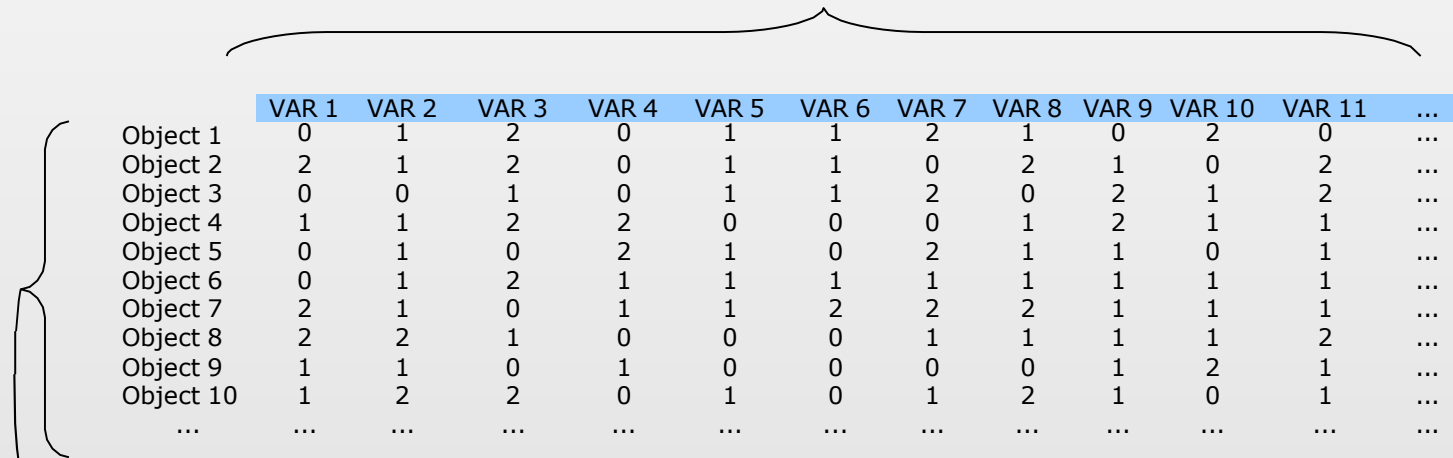


# How to get datasets for Machine Learning

# Glossary

- Data=a table (dataset, database, sample)

Variables (attributes, features) =  
measurements made on objects



	VAR 1	VAR 2	VAR 3	VAR 4	VAR 5	VAR 6	VAR 7	VAR 8	VAR 9	VAR 10	VAR 11	...
Object 1	0	1	2	0	1	1	2	1	0	2	0	...
Object 2	2	1	2	0	1	1	0	2	1	0	2	...
Object 3	0	0	1	0	1	1	2	0	2	1	2	...
Object 4	1	1	2	2	0	0	0	1	2	1	1	...
Object 5	0	1	0	2	1	0	2	1	1	0	1	...
Object 6	0	1	2	1	1	1	1	1	1	1	1	...
Object 7	2	1	0	1	1	2	2	2	1	1	1	...
Object 8	2	2	1	0	0	0	1	1	1	1	2	...
Object 9	1	1	0	1	0	0	0	0	1	2	1	...
Object 10	1	2	2	0	1	0	1	2	1	0	1	...
...	...	...	...	...	...	...	...	...	...	...	...	...

Objects (samples, observations,  
individuals, examples, patterns)

