**Best Public Datasets for Machine Learning and Data Science**

**Best public datasets for** [**machine learning**](https://mld.ai/mldcmu)**, data science, sentiment analysis, computer vision, natural language processing (NLP), clinical data, and others.**

**Author(s):** Stacy Stanford, [Roberto Iriondo](https://mktg.best/vguzs), [Pratik Shukla](https://www.linkedin.com/in/pratik-shukla28/)

Last updated January 6, 2021

This resource is continuously updated. If you know of any other suitable and open datasets, please let us know by emailing us at [**pub@towardsai.net**](mailto:pub@towardsai.net)or by dropping a comment below.

**Dataset Finders**

[**Google Dataset Search**](https://toolbox.google.com/datasetsearch): Similar to how [Google Scholar](https://scholar.google.com/) works, Dataset Search lets you find datasets wherever they are hosted, whether it’s a publisher’s site, a digital library, or an author’s web page. It’s a phenomenal dataset finder, and it contains over 25 million datasets.

[**Kaggle**](https://www.kaggle.com/): Kaggle provides a vast container of datasets, sufficient for the enthusiast to the expert.

[**UCI Machine Learning Repository**](http://mlr.cs.umass.edu/ml/): The Machine Learning Repository at UCI provides an up to date resource for open-source datasets.

[**VisualData**](https://www.visualdata.io/): Discover computer vision datasets by category; it allows searchable queries.

[**CMU Libraries**](https://guides.library.cmu.edu/machine-learning/datasets)**:** Discover high-quality datasets thanks to the collection of Huajin Wang, at CMU.

[**The Big Bad NLP Database**](https://datasets.quantumstat.com/): This cool dataset list contains datasets for various natural language processing tasks, created and curated by [**Quantum Stat**](https://quantumstat.com/).

**General Datasets**

**Housing Datasets**

[**Boston Housing Dataset:**](https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html) **C**ontains information collected by the US Census Service concerning housing in the area of Boston Mass. It was obtained from the [StatLib archive](http://lib.stat.cmu.edu/datasets/boston) and has been used extensively throughout the literature to benchmark algorithms.

**Geographic Datasets**

[**Google-Landmarks-v2:**](https://www.kaggle.com/xiuchengwang/python-dataset-download) An improved dataset for landmark recognition and retrieval. This dataset contains 5M+ images of 200k+ landmarks from across the world, sourced and annotated by the [W](https://twitter.com/WikiCommons)iki Commons community.

**Machine Learning Datasets:**

[**Mall Customers Dataset**](https://www.kaggle.com/shwetabh123/mall-customers)**:** The Mall customers dataset contains information about people visiting the mall in a particular city. The dataset consists of various columns like gender, customer id, age, annual income, and spending score. It’s generally used to segment customers based on their age, income, and interest.

[**IRIS Dataset:**](https://archive.ics.uci.edu/ml/datasets/Iris) The iris dataset is a simple and beginner-friendly dataset that contains information about the flower petal and sepal width. The data is divided into three classes, with 50 rows in each class. It’s generally used for classification and regression modeling.

[**MNIST Dataset:**](http://yann.lecun.com/exdb/mnist/) This is a database of handwritten digits. It contains 60,000 training images and 10,000 testing images. This is a perfect dataset to start implementing image classification where you can classify a digit from 0 to 9.

[**Boston Housing Dataset:**](https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html) **C**ontains information collected by the US Census Service concerning housing in the area of Boston Mass. It was obtained from the [StatLib archive](http://lib.stat.cmu.edu/datasets/boston) and has been used extensively throughout the literature to benchmark algorithms.

[**Fake News Detection Dataset**](https://www.kaggle.com/c/fake-news/data)**:** It is a CSV file that has 7796 rows with four columns. There are four columns: news, title, news text, result.

[**Wine quality dataset**](https://archive.ics.uci.edu/ml/datasets/wine+quality)**:** The dataset contains different chemical information about the wine. The dataset is suitable for classification and regression tasks.

