

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*.

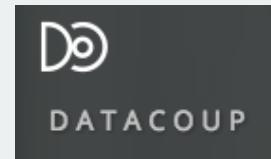
Why do we want *other people's* data?

- 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

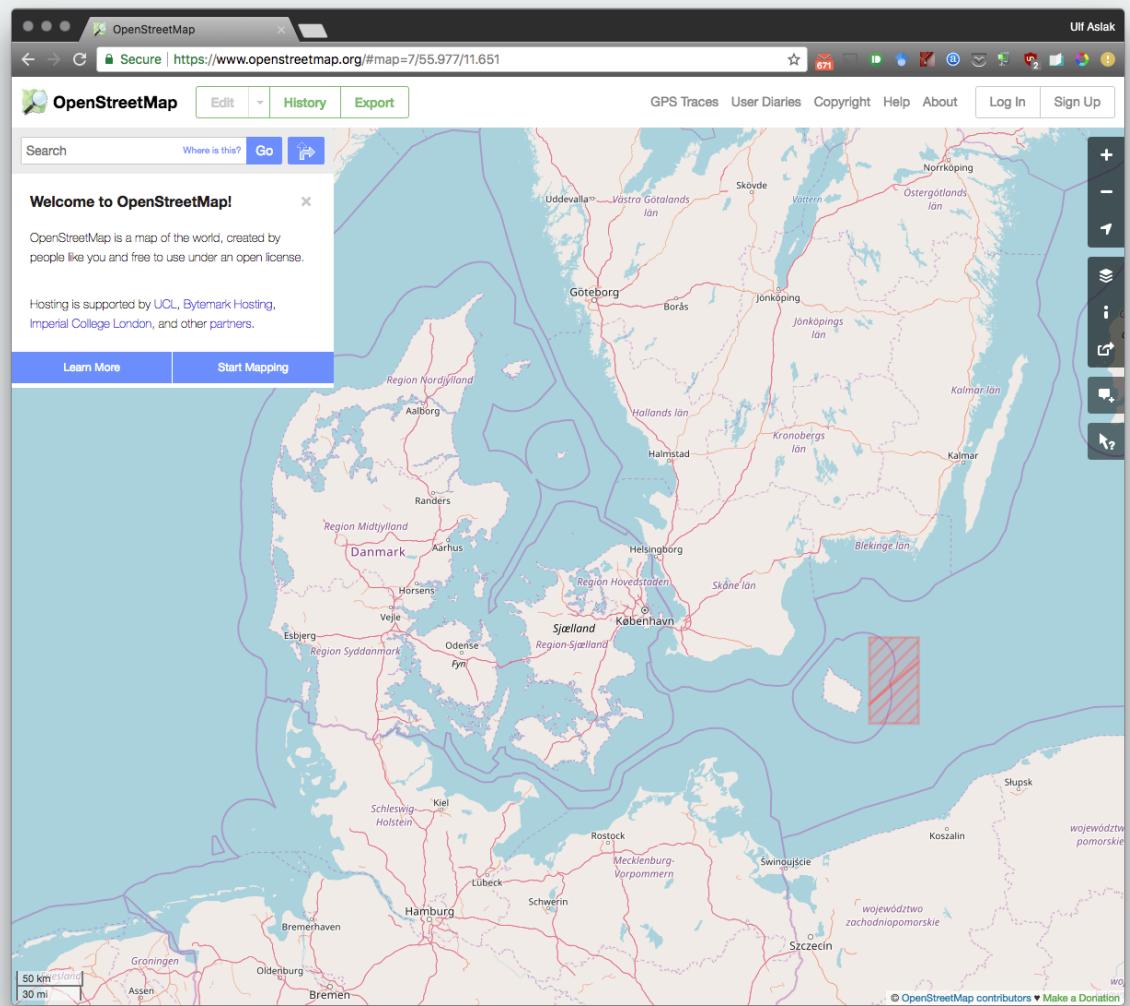
There are three main ways of getting existing data

