

PROJECTS AND DATASETS

The following is a complete set of resources for datasets and machine/Deep learning projects.

General-Purpose Dataset Repositories

1. **Kaggle Datasets**

- Website: <https://www.kaggle.com/datasets>
- Description: A vast collection of datasets for competitions, research, and practice. Covers almost every domain.

2. **UCI Machine Learning Repository**

- Website: <https://archive.ics.uci.edu/ml/index.php>
- Description: A classic repository of datasets for machine learning, including classification, regression, and clustering tasks.

3. **Google Dataset Search**

- Website:
<https://datasetsearch.research.google.com/>
- Description: A search engine for datasets across the web, including academic, government, and commercial sources.

4. **AWS Open Data Registry**

- Website: <https://registry.opendata.aws/>
- Description: A collection of datasets hosted on AWS, covering domains like genomics, satellite imagery, and more.

5. **Microsoft Research Open Data**

- Website: <https://msropendata.com/>
- Description: Datasets released by Microsoft Research for academic and research purposes.

6. **Data.gov**

- Website: <https://www.data.gov/>
- Description: The U.S. government's open data portal, offering datasets on topics like climate, education, and healthcare.

7. **OpenML**

- Website: <https://www.openml.org/>
- Description: A platform for sharing datasets, algorithms, and experiments in machine learning.

Computer Vision Datasets

1. **ImageNet**

- Website: <http://www.image-net.org/>
- Description: A large-scale dataset for image classification, object detection, and segmentation.

2. **COCO (Common Objects in Context)**

- Website: <https://cocodataset.org/>
- Description: A dataset for object detection, segmentation, and captioning.

3. **CIFAR-10 and CIFAR-100**

- Website:
<https://www.cs.toronto.edu/~kriz/cifar.html>
- Description: Small image datasets for classification tasks.

4. **Pascal VOC**

- Website: <http://host.robots.ox.ac.uk/pascal/VOC/>
- Description: A dataset for object detection and segmentation.

5. **MNIST**

- Website: <http://yann.lecun.com/exdb/mnist/>
- Description: A classic dataset of handwritten digits for classification tasks.

6. ****Open Images Dataset****

- Website: <https://storage.googleapis.com/openimages/web/index.html>
- Description: A large dataset with annotated images for object detection and segmentation.

7. ****Fashion-MNIST****

- Website: <https://github.com/zalandoresearch/fashion-mnist>
- Description: A dataset of fashion items, designed as a drop-in replacement for MNIST.

8. ****Cityscapes Dataset****

- Website: <https://www.cityscapes-dataset.com/>
- Description: A dataset for urban scene understanding, including semantic segmentation.

Natural Language Processing (NLP) Datasets

1. ****GLUE (General Language Understanding Evaluation)****

- Website: <https://gluebenchmark.com/>
- Description: A benchmark dataset for evaluating NLP models on tasks like sentiment analysis and textual entailment.

2. ****SQuAD (Stanford Question Answering Dataset)****

- Website: <https://rajpurkar.github.io/SQuAD-explorer/>
- Description: A dataset for question answering tasks.

3. ****Common Crawl****

- Website: <https://commoncrawl.org/>
- Description: A massive dataset of web-crawled data for training large language models.

4. ****WikiText****

- Website: <https://blog.salesforceairesearch.com/the-wikitext-long-term-dependency-language-modeling-dataset/>
- Description: A dataset for language modeling, derived from Wikipedia articles.

5. ****Yelp Open Dataset****

- Website: <https://www.yelp.com/dataset>
- Description: A dataset of Yelp reviews for sentiment analysis and recommendation systems.

6. ****IMDB Movie Reviews Dataset****

- Website:
<https://ai.stanford.edu/~amaas/data/sentiment/>
- Description: A dataset for binary sentiment classification.

7. ****Gutenberg Project****

- Website: <https://www.gutenberg.org/>
- Description: A collection of free eBooks for text-based NLP tasks.

Audio and Speech Datasets

1. ****LibriSpeech****

- Website: <http://www.openslr.org/12/>

- Description: A dataset of audiobooks for speech recognition tasks.

2. **Common Voice by Mozilla**

- Website: <https://commonvoice.mozilla.org/>
- Description: A crowdsourced dataset of voice recordings for speech recognition.

3. **UrbanSound8K**

- Website: <https://urbansounddataset.weebly.com/>
- Description: A dataset of urban sounds for audio classification.

4. **VoxCeleb**

- Website:
<https://www.robots.ox.ac.uk/~vgg/data/voxceleb/>
- Description: A dataset for speaker identification and verification.

Healthcare and Biomedical Datasets

1. **MIMIC-III (Medical Information Mart for Intensive Care III)**

- Website: <https://mimic.physionet.org/>
- Description: A dataset of de-identified health data for predictive modeling in healthcare.

2. **TCGA (The Cancer Genome Atlas)**

- Website: <https://www.cancer.gov/tcga>
- Description: A dataset for cancer research, including genomic and clinical data.

3. **CheXpert**

- Website:

<https://stanfordmlgroup.github.io/competitions/chexpert/>

- Description: A dataset of chest X-rays for medical image analysis.

4. **PhysioNet**

- Website: <https://physionet.org/>

- Description: A repository of physiological signals and medical data.

Reinforcement Learning Datasets

1. **OpenAI Gym**

- Website: <https://www.gymnasium.dev/>

- Description: A toolkit for developing and comparing reinforcement learning algorithms.

2. **DeepMind Control Suite**

- Website:

[https://github.com/deepmind/dm_control](https://github.com/deepmind/dm_control)

- Description: A suite of tasks for testing reinforcement learning agents.

Time Series and Financial Datasets

1. **Yahoo Finance**

- Website: <https://finance.yahoo.com/>

- Description: Historical stock price data for time series forecasting.

2. **UCR/UEA Time Series Classification Archive**

- Website:

<https://www.timeseriesclassification.com/>

- Description: A collection of time series datasets for classification tasks.

3. **Kaggle Time Series Datasets**

- Website: [[https://www.kaggle.com/datasets?tags=13205-](https://www.kaggle.com/datasets?tags=13205-Time+Series)

[Time+Series](https://www.kaggle.com/datasets?tags=13205-Time+Series)](<https://www.kaggle.com/datasets?tags=13205-Time+Series>)

- Description: A variety of time series datasets for forecasting and analysis.

Social Network and Graph Datasets

1. **Stanford Large Network Dataset Collection (SNAP)**

- Website: <https://snap.stanford.edu/data/>

- Description: A collection of datasets for social network analysis and graph-based machine learning.

2. **Open Graph Benchmark (OGB)**

- Website: <https://ogb.stanford.edu/>

- Description: Benchmark datasets for graph-based machine learning tasks.

Miscellaneous Datasets

1. **NASA Open Data Portal**

- Website: <https://data.nasa.gov/>

- Description: Datasets related to space exploration, climate, and more.

2. **World Bank Open Data**

- Website: <https://data.worldbank.org/>

- Description: Global development data for economic and social research.

3. ****UNICEF Data****

- Website: <https://data.unicef.org/>
- Description: Datasets related to child welfare and global development.