[**SOCR data — Heights and Weights Dataset**](http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_020108_HeightsWeights)**:** This is a basic dataset for beginners. It contains only the height and weights of 25,000 different humans of 18 years of age. This dataset can be used to build a model that can predict the height or weight of a human.

[**Titanic Dataset**](https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/problem12.html)**:** The dataset contains information like name, age, sex, number of siblings aboard, and other information about 891 passengers in the training set and 418 passengers in the testing set.

[**Credit Card Fraud Detection Dataset**](https://www.kaggle.com/mlg-ulb/creditcardfraud)**:** The dataset contains transactions made by credit cards; they are labeled as fraudulent or genuine. This is important for companies that have transaction systems to build a model for detecting fraudulent activities.

**Computer Vision Datasets**

[**xView**](http://xviewdataset.org/#dataset)**:** xView is one of the most massive publicly available datasets of overhead imagery. It contains images from complex scenes around the world, annotated using bounding boxes.

[**ImageNet**](http://image-net.org/): The largest image dataset for computer vision. It provides an accessible image database that is organized hierarchically, according to WordNet.

[**Kinetics-700**](https://deepmind.com/research/open-source/open-source-datasets/kinetics/)**:** A large-scale dataset of video URLs from Youtube. Including human-centered actions. It contains over 700,000 videos.

[**Google’s Open Images**](https://research.googleblog.com/2016/09/introducing-open-images-dataset.html)**:** A vast dataset from Google AI containing over 10 million images.

[**Cityscapes Dataset**](https://www.cityscapes-dataset.com/)**:** This is an open-source dataset for Computer Vision projects. It contains high-quality pixel-level annotations of video sequences taken in 50 different city streets. The dataset is useful in semantic segmentation and training deep neural networks to understand the urban scene.

[**IMDB-Wiki dataset**](https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/)**:** The IMDB-Wiki dataset is one of the most extensive open-source datasets for face images with labeled gender and age. The images are collected from IMDB and Wikipedia. It has five million-plus labeled images.

[**Color Detection Dataset**](https://github.com/codebrainz/color-names/blob/master/output/colors.csv)**:** The dataset contains a CSV file that has 865 color names with their corresponding RGB(red, green, and blue) values of the color. It also has the hexadecimal value of the color.

[**Stanford Dogs Dataset:**](http://vision.stanford.edu/aditya86/ImageNetDogs/)It contains 20,580 images and 120 different dog breed categories.

**Sentiment Analysis Datasets**

[**Lexicoder Sentiment Dictionary**](http://www.lexicoder.com/)**:** This dataset is specific for sentiment analysis. The dataset contains over 3000 negative words and over 2000 positive sentiment words.

[**IMDB reviews**](https://www.kaggle.com/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews): An interesting dataset with over 50,000 movie reviews from Kaggle.

[**Stanford Sentiment Treebank**](http://nlp.stanford.edu/sentiment/code.html): Standard sentiment dataset with sentiment annotations.

[**Twitter US Airline Sentiment**](https://www.kaggle.com/crowdflower/twitter-airline-sentiment)**:** Twitter data on US airlines from February 2015, classified as positive, negative, and neutral tweets

**Natural Language Processing (NLP) Datasets**

[**The Big Bad NLP Database**](https://datasets.quantumstat.com/): This cool dataset list contains datasets for various natural language processing tasks, created and curated by [**Quantum Stat**](https://quantumstat.com/).

[**HotspotQA Dataset**](https://hotpotqa.github.io/): Question answering dataset featuring natural, multi-hop questions, with intense supervision for supporting facts to enable more explainable question answering systems.

[**Amazon Review**s](https://snap.stanford.edu/data/web-Amazon.html): A vast dataset from Amazon, containing over 45 million Amazon reviews.

[**Rotten Tomatoes Reviews**](https://drive.google.com/file/d/1w1TsJB-gmIkZ28d1j7sf1sqcPmHXw352/view): Archive of more than 480,000 critic reviews (fresh or rotten).

[**SMS Spam Collection in English**](http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/): A dataset that consists of 5,574 English SMS spam messages.