- 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

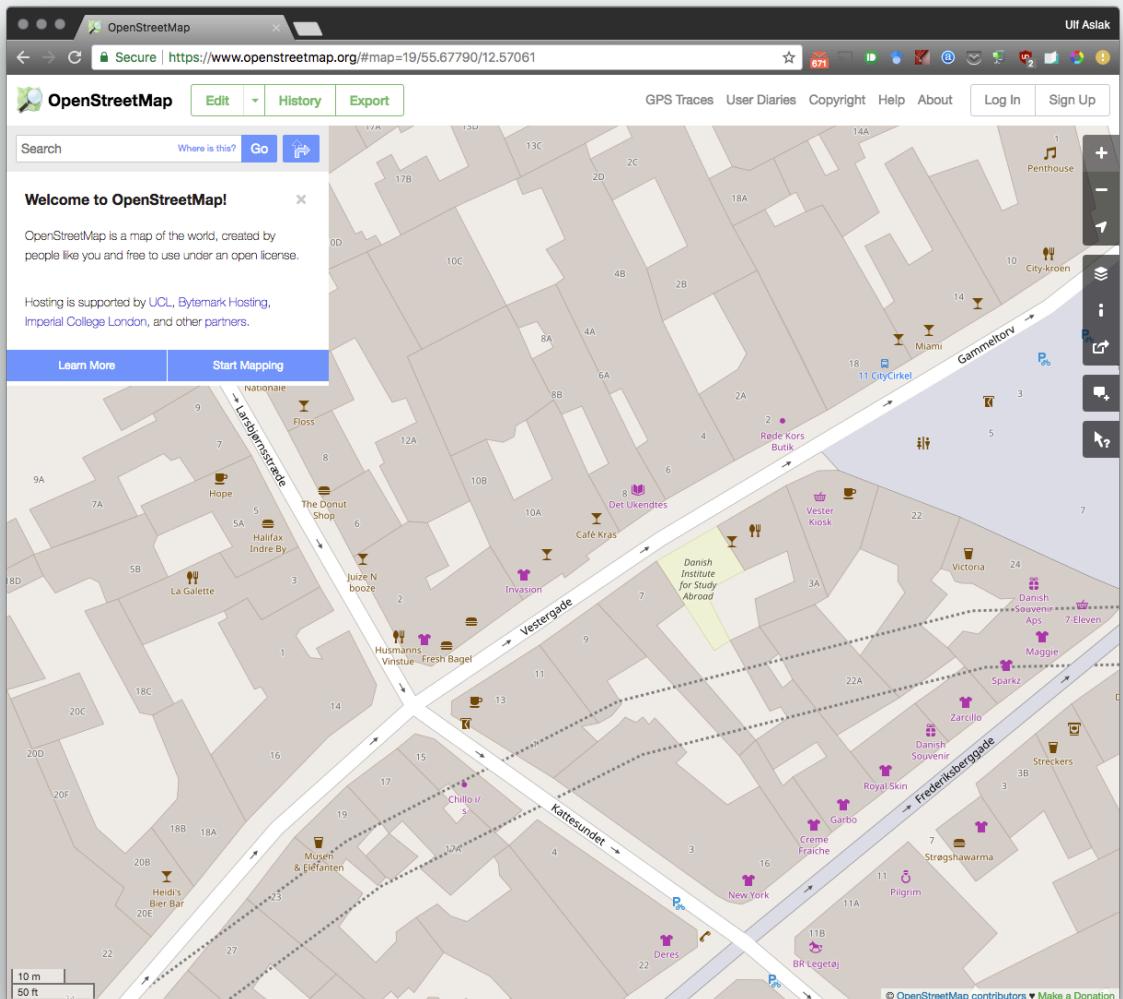
Open data

- **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

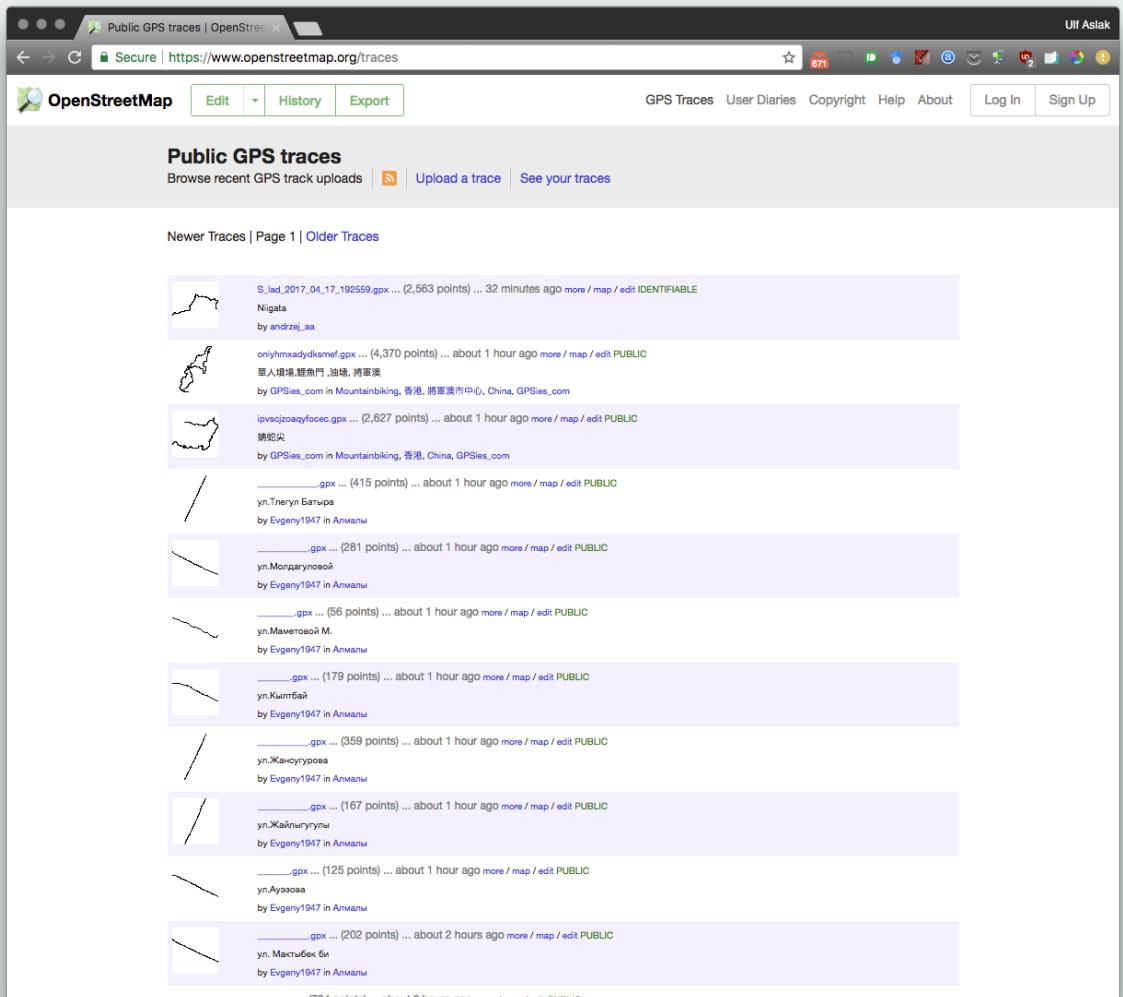
OpenStreetMap's community is diverse, passionate, and growing every day. Our contributors include enthusiast mappers, GIS professionals, engineers running the OSM servers, humanitarians mapping disaster-affected areas, and many more. To learn more about the community, see the [OpenStreetMap Blog](#),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

