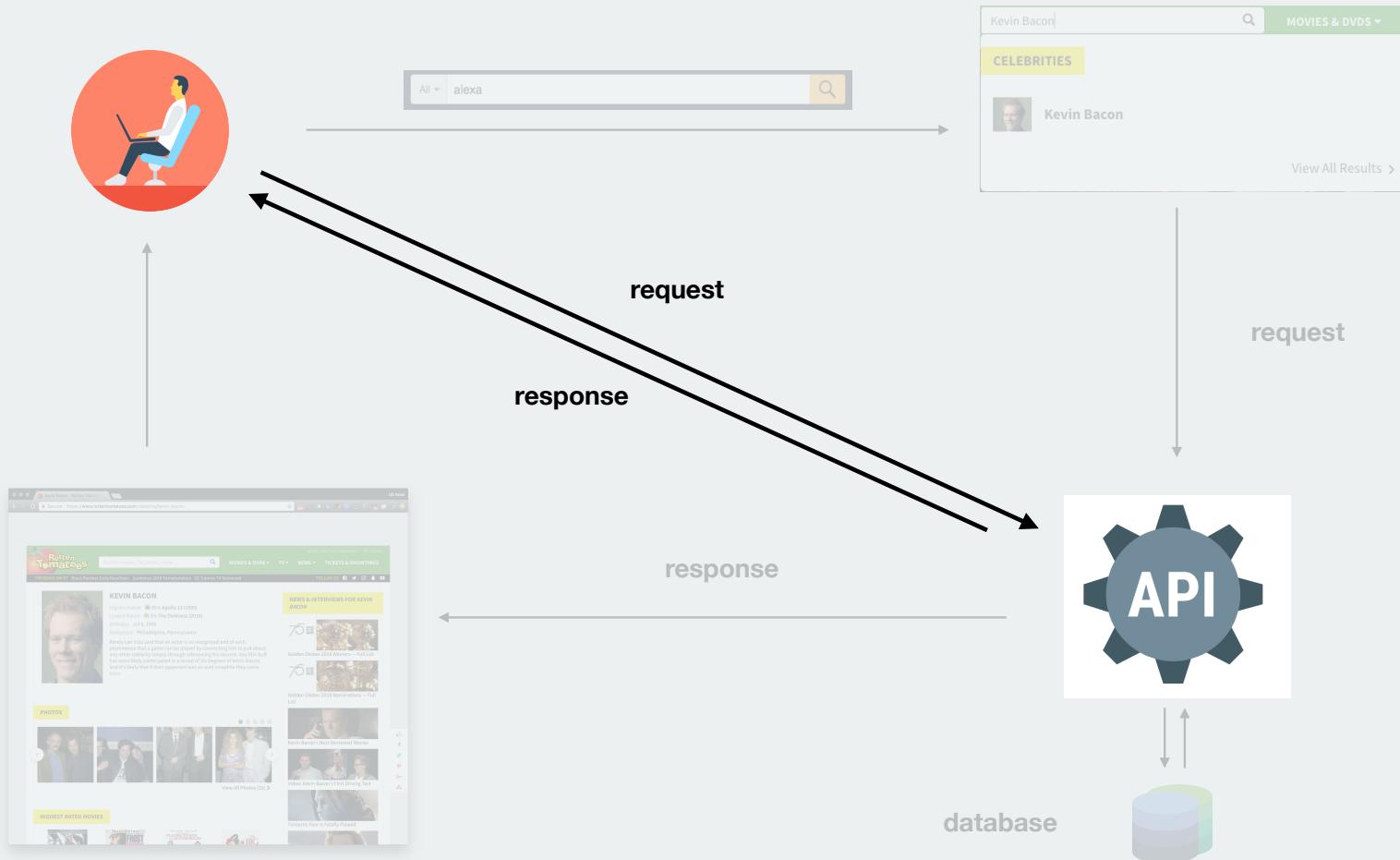
Scraping

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 for Black Panther Early Reactions. The main content area features a large photo of Kevin Bacon and his bio. His highest rated movie is listed as 95% for Apollo 13 (1995). Below his bio 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 him driving.