[**Enron Email Dataset**](https://www.cs.cmu.edu/~enron/)**:** It contains around 0.5 million emails of over 150 users.

[**Recommender Systems Dataset**](https://cseweb.ucsd.edu/~jmcauley/datasets.html)**:** It contains various datasets from popular websites like Goodreads book reviews, Amazon product reviews, bartending data, data from social media, and others that are used in building a recommender system.

[**UCI Spambase Dataset**](https://archive.ics.uci.edu/ml/datasets/Spambase)**:** Classifying emails as spam or non-spam is a prevalent and useful task. The dataset contains 4601 emails and 57 meta-information about the emails. You can build models to filter out the spam.

[**IMDB reviews**](http://ai.stanford.edu/~amaas/data/sentiment/)**:** The large movie review dataset consists of movie reviews from IMDB website with over 25,000 reviews for training and 25,000 for the testing set.

**Self-driving (Autonomous Driving) Datasets**

[**Waymo Open Dataset**](https://waymo.com/open/)**:** This is a fantastic dataset resource from the folks at Waymo. Includes a vast dataset of autonomous driving, enough to train deep nets from zero.

[**Berkeley DeepDrive BDD100k**:](http://bdd-data.berkeley.edu/) One of the largest datasets for self-driving cars, containing over 2000 hours of driving experiences across New York and California.

[**Bosch Small Traffic Light Dataset**](https://hci.iwr.uni-heidelberg.de/node/6132): Dataset for small traffic lights for deep learning.

[**LaRa Traffic Light Recognition**](http://www.lara.prd.fr/benchmarks/trafficlightsrecognition): Another dataset for traffic lights. This dataset is gathered from Paris.

[**WPI datasets**](http://computing.wpi.edu/dataset.html): Datasets for traffic lights, pedestrian, and lane detection.

[**Comma.ai**](https://archive.org/details/comma-dataset)**:** It contains details such as a car’s speed, acceleration, steering angle, and GPS coordinates.

[**MIT AGE Lab**](http://lexfridman.com/automated-synchronization-of-driving-data-video-audio-telemetry-accelerometer/)**:** A sample of the 1,000+ hours of multi-sensor driving datasets collected at AgeLab.

[**LISA: Laboratory for Intelligent & Safe Automobiles, UC San Diego Datasets**](http://cvrr.ucsd.edu/LISA/datasets.html)**:** This dataset includes traffic signs, vehicle detection, traffic lights, and trajectory patterns.

[**Cityscape Dataset**](https://www.cityscapes-dataset.com/)**:** This is an extensive dataset that has street scenes in 50 different cities.

**Clinical Datasets**

[**MaskedFace-Net**](https://github.com/cabani/MaskedFace-Net)**:** MaskedFace-Net is a real dataset containing human faces with correct and incorrectly worn masks. It contains over 137k images which are based on the [**Flick-Faces-HQ dataset**](https://github.com/NVlabs/ffhq-dataset) [21]. For more details about the dataset and its uses, please visit the [**documentation on Github**](https://github.com/cabani/MaskedFace-Net).

[**COVID-19 Dataset**](https://www.semanticscholar.org/cord19): The Allen Institute of AI research has released a vast research dataset of over 45,000 scholarly articles about COVID-19.

[**MIMIC-III**](https://mimic.physionet.org/): Openly available dataset developed by the MIT Lab for Computational Physiology, comprising de-identified health data associated with ~40,000 critical care patients. It includes demographics, vital signs, laboratory tests, medications, and more.

**Datasets for Recommender Systems**

[**MovieLens**](https://grouplens.org/datasets/movielens/)**:** It contains rating data sets from the MovieLens web site.

[**Jester**](http://www.ieor.berkeley.edu/~goldberg/jester-data/)**:** It contains 4.1 Million continuous ratings (-10.00 to +10.00) of 100 jokes from 73,421 users. It’s mostly used for the collaborative filter.

[**Million Song Dataset**](https://www.kaggle.com/c/msdchallenge#description)**:** It can be used for both collaborative and content-based filtering.