WHERE TO FIND DATA

FOR YOUR DATA SCIENCE PROJECTS

SOME IDEAS FOR FINDING DATA ONLINE

The Artists of Data Science
with Harpreet Sahota

KAGGLE.COM

Companies post a dataset and a question, and usually offer a prize for the best answer.

Kaggle also has discussion forums and "kernels" in which people share their code so you can learn how others approached the dataset.

As a result, Kaggle has thousands of datasets with accompanying questions and examples of how other people analyzed them.

The biggest benefit of Kaggle is also its biggest drawback: by handing you a (generally cleaned) dataset and problem, it's done a lot of the work for you.You also have thousands of people tackling the same problem, so it's difficult to make a unique contribution.

One way to use Kaggle is to take a dataset but pose a different question or do an exploratory analysis.

But generally, we think that Kaggle is best for learning by tackling a project and then seeing how you performed compared with others, thus learning from what their models did, rather than as a piece of your portfolio..

THE NEWS

Recently, many news companies have started making their data public.

FiveThirtyEight.com, for example, a website that focuses on opinion-poll analysis, politics, economics, and sports blogging, publishes data it can use for articles and even links to the raw data directly from the article website.

Although these datasets often require manual cleaning, the fact that they're in the news means that an obvious question is probably associated with them.

"IN GOD WE TRUST, ALL OTHERS BRING DATA."
— **W EDWARDS DEMING**

OPEN DATA PORTALS

A lot of government data is available online.

You can use census data, employment data, the general social survey, and tons of local government data such as New York City's 911 calls or traffic counts.

Sometimes you can download this data directly as a CSV file; at other times, you need to use an API.

You can even submit Freedom of Information Act requests to government agencies to get data that isn't publicly listed.

Government information is great because it's often detailed and deals with unusual subjects, such as data on the registered pet names of every animal in Elk Grove, California.

The downside - or potential upside as it poses a great challenge - of government information is that it often isn't well formatted, such as tables stored within PDF files.

"YOU CAN HAVE DATA WITHOUT INFORMATION, BUT YOU CANNOT HAVE INFORMATION WITHOUT DATA"
-**DANIEL KEYS MORAN**

APIS

APIs (application programming interfaces) are developer tools that allow you to access data directly from companies.

You know how you can type in a URL and get to a website?

APIs are like URLs, but instead of a website, you get data.

Some examples of companies with helpful APIs are **The New York Times, Yelp**, **Spotify, Netflix, or The Weather Channel.**

Some APIs even have R or Python packages that specifically make it easier to work with them. rtweet for R, for example, lets you pull Twitter data quickly so that you can find tweets with a specific hashtag, what the trending topics in Sacramento are, or what tweets Naval Ravikant is favoriting.

Keep in mind that there are limitations and terms of service to how you can use these APIs.

APIs are great for providing extremely robust, organized data from many sources.

YOUR OWN DATA

There are many places where you can download data about yourself; social media websites and email services are two big ones.

But if you use apps to keep track of your physical activity, reading list, budget, sleep, or anything else, you can usually download that data as well.

Maybe you could build a chatbot based on your emails with your colleagues or friends. Or you could look at the most common words you use in your tweets and how those words have changed over time.

Perhaps you could track your caffeine intake and exercise for a month to see whether you can predict how much and well you sleep.

The advantage of using your own data is that your project is guaranteed to be unique: no one else will have looked at that data before!

WEB SCRAPING

Web scraping is a way to extract data from websites that don't have an API, essentially by automating visiting web pages and copying the data.

You could create a program to search a movie website for a list of 100 actors, load their actor profiles, copy the lists of movies they're in, and put that data

in a spreadsheet. You do have to be careful, though: scraping a website can be against the website's terms of use, and you could be banned. You can check the robots.txt file of a website to find out what is allowed.

You also want to be nice to websites: if you hit a site too many times, you can bring it down.

But assuming that the terms of service allow it and you build in time between your hits, scraping can be a great way to get unique data.