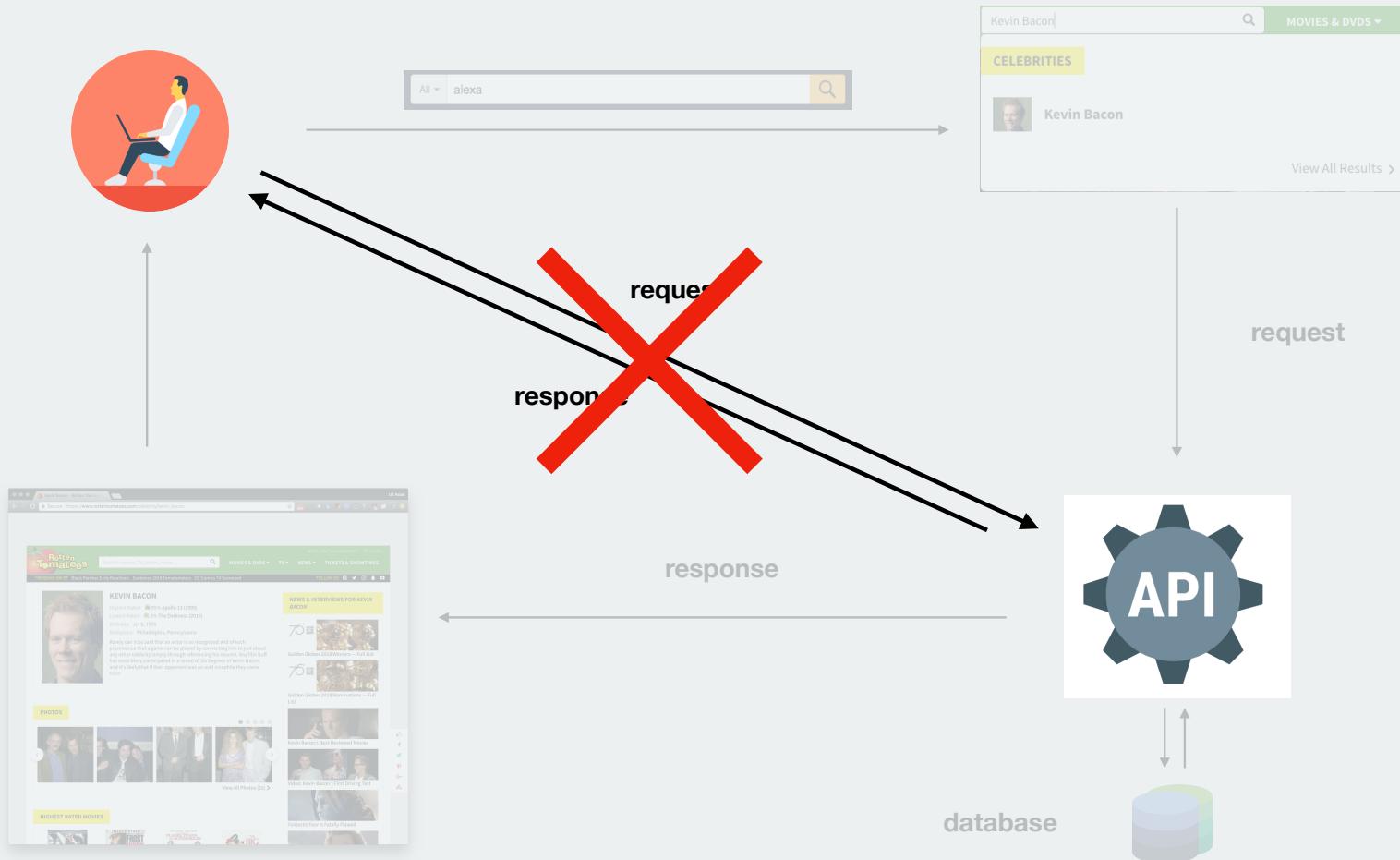
Scraping



Scraping



Scraping



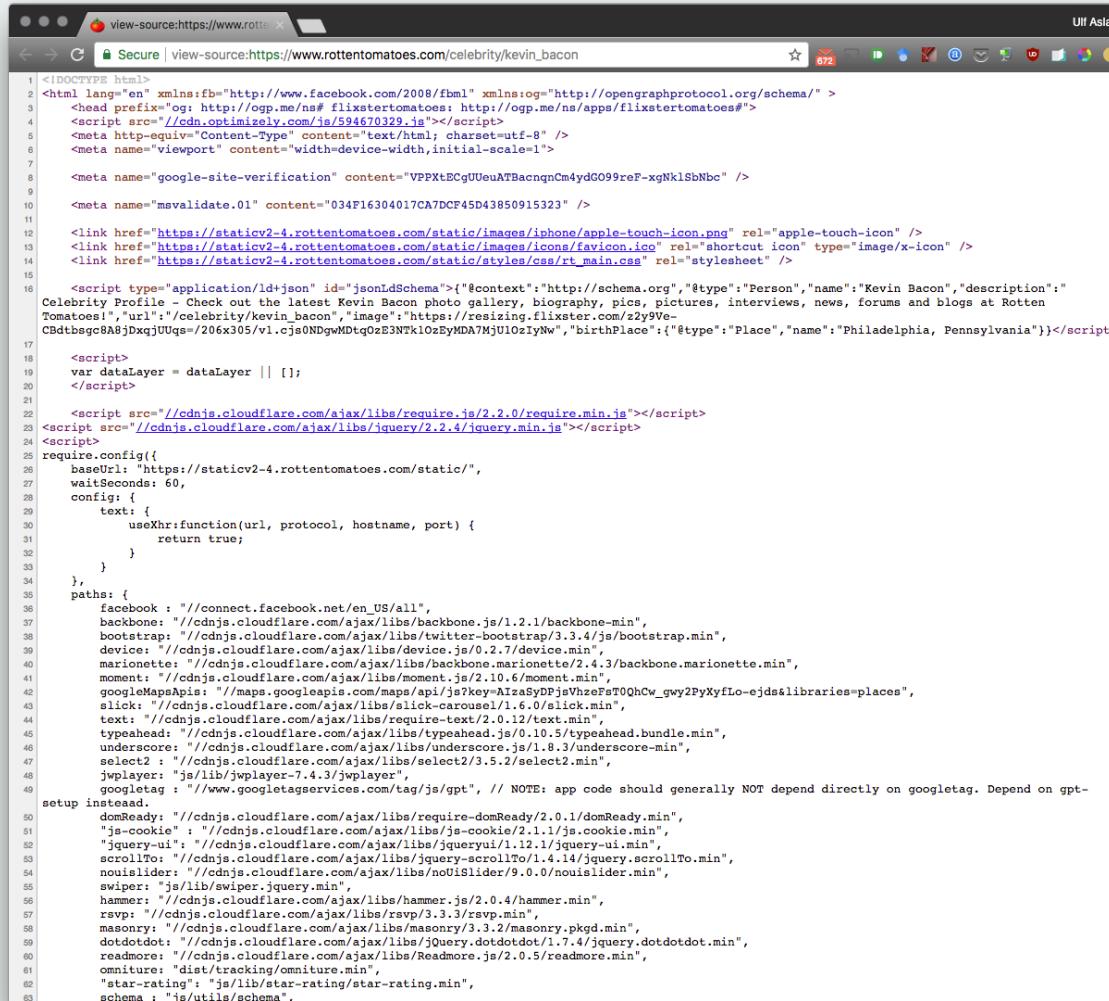
Scraping

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 for Black Panther Early Reactions. The main content area features a large photo of Kevin Bacon and his bio information. To the right, there's a sidebar with news and interviews about him, including links to the Golden Globes 2018 winners and nominations. Below the bio, there's a "PHOTOS" section with a grid of smaller images and a "HIGHEST RATED MOVIES" section with movie thumbnails.

Scraping

The screenshot shows the Rotten Tomatoes website for Kevin Bacon. The main content includes a large photo of Kevin Bacon, his biography (Highest Rated: 95% for Apollo 13, Lowest Rated: 3% for The Darkness), and a section on his highest-rated movies. To the right, there's a sidebar for news and interviews. A context menu is open over the page, listing options such as Back, Forward, Print, Save As..., and various sharing links for social media and services like Pushbullet and Pocket.