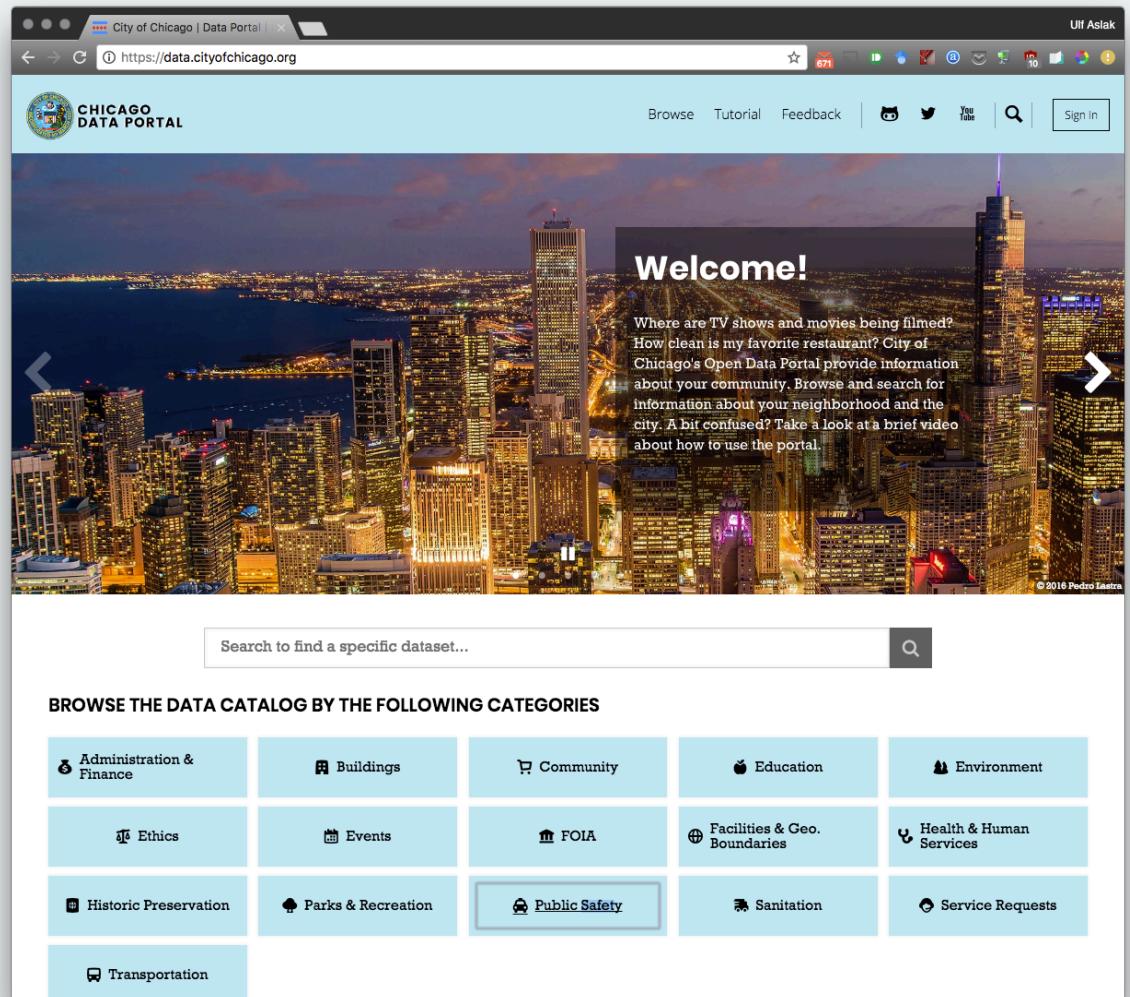
Open data

- **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 January 31, 2018	Metadata
Data Last Updated January 31, 2018	Time Period 2001 to present, minus the most recent seven days
Metadata Last Updated September 27, 2017	Frequency Data are updated daily.
Date Created September 30, 2011	Data Owner Police
Views 1,011,200	Topics
Downloads 1,011,200	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 is a navigation bar with links for 'Dataset' (which is highlighted), 'organizations', 'groups', and 'About'. There is 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 'Sort by: Relevance' dropdown menu. Below this, there are several sections:

- Air pollution**: A brief description stating that 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 is 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 is a sidebar with a tree view under 'organizations' and 'groups'. Under 'organizations', categories include 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), and a 'Show more Organizations' link. Under 'groups', categories include Geography (56) and Transport and infras ... (27).

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. The search bar at the top contains the query "homeland security", "medicare". Below the search bar, the results are displayed under the heading "CONGRESSIONAL RECORD". There are 1-100 of 577,488 results shown per page. The results are organized into several sections, each with a title and a list of items. For example, the first section is titled "CONGRESSIONAL RECORD" and lists "1. Daily Digest - Next Meeting of the SENATE + Next Meeting of the HOUSE OF REPRESENTATIVES + Other End Matter". Other sections include "COMMITTEE MEETINGS FOR 2018-02-02", "NEW PUBLIC LAWS", "House Committee Meetings", "Senate Committee Meetings", and "DAILY DIGEST - HIGHLIGHTS + SENATE". Each item in the list includes the issue and section date, the page number, and a link to the PDF.

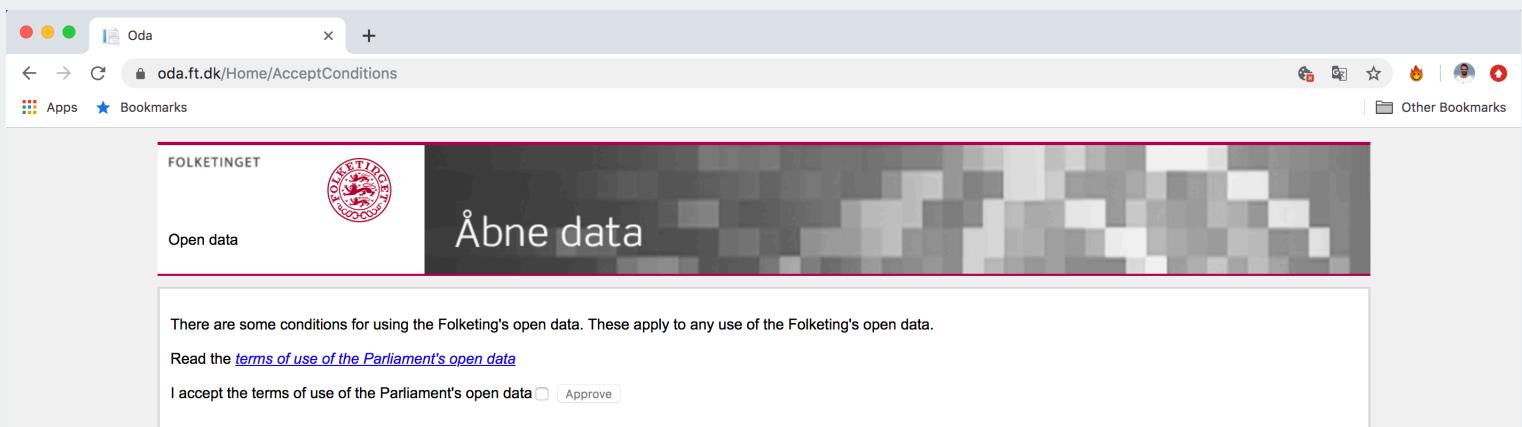
Open data

- 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.