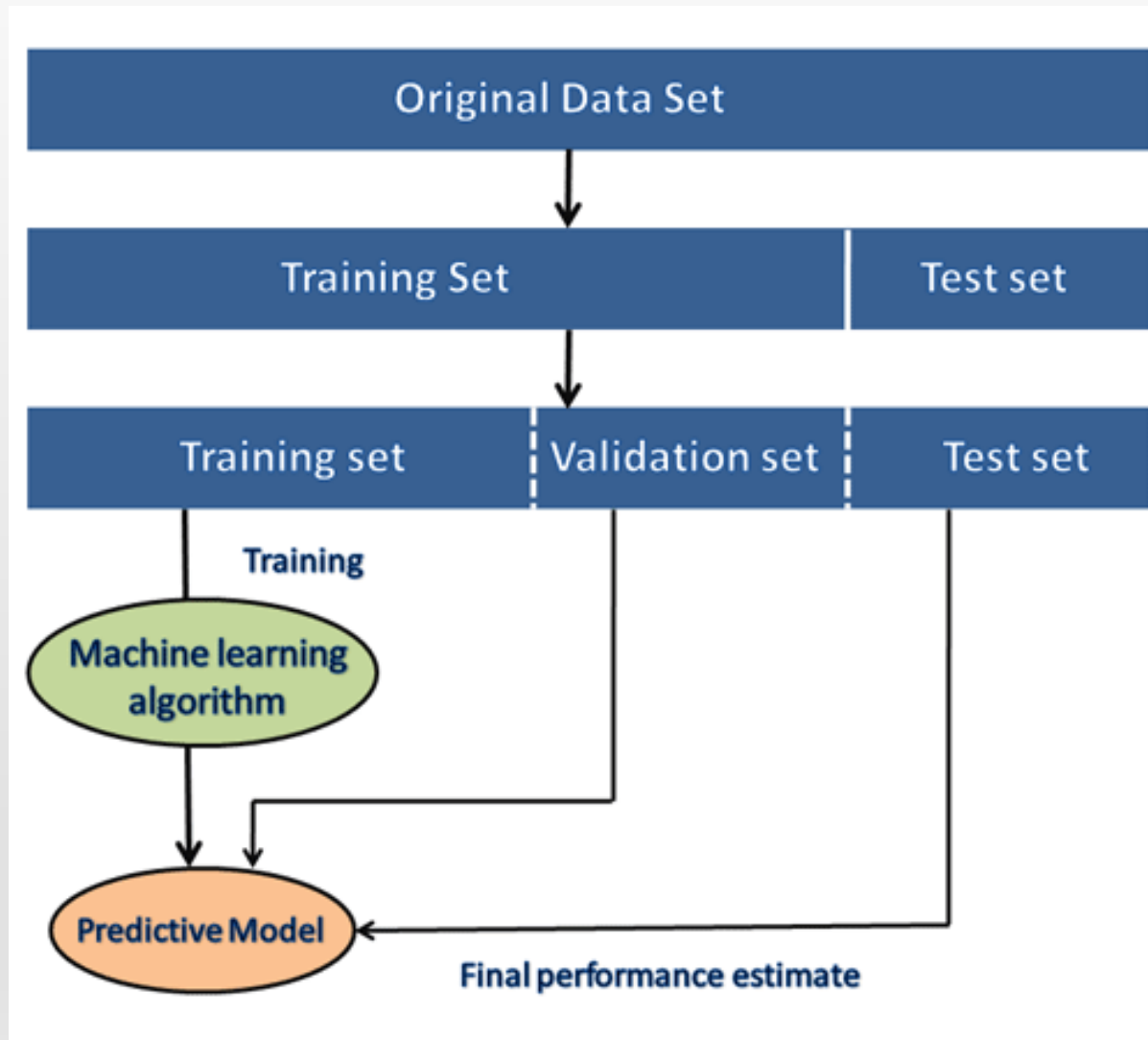
Dimension=number of variables  
Size=number of objects

- Objects: samples, patients, documents, images...  
Variables: genes, proteins, words, pixels...

# Types

- **Numerical data:**Such as house price, temperature, etc.
- **Categorical data:**Such as Yes/No, True/False, Blue/green, etc.
- **Ordinal data:**These data are similar to categorical data but can be measured on the basis of comparison.