Scraping



```

1 <!DOCTYPE html>
2 <html lang="en" xmlns:fb="http://www.facebook.com/2008/fbml" xmlns:og="http://opengraphprotocol.org/schema/">
3   <head prefix="og: http://ogp.me/ns# flixstertomatoes: http://ogp.me/ns/apps/flixstertomatoes#">
4     <script src="//cdn.optimizely.com/js/594670329.js"></script>
5     <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
6     <meta name="viewport" content="width=device-width,initial-scale=1">
7
8     <meta name="google-site-verification" content="VPPXtECgJUeuATBacnqnCm4ydG099reF-xgNklSbNbc" />
9
10    <meta name="nsvalidate.01" content="034f16304017CA7DCF45DA43850915323" />
11
12    <link href="https://staticcv2-4.rottentomatoes.com/static/images/iphone/apple-touch-icon.png" rel="apple-touch-icon" />
13    <link href="https://staticcv2-4.rottentomatoes.com/static/images/icons/favicon.ico" rel="shortcut icon" type="image/x-icon" />
14    <link href="https://staticcv2-4.rottentomatoes.com/static/styles/css/rt_main.css" rel="stylesheet" />
15
16    <script type="application/ld+json" id="jsonDataSchema">{"@context": "http://schema.org", "@type": "Person", "name": "Kevin Bacon", "description": "Celebrity Profile - Check out the latest Kevin Bacon photo gallery, biography, pics, pictures, interviews, news, forums and blogs at Rotten Tomatoes!", "url": "celebrity/kevin_bacon", "image": "https://resizing.flixster.com/zzy9Ve-CBdtdbsgc8A8jDxqjUUqs-/206x305/v1.cjs0NDgwmDtqOzE3NTk1OzEyMDA7MjU0zIyNw", "birthPlace": {"@type": "Place", "name": "Philadelphia, Pennsylvania"}</script>
17
18    <script>
19      var dataLayer = dataLayer || [];
20    </script>
21
22    <script src="//cdnjs.cloudflare.com/ajax/libs/require.js/2.2.0/require.min.js"></script>
23    <script src="//cdnjs.cloudflare.com/ajax/libs/jquery/2.2.4/jquery.min.js"></script>
24  <script>
25    require.config({
26      baseUrl: "https://staticcv2-4.rottentomatoes.com/static/",
27      waitSeconds: 60,
28      config: {
29        text: {
30          useXhr: function(url, protocol, hostname, port) {
31            return true;
32          }
33        },
34      },
35      paths: {
36        facebook: "//connect.facebook.net/en_US/all",
37        backbone: "//cdnjs.cloudflare.com/ajax/libs/backbone.js/1.2.1/backbone-min",
38        bootstrap: "//cdnjs.cloudflare.com/ajax/libs/twitter-bootstrap/3.3.4/js/bootstrap.min",
39        device: "//cdnjs.cloudflare.com/ajax/libs/device.js/0.2.7/device.min",
40        marionette: "//cdnjs.cloudflare.com/ajax/libs/backbone.marionette/2.4.3/backbone.marionette.min",
41        moment: "//cdnjs.cloudflare.com/ajax/libs/moment.js/2.10.6/moment.min",
42        googleMapsApis: "//maps.googleapis.com/maps/api/js?key=AIZasSyDPjvWhzeFaTQhCw_gwy2PyXyfLo-ejds&libraries=places",
43        slick: "//cdnjs.cloudflare.com/ajax/libs/slick-carousel/1.6.0/slick.min",
44        text: "//cdnjs.cloudflare.com/ajax/libs/require-text/2.0.12/text.min",
45        typeahead: "//cdnjs.cloudflare.com/ajax/libs/typeahead.js/0.10.5/typeahead.bundle.min",
46        underscore: "//cdnjs.cloudflare.com/ajax/libs/underscore.js/1.8.3/underscore-min",
47        select2: "//cdnjs.cloudflare.com/ajax/libs/select2/3.5.2/select2.min",
48        jwplayer: "js/lib/jwplayer-7.4.3/jwplayer",
49        googletag: "//www.googletagservices.com/tag/js/gpt", // NOTE: app code should generally NOT depend directly on googletag. Depend on gpt-setup instead.
50        domReady: "//cdnjs.cloudflare.com/ajax/libs/require-domReady/2.0.1/domReady.min",
51        "js-cookie": "//cdnjs.cloudflare.com/ajax/libs/js-cookie/2.1.1/js.cookie.min",
52        "jquery-ui": "//cdnjs.cloudflare.com/ajax/libs/jqueryui/1.12.1/jquery-ui.min",
53        scrollTo: "//cdnjs.cloudflare.com/ajax/libs/jquery-scrollTo/1.4.14/jquery.scrollTo.min",
54        nouislider: "//cdnjs.cloudflare.com/ajax/libs/nouislider/9.0.0/nouislider.min",
55        swiper: "js/lib/swiper.jquery.min",
56        hammer: "//cdnjs.cloudflare.com/ajax/libs/hammer.js/2.0.4/hammer.min",
57        rsvp: "//cdnjs.cloudflare.com/ajax/libs/rsvp/3.3.3/rsvp.min",
58        masonry: "//cdnjs.cloudflare.com/ajax/libs/masonry/3.3.2/masonry.pkgd.min",
59        dotdotdot: "//cdnjs.cloudflare.com/ajax/libs/jquery.dotdotdot/1.7.4/jquery.dotdotdot.min",
60        readmore: "//cdnjs.cloudflare.com/ajax/libs/Readmore.js/2.0.5/readmore.min",
61        omniture: "dist/tracking/omniture.min",
62        "star-rating": "js/lib/star-rating/star-rating.min",
63        schema : "js/utils/schema",
64      }
65    }
66  
```