"IF WE HAVE DATA, LET'S LOOK AT DATA. IF ALL WE HAVE ARE OPINIONS, LET'S GO WITH MINE."
— **JIM BARKSDALE**

The Artists of Data Science
with Harpreet Sahota

OPEN ML

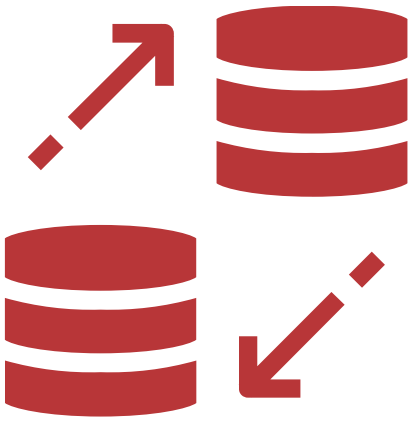
OpenML.org is an open science online platform for machine learning, which holds open data, open algorithms and tasks.

One of the core components of OpenML are datasets. People can upload their datasets, and the system automatically organizes these on line.

Each dataset has it's own unique ID. Information about the dataset, the data features and the data qualities can be obtained automatically by means of API functions, or downloaded manually as a CSV file.

Every dataset gets a dedicated page with all known information, including a wiki, visualizations, statistics, user discussions, and the tasks in which it is used.

At the time of this writing, OpenML has nearly 22,000 datasets available!



UC IRVINE

The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.

The archive was created as an ftp archive in 1987 by David Aha and fellow graduate students at UC Irvine.

Since that time, it has been widely used by students, educators, and researchers all over the world as a primary source of machine learning data sets.

As an indication of the impact of the archive, it has been cited over 1000 times, making it one of the top 100 most cited "papers" in all of computer science.

The current version of the web site was designed in 2007 by Arthur Asuncion and David Newman.

Visit it here:
https://archive.ics.uci.edu/ml/datasets.php

""ABOVE ALL ELSE, SHOW THE DATA."
– EDWARD R. TUFTE

PROPUBLICA

ProPublica is an independent, nonprofit newsroom that produces investigative journalism with moral force.

The ProPublica Data Store gives you access to the data behind our reporting and helps to sustain the challenging, expensive work of investigative reporting.

They provide free access to the raw data behind our work, as well as premium data products and custom data services. T

hese and other initiatives support ProPublica's mission of investigative journalism in the public interest.

Visit **propublica.org/datastore to** browse data sets about Health, Criminal Justice, Education, Politics, Business, Transportation, Military, Environment, Finance, or Religion.

GOOGLE DATASET SEARCH

You can get to Google's dataset search directly by visiting: **datasetsearch.research.google.com**

Similar to how Google Scholar works, Dataset Search lets you find datasets wherever they're hosted, whether it's a publisher's site, a digital library, or an author's personal web page.

Dataset Search enables users to find datasets stored across the Web through a simple keyword search.

The tool surfaces information about datasets hosted in thousands of repositories across the Web, making these datasets universally accessible and useful.

"ERRORS USING INADEQUATE DATA ARE MUCH LESS THAN THOSE USING NO DATA AT ALL."
– **CHARLES BABBAGE**

The Artists of Data Science
with Harpreet Sahota

OPEN DATA STACK EXCHANGE

Open Data Stack Exchange is a question and answer site for developers and researchers interested in open data.

It's built and run by the community as part of the Stack Exchange network of Q&A sites.

With the help of the research and data community, they're working together to build a library of detailed answers to every question about open data.

The site is all about getting answers. It's not a discussion forum.

Good answers are voted up and rise to the top. The best answers show up first so that they are always easy to find.

At this time the site is in beta mode, but it is still *very* useful to help you find data - or ask someone if they know where the type of data you're looking for can be found.

Visit **opendata.stackexchange.com** to check it out.

REDDIT DATASETS

Reddit's /r/datasets is a place to share, find, and discuss Datasets.

Users have posted an eclectic mix of datasets about gun ownership, NYPD crime rates, college student study habits and caffeine concentrations in popular beverages.