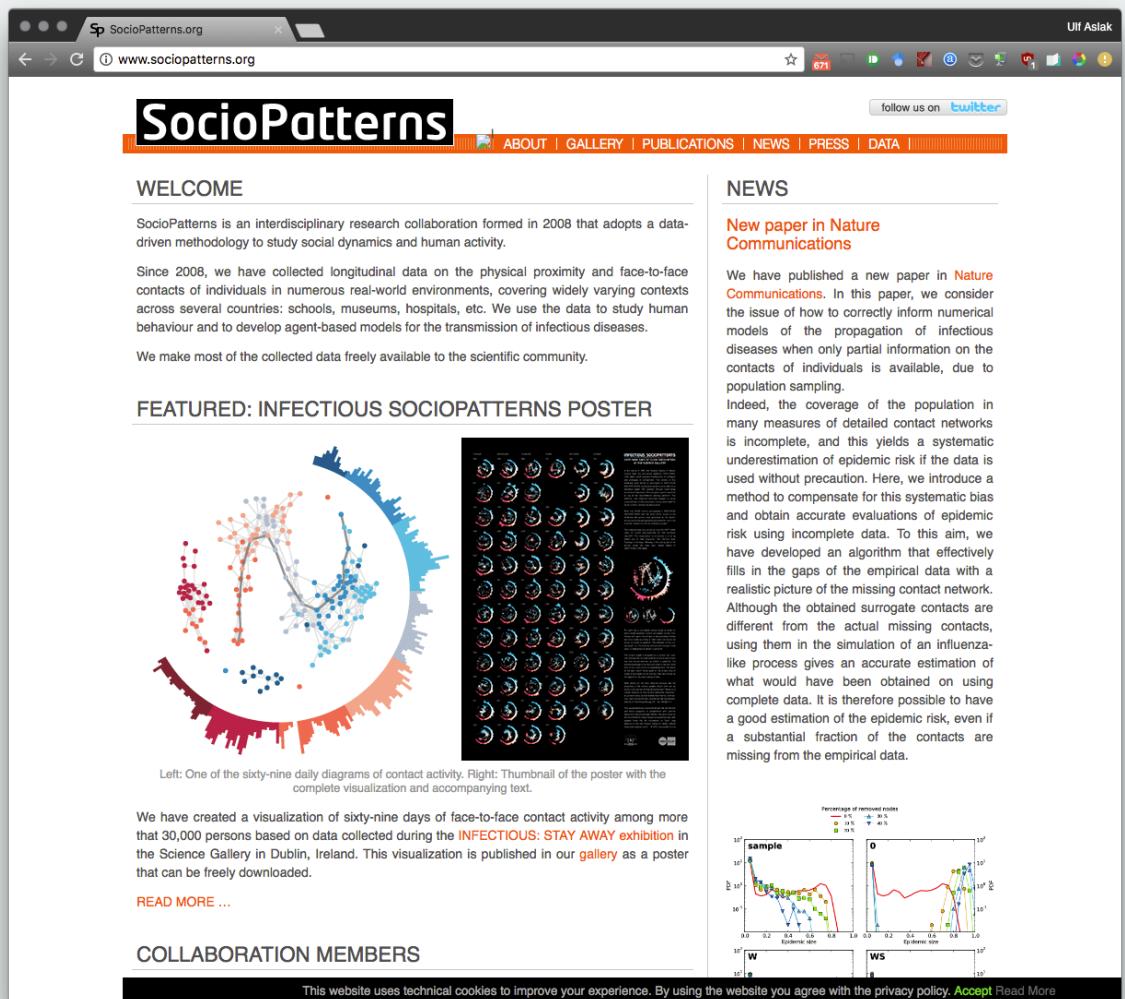
Open data

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



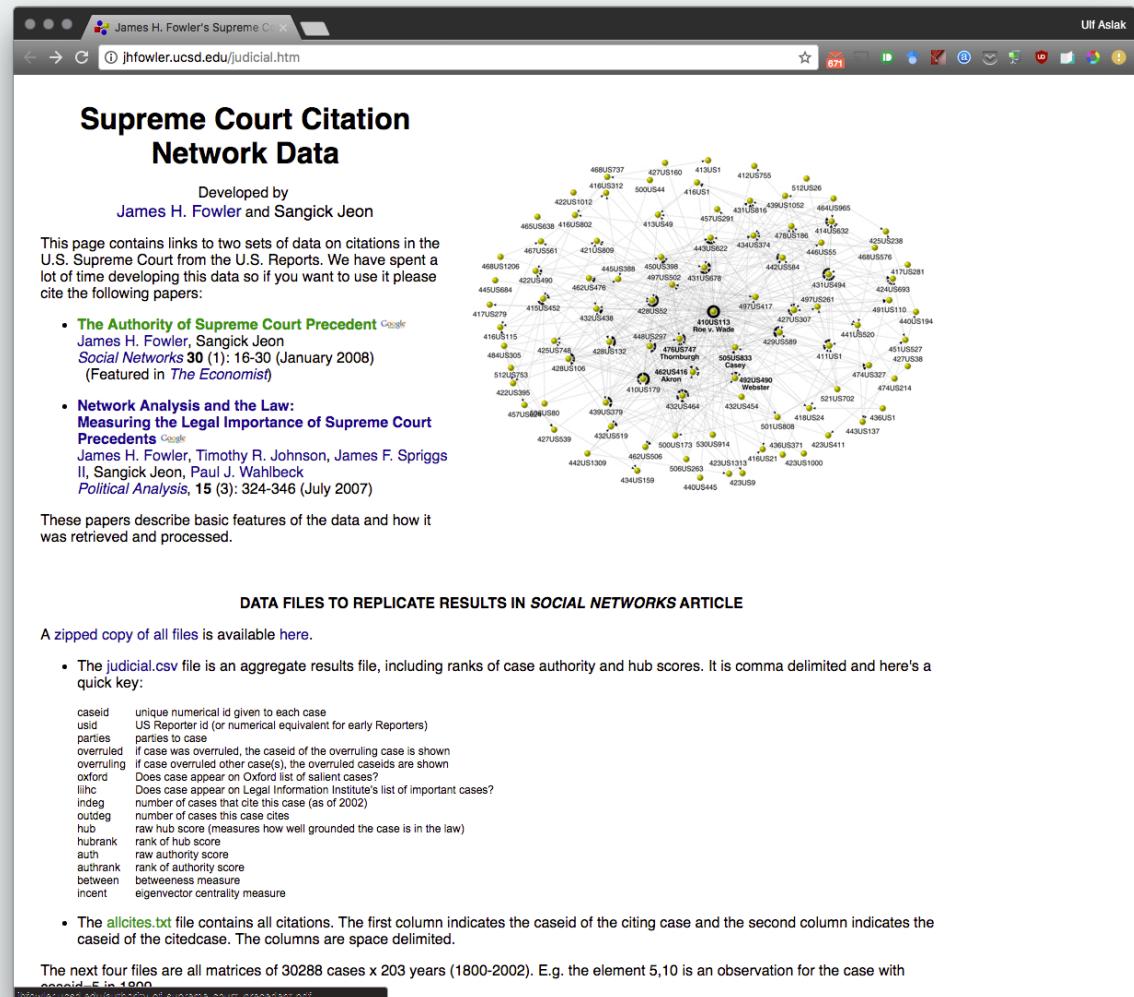
Open data

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



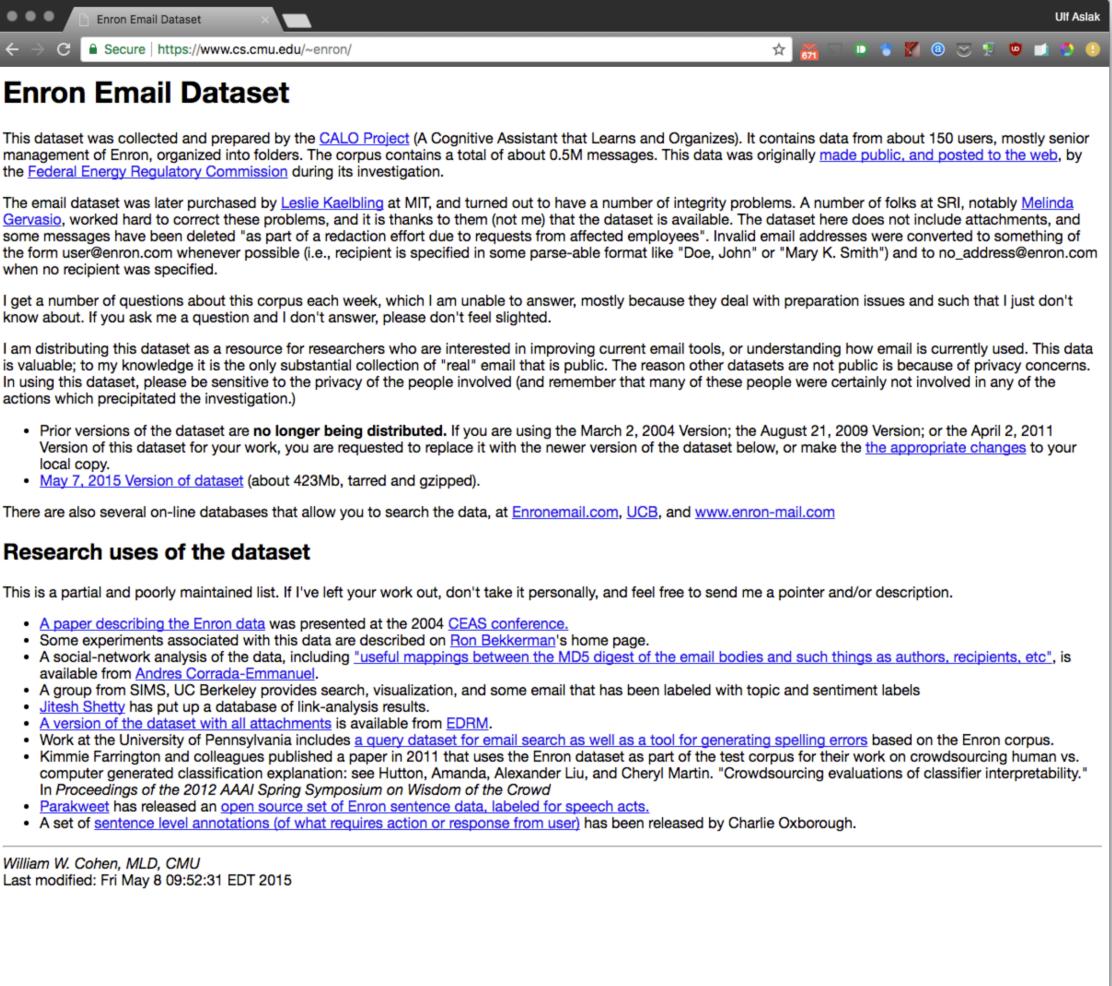
Open data

- 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

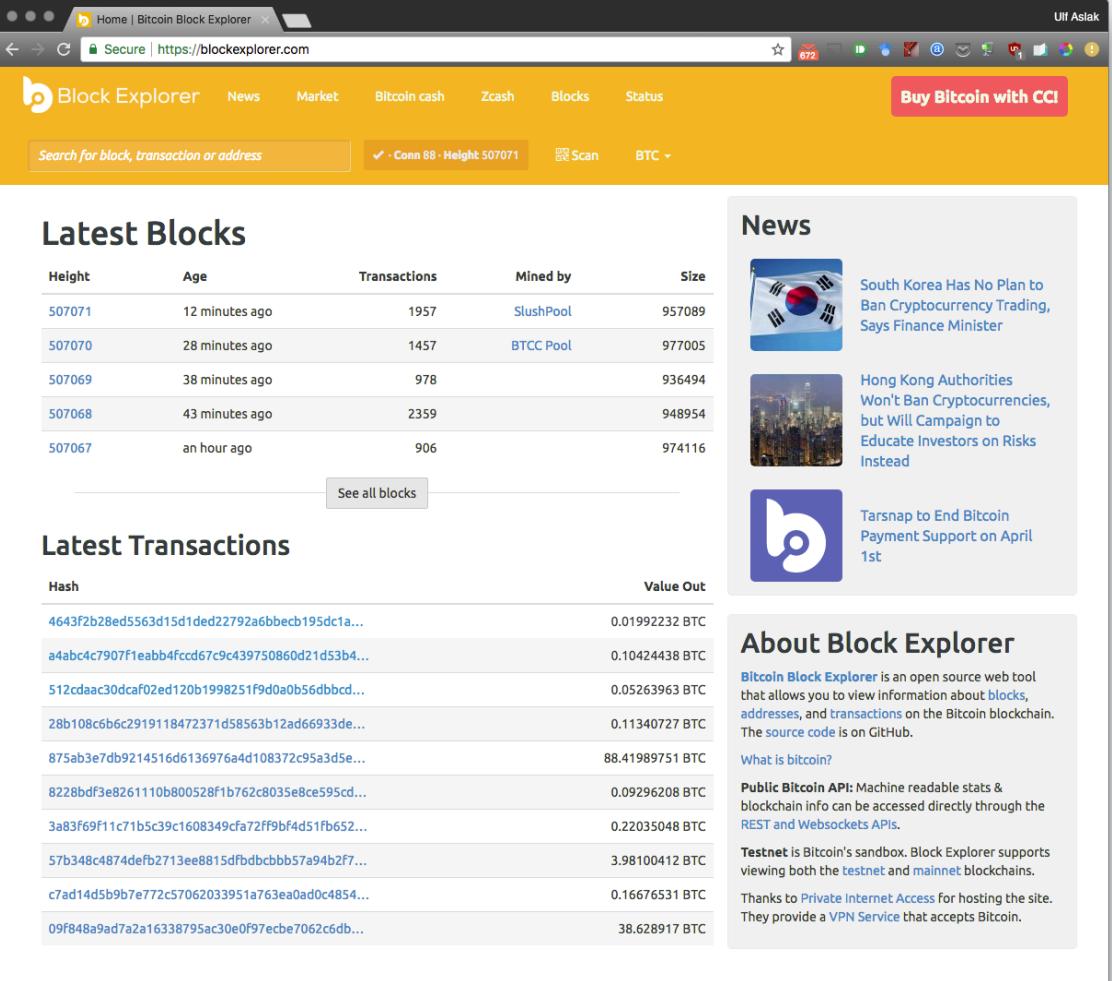
- Geographical data
- City data
- Political data
- Research data
- **Competition datasets**
- Transactional data

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, it says "10,348 Datasets". A search bar and a "Sort by" dropdown set to "Hotness" are also present. The main content area displays a list of datasets with the following details:

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

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 dropdown menu showing "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?](#)

