Chapter 2: Sources of Data

Public datasets

- Kaggle Datasets: Kaggle is a popular online community for data scientists and machine learning enthusiasts. It hosts a large repository of public datasets, covering a wide range of domains, including:
  - Business
  - Computer vision
  - Natural language processing
  - Medical imaging
  - Physical sciences
  - Social sciences
- UCI Machine Learning Repository: The UCI Machine Learning Repository is another popular source of public datasets. It contains over 500 datasets, covering a wide range of machine learning tasks, including:
  - Classification
  - Regression
  - Clustering
  - Anomaly detection
- Google Cloud Public Datasets: Google Cloud Public Datasets is a service that provides access to over 100 public datasets, hosted on Google Cloud Storage. The datasets are organized by category, including:
  - Business
  - Government
  - Science
  - Education
- Microsoft Datasets: Microsoft Datasets is a service that provides access to a variety of public datasets, hosted on Azure Blob Storage. The datasets are organized by category, including:
  - Business
  - Government
  - Science
  - Open source
- AWS Public Datasets: AWS Public Datasets is a service that provides access to a variety of public datasets, hosted on Amazon S3. The datasets are organized by category, including:
  - Business
  - Government

    - Science
    - Open source

Other sources of public datasets

- Data.gov: Data.gov is a repository of over 200,000 datasets from the US government.
- OpenDataSoft: OpenDataSoft is a platform for publishing and sharing open data. It hosts over 200,000 datasets from around the world.
- World Bank Data: World Bank Data is a repository of data from the World Bank. It covers a wide range of economic, social, and environmental indicators.
- State of California Public Datasets: State of California provides access to data.

Private datasets

In addition to public datasets, there are also many private datasets that can be used for machine learning practice. These datasets may be collected by companies, organizations, or individuals. An example is Quandl, a platform for financial and economic data. It hosts over 10 million datasets from a variety of sources.

If you are interested in using private datasets for machine learning practice, you can try to reach out to the owners of the datasets to see if they would be willing to share them with you. You may also be able to find private datasets that are available for purchase.

Other sources of data

In addition to public and private datasets, there are also a number of other sources of data that can be used for machine learning practice. These include:

- Web APIs: Many websites and online services provide access to data through web APIs. For example, you could use the Google Maps API to access geospatial data, or the Twitter API to access social media data.
- Social media: Social media platforms, such as Twitter, Facebook, and Instagram, contain a wealth of data that can be used for machine learning practice. For example, you could use social media data to build models for sentiment analysis, image classification, or text generation.
- Sensor data: Sensors can be used to collect data from the physical world. For example, you could use temperature sensors to collect data on climate change, or motion sensors to collect data on human activity.

Tips for finding datasets

Here are a few tips for finding datasets for machine learning practice:

- Use a dataset search engine: There are a number of dataset search engines available online, such as the Google Dataset Search Engine and the AWS Public Datasets Catalog. These search engines can help you to find datasets that are relevant to your interests.
- Browse online repositories: There are a number of online repositories that host datasets for machine learning practice, such as Kaggle Datasets and the UCI Machine Learning Repository.
- Reach out to experts: If you are looking for a specific type of dataset, you can try to reach out to experts in the field to see if they can help you to find it.
- Collect your own data: If you cannot find a suitable dataset, you can always collect your own data. This can be done using sensors, web APIs, or by manually scraping data from the web.

Once you have found a dataset, it is important to clean and prepare it before using it for machine learning. This may involve removing outliers, filling in missing values, and converting the data into a format that is compatible with your machine learning algorithm.