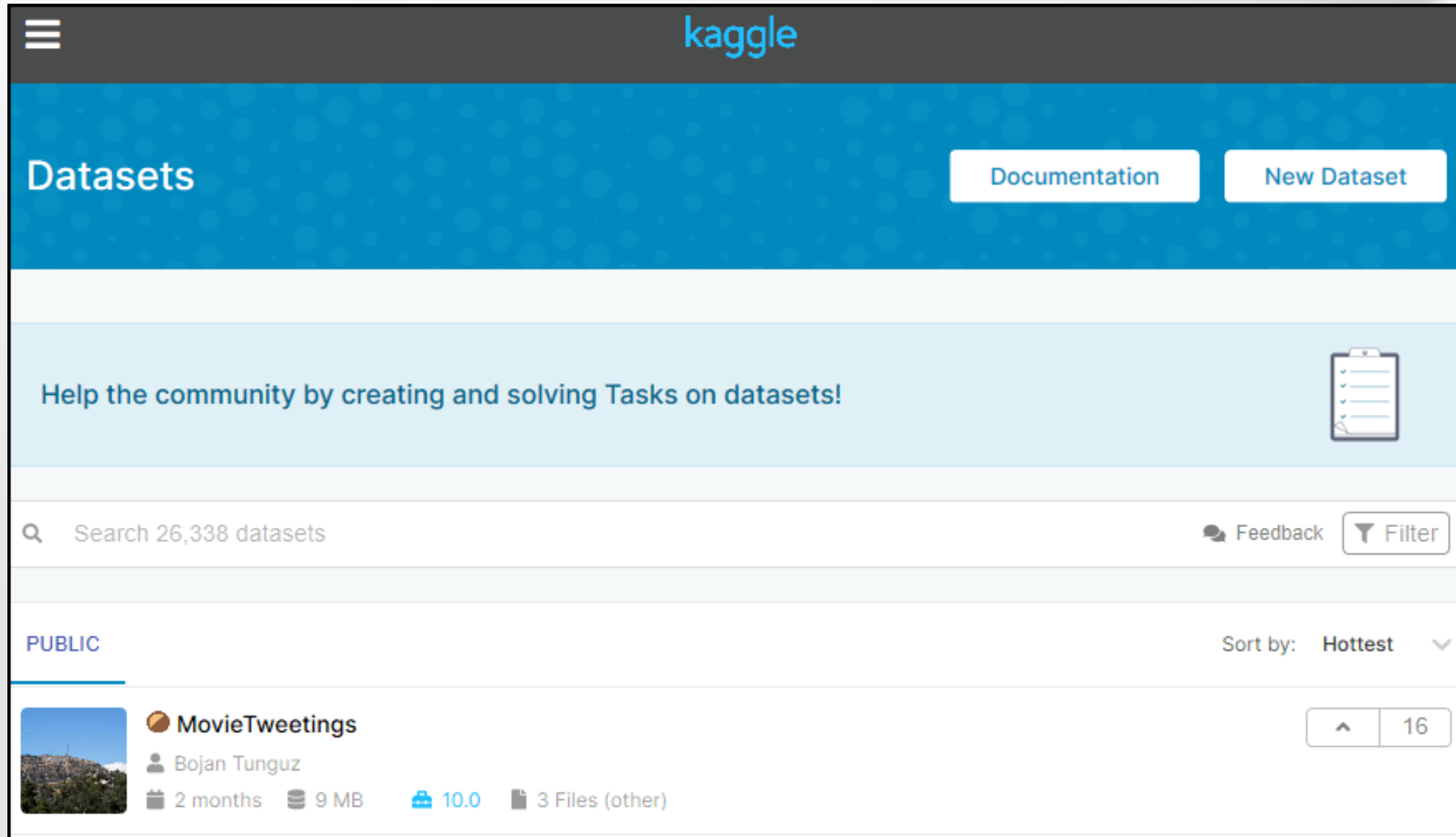
# Need of Dataset



- Training dataset
- Test Dataset

# Popular sources for Machine Learning datasets

# Kaggle Datasets



The link for the Kaggle dataset is <https://www.kaggle.com/datasets>.

# UCI Machine Learning Repository



[About](#) [Citation Policy](#) [Donate a Data Set](#) [Contact](#)






☒ Repository ☐ Web 

[View ALL Data Sets](#)

**Machine Learning Repository**  
Center for Machine Learning and Intelligent Systems

Browse Through: 488 Data Sets


[Table View](#) [List View](#)