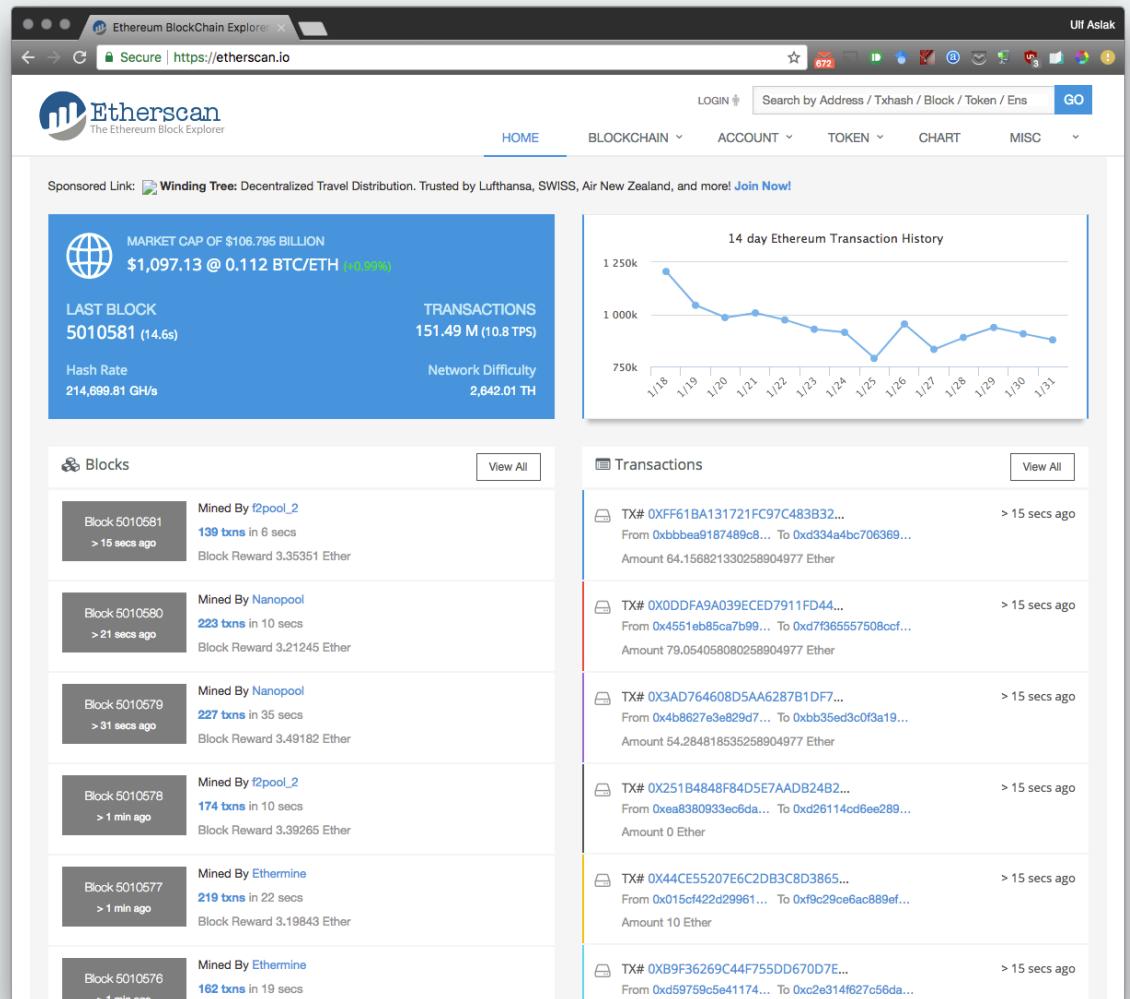
[Public Bitcoin API:](#) Machine readable stats & blockchain info can be accessed directly through the [REST](#) and [Websockets APIs](#).