You're sure to find awesome data here. Visit **reddit.com/r/datasets** to learn more.

"WITHOUT A SYSTEMATIC WAY TO START AND KEEP DATA CLEAN, BAD DATA WILL HAPPEN."
— DONATO DIORIO

AWS OPEN DATA

Amazon makes large data sets available on its Amazon Web Services platform: **aws.amazon.com/opendata**

You can download the data and work with it on your own computer, or analyze the data in the cloud using EC2 and Hadoop via EMR.

You can read more about how the program works here.Amazon has a page that lists all of the data sets for you to browse.

You'll need an AWS account, although Amazon gives you a free access tier for new accounts that will enable you to explore the data without being charged.

Here are some examples:

- Lists of n-grams from Google Books — common words and groups of words from a huge set of books.
- Common Crawl Corpus — data from a crawl of over 5 billion web pages.
- Landsat images — moderate resolution satellite images of the surface of the Earth.

ACADEMIC TORRENTS

Academic Torrents is designed to facilitate storage of all the data used in research, including datasets as well as publications.

It's a distributed system for sharing enormous datasets - for researchers, by researchers.

The result is a scalable, secure, and fault-tolerant repository for data, with blazing fast download speeds.

Checkout **academictorrents.com** to see all they have to offer as well as documentation about their API.

"WITH DATA COLLECTION, 'THE SOONER THE BETTER' IS ALWAYS THE BEST ANSWER."
– **MARISSA MAYER**

WIKIPEDIA

As part of Wikipedia's commitment to advancing knowledge, they offer all of their content for free, and regularly generate dumps of all the articles on the site.

Additionally, Wikipedia offers edit history and activity, so you can track how a page on a topic evolves over time, and who contributes to it.

You can find the various ways to download the data on the Wikipedia site. You'll also find scripts to reformat the data in various ways.

Here are some examples:

- All images and other media from Wikipedia — all the images and other media files on Wikipedia.
-
- Full site dumps — of the content on Wikipedia, in various formats.

BIGQUERY PUBLIC DATA

Much like Amazon, Google also has a cloud hosting service, called Google Cloud Platform.

With GCP, you can use a tool called BigQuery to explore large data sets.

Visit here to learn more: **cloud.google.com/bigquery/public-data**

Here are some examples:

- USA Names — contains all Social Security name applications in the US, from 1879 to 2015.
- Github Activity — contains all public activity on over 2.8 million public Github repositories.
- Historical Weather — data from 9000 NOAA weather stations from 1929 to 2016.

"WE'RE ENTERING A NEW WORLD IN WHICH DATA MAY BE MORE IMPORTANT THAN SOFTWARE."
– TIM O'REILLY

QUANDL

Quandl is a repository of economic and financial data.

Some of this information is free, but many data sets require purchase.

Quandl is useful for building models to predict economic indicators or stock prices.

Due to the large amount of available data sets, it's possible to build a complex model that uses many data sets to predict values in another.

Visit **quandl.com/search** to browse available dataset.

Here are some examples you might find:

- Entrepreneurial activity by race and other factors — contains data from the Kauffman foundation on entrepreneurs in the US.
- Chinese macroeconomic data — indicators of Chinese economic health.
- US Federal Reserve data — US economic indicators, from the Federal Reserve..

DATA.GOV

Data.gov makes it possible to download data from multiple US government agencies.

Data can range from government budgets to school performance scores.

Much of the data requires additional research, and it can sometimes be hard to figure out which data set is the "correct" version.

Anyone can download the data, although some data sets require additional hoops to be jumped through, like agreeing to licensing agreements.

Here are some examples:

- Food Environment Atlas — contains data on how local food choices affect diet in the US.
- School system finances — a survey of the finances of school systems in the US.
- Chronic disease data — data on chronic disease indicators in areas across the US.

"THINK ANALYTICALLY, RIGOROUSLY, AND SYSTEMATICALLY ABOUT A BUSINESS PROBLEM AND COME UP WITH A SOLUTION THAT LEVERAGES THE AVAILABLE DATA."
– MICHAEL O'CONNELL