Default Task	Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
<a href="#">Classification (360)</a> <a href="#">Regression (107)</a> <a href="#">Clustering (90)</a> <a href="#">Other (55)</a>	 <a href="#">Abalone</a>	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
<b>Attribute Type</b>	 <a href="#">Adult</a>	Multivariate	Classification	Categorical, Integer	48842	14	1996
<a href="#">Categorical (38)</a> <a href="#">Numerical (325)</a> <a href="#">Mixed (55)</a>	 <a href="#">Annealing</a>	Multivariate	Classification	Categorical, Integer, Real	798	38	
<b>Data Type</b>	 <a href="#">Anonymous Microsoft Web Data</a>		Recommender-Systems	Categorical	37711	294	1998
<a href="#">Multivariate (374)</a> <a href="#">Univariate (24)</a> <a href="#">Sequential (51)</a> <a href="#">Time-Series (99)</a> <a href="#">Text (55)</a> <a href="#">Domain-Theory (23)</a> <a href="#">Other (21)</a>	 <a href="#">Arrhythmia</a>	Multivariate	Classification	Categorical, Integer, Real	452	279	1998
<b>Area</b>							
<a href="#">Life Sciences (110)</a> <a href="#">Physical Sciences (52)</a> <a href="#">CS / Engineering (178)</a>							

The link for the UCI machine learning repository is <https://archive.ics.uci.edu/ml/index.php>

# Datasets via AWS

## Registry of Open Data on AWS



### About

This registry exists to help people discover and share datasets that are available via AWS resources. [Learn more about sharing data on AWS.](#)

See [all usage examples for datasets listed in this registry.](#)

See datasets from [Facebook Data for Good](#), [NOAA Big Data Project](#), and [Space Telescope Science Institute](#).

---

### Search datasets (currently 120 matching datasets)

---

### Add to this registry

If you want to add a dataset or example of how to use a dataset to this registry, please follow the instructions on the [Registry of Open Data on AWS GitHub page](#).

## Sentinel-2

[disaster response](#) [earth observation](#) [geospatial](#) [natural resource](#) [satellite imagery](#) [sustainability](#)

The [Sentinel-2 mission](#) is a land monitoring constellation of two satellites that provide high resolution optical imagery and provide continuity for the current SPOT and Landsat missions. The mission provides a global coverage of the Earth's land surface every 5 days, making the data of great use in on-going studies. L1C data are available from June 2015 globally. L2A data are available from April 2017 over wider Europe region and globally since December 2018.

[Details](#) →

### Usage examples

- [Sentinel Playground](#) by Sinergise
- [Learning Custom Scripts to Make Useful and Beautiful Satellite Images](#) by Monja Šebela
- [Sterling Geo Using Sentinel-2 on Amazon Web Services to Create NDVI](#) by Sterling Geo
- [FME Landsat-8/Sentinel-2 File Selector](#) by Safe Software

The link for the resource is <https://registry.opendata.aws/>



# Google's Dataset Search Engine

The screenshot shows the Google Dataset Search interface. At the top, the search bar contains the word "classification". Below the search bar, there are filters for "Updated Date", "Download Format", "Usage Rights", and "Free". The results section shows "100+ results found". On the left, a sidebar highlights the "Classification" dataset with a blue circle containing the letter "D". The main content area displays the "Classification" dataset details, including links to explore it at catalog.data.gov, Rally - Open Data Portal, and data.wu.ac.at. It also shows the dataset was updated on May 2, 2019, and is provided by Dashlink. A description of supervised learning is provided at the bottom.