[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

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

Batman

Art by Tony Daniel

Publication information

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

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 code demonstrates how to make a GET request to the Wikipedia API. The URL is constructed as a single string. Two annotations with arrows point to specific parts of this string: 'API address' points to the prefix 'https://en.wikipedia.org/w/api.php?', and 'Query parameters' points to the suffix '&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 dates|date=October 2015}}\n{{Infobox comic book 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* [[Outsider (team)|Outsider]]}}
```

(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". The main content area is divided into several sections:

- MOVIES OPENING THIS WEEK**: Shows movies like Winchester, A Fantastic Woman, and Braven opening on FEB 2.
- TOP BOX OFFICE**: Shows movies like Maze Runner: The Death Cure, Jumanji: Welcome to the Jungle, and Hostiles at the top of the box office.
- COMING SOON TO THEATERS**: Shows Fifty Shades Freed opening on FEB 9.
- NEW TV TONIGHT**: Shows new TV shows like A.P. Bio, The Good Place, and How to Get Away With Murder.
- MOST POPULAR TV ON RT**: Shows popular TV shows like Black Lightning, Counterpart, and The End of the F***ing World.
- TOP DVD & STREAMING MOVIES**: Shows movies like Babylon Berlin and American Crime Story.

On the right side of the page, there are social media sharing icons for Facebook, Twitter, and Google+.

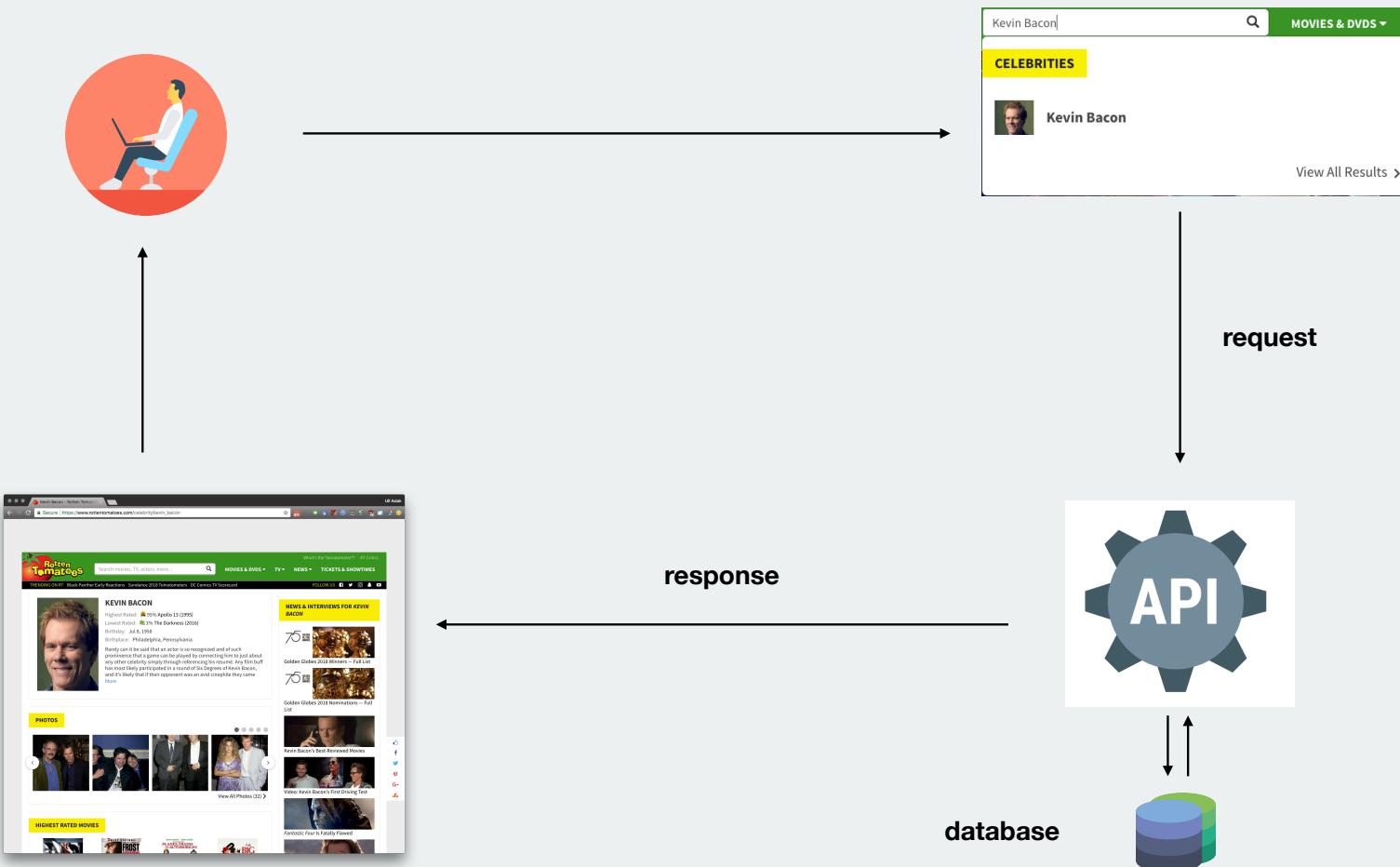
Scraping

