Data Sources

Where to get the Data for Data Science and Machine Learning projects?

1. Kaggle

- Description: One of the largest platforms for data science and machine learning competitions, Kaggle offers a vast collection of datasets across various domains.

- Website: https://www.kaggle.com/datasets

- Popular Uses: Competitions, projects, and learning resources.

2. UCI Machine Learning Repository

- Description: A long-standing resource for machine learning datasets, hosted by the University of California, Irvine. It provides hundreds of datasets used in academic research.

- Website: https://archive.ics.uci.edu/ml/index.php

- Popular Uses: Academic research, training models, teaching.

3. Google Dataset Search

- Description: A specialized search engine from Google that helps you find datasets across the web. It indexes datasets from multiple sources and repositories.

- Website: https://datasetsearch.research.google.com/

- Popular Uses: Finding niche and domain-specific datasets.

4. AWS Open Data Registry

- Description: Amazon provides a collection of publicly available datasets, which can be accessed and analyzed via their cloud services.

- Website: https://registry.opendata.aws/

- Popular Uses: Large-scale datasets for cloud-based machine learning.

5. US Government

- Description: The U.S. government's open data portal, with thousands of datasets from federal agencies, on topics ranging from health to education and the environment.

- Website: https://www.data.gov/

- Popular Uses: Policy analysis, research, data visualization.

6. Google BigQuery Public Datasets

- Description: Google provides a collection of public datasets that are integrated with BigQuery, enabling powerful SQL queries on massive data.

- Website: https://cloud.google.com/bigquery/public-data

- Popular Uses: Large-scale data analysis, SQL-based machine learning.

7. FiveThirtyEight

- Description: Datasets behind stories published on FiveThirtyEight, a website that focuses on data-driven journalism.

- Website: https://data.fivethirtyeight.com/

- Popular Uses: Social science, political data analysis.

8. OpenML

- Description: An open platform for sharing and discovering datasets, algorithms, and machine learning experiments.

- Website: https://www.openml.org/

- Popular Uses: Benchmarking machine learning models, collaborative data science.

9. Academic Torrents

- Description: A distributed system for sharing academic datasets, which includes a range of machine learning and data science datasets.

- Website: https://academictorrents.com/

- Popular Uses: High-speed access to large datasets.

10. Stanford Large Network Dataset Collection (SNAP)

- Description: A collection of network datasets from Stanford, suitable for research in graph-based machine learning.

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

- Popular Uses: Graph-based machine learning, social networks analysis.

11. The World Bank Open Data

- Description: Offers free and open access to global development data, which can be useful for socio-economic and policy-driven machine learning models.

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

- Popular Uses: Economic modeling, policy analysis.

12. Microsoft Azure Open Datasets

- Description: Microsoft's open dataset repository focused on machine learning, including environmental, healthcare, and financial data.

- Website: https://azure.microsoft.com/en-us/services/open-datasets/

- Popular Uses: Cloud-based machine learning, industry-specific applications.

13. Papers with Code

- Description: Provides links to datasets used in recent machine learning research papers, along with the code implementations.

- Website: https://paperswithcode.com/datasets

- Popular Uses: Replicating state-of-the-art ML research, benchmarking.

14. Quandl

- Description: A platform that provides financial, economic, and alternative datasets, with a mix of free and premium data.

- Website: https://www.quandl.com/

- Popular Uses: Financial machine learning, algorithmic trading.

These resources provide diverse datasets across fields like finance, healthcare, social science, natural language processing, and more. Depending on your project, you can choose from these sources to get structured and unstructured data for your models.