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 various titles with their release dates or air dates and Rotten Tomatoes scores. The page has a green header and a white background with some dark blue and grey accents.

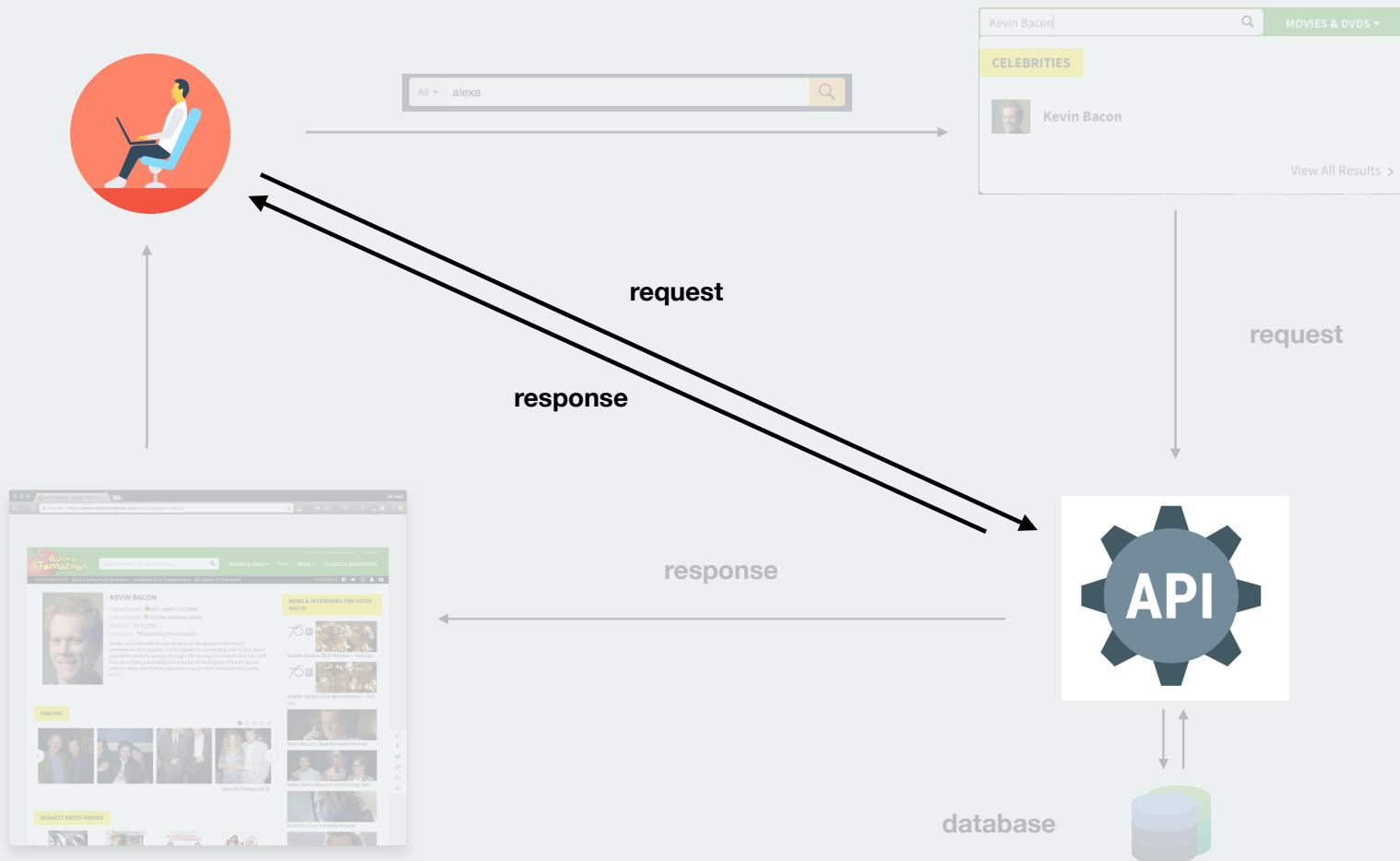
Scraping

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 features a large photo of Kevin Bacon and his biography. It includes facts like his highest rated movie (95% for Apollo 13) and lowest rated movie (3% for The Darkness). 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 with news and interview snippets about Kevin Bacon, including links to the Golden Globes 2018 winners and nominations. There are also video links for his best-reviewed movies and first driving test, along with a thumbnail for the movie 'Fantastic Four'. On the far right, there are social media sharing icons.

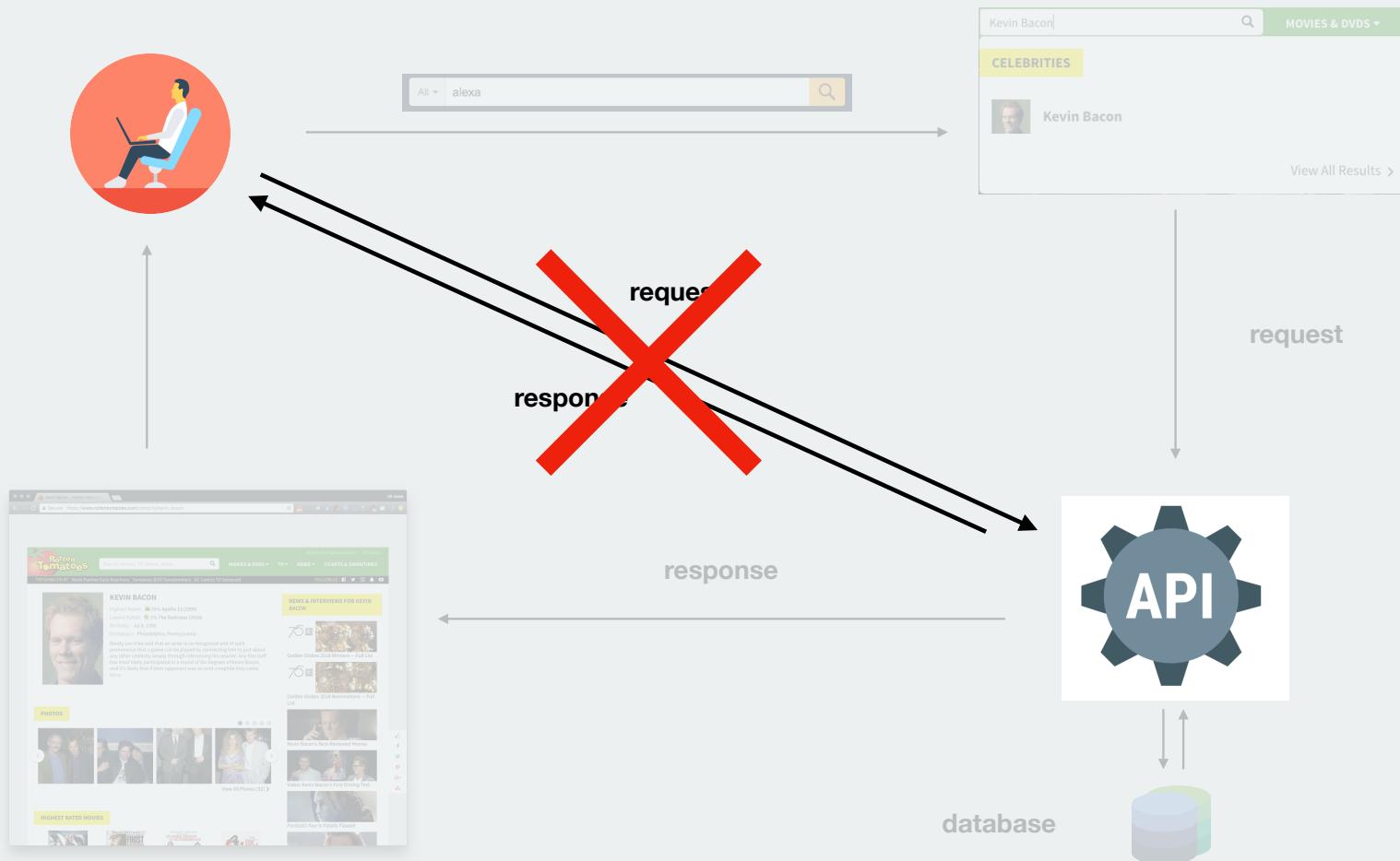
Scraping



Scraping



Scraping



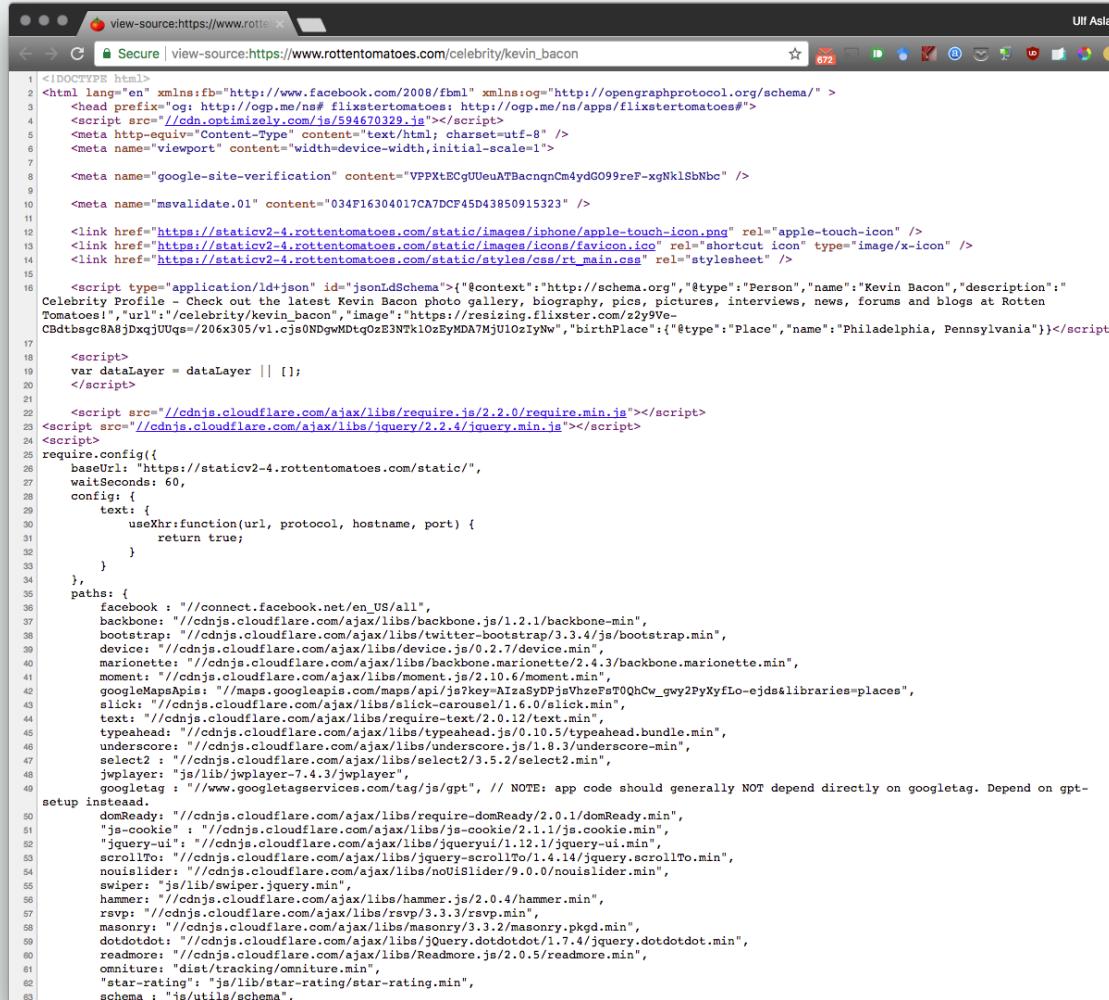
Scraping

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. Further down, there are sections for "HIGHEST RATED MOVIES" and "Video: Kevin Bacon's First Driving Test". On the far right, there are social media sharing icons.

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



```

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://og.me/ns# flixstertomatoes: http://og.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="msvalidate.01" content="034F16304017CA7DCP45D43850915323" />
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/z2y9Ve-CBdtbsgc8A8jDxqjUqs-/206x305/v1.cjsaONDgwMDtqzE3NTk1OzEyMDA7MjU0zIyNw","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=AIZasSyDPjzVhzeFaTQhCw_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/3.1.3/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    </script>
67  
```

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 :)