Google Dataset Search

classification

Updated Date Download Format Usage Rights Free

100+ results found

**Classification**  
catalog.data.gov  
data.nasa.gov  
+1more  
Updated May 2, 2019

**Mushroom Classification**  
www.kaggle.com  
Updated Dec 1, 2016

**Question Classification**  
www.kaggle.com

**Classification**

Explore at catalog.data.gov

Explore at Rally - Open Data Portal

Explore at data.wu.ac.at

Dataset updated May 2, 2019

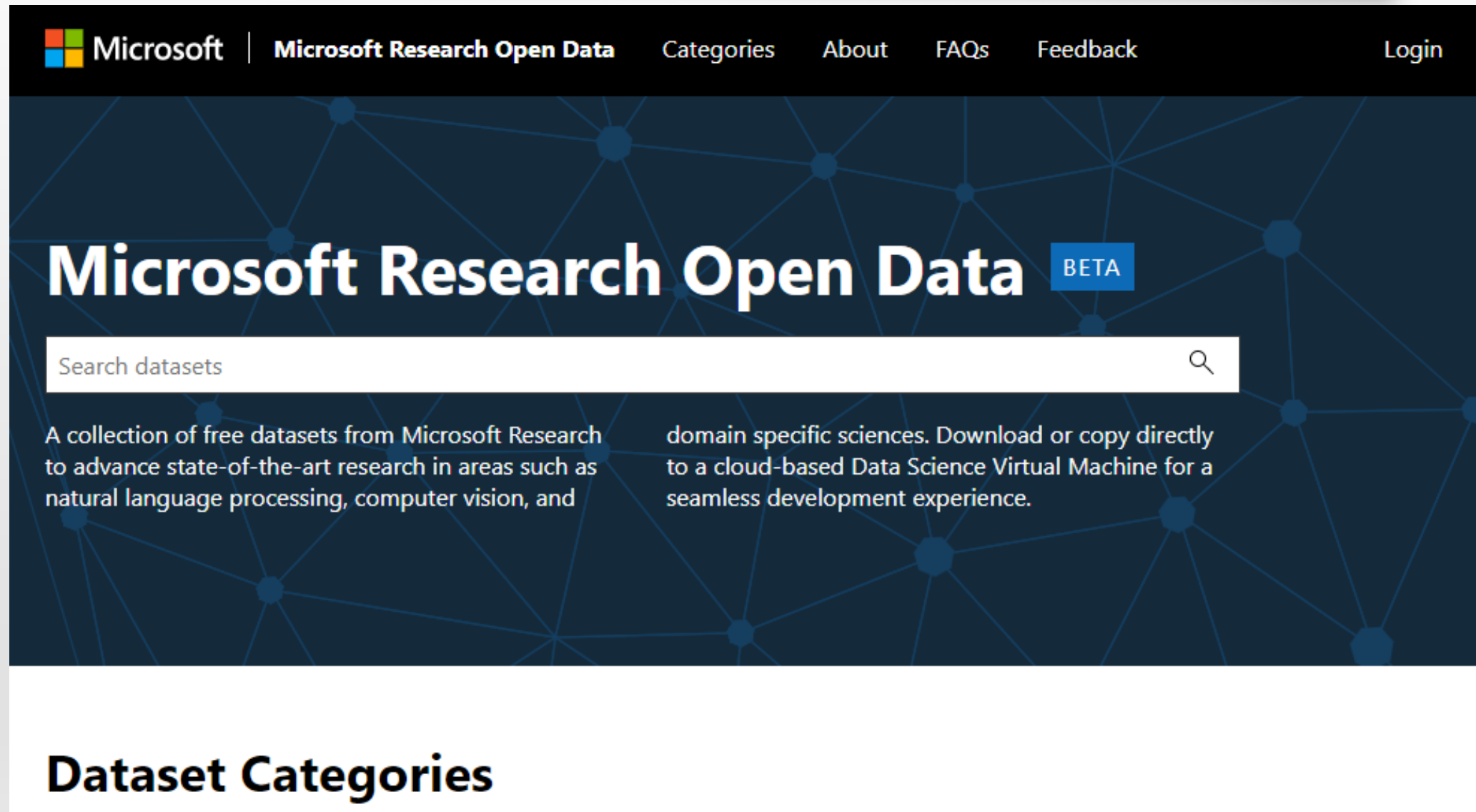
Dataset provided by  
Dashlink

**Description**

A supervised learning task involves constructing a mapping from an input data space (normally described by several features) to an output space. A set of training examples—examples with known output values—is used by a learning

The link for the Google dataset search engine is <https://toolbox.google.com/datasetsearch>

# Microsoft Datasets

The image is a screenshot of the Microsoft Research Open Data website. The top navigation bar is dark blue with the Microsoft logo and links for 'Microsoft Research Open Data', 'Categories', 'About', 'FAQs', 'Feedback', and 'Login'. The main header area has a dark blue background with a network diagram pattern. It features the title 'Microsoft Research Open Data' in large white text, followed by a blue 'BETA' badge. Below the title is a white search bar with the placeholder text 'Search datasets' and a magnifying glass icon. Two columns of text describe the datasets: 'A collection of free datasets from Microsoft Research to advance state-of-the-art research in areas such as natural language processing, computer vision, and' and 'domain specific sciences. Download or copy directly to a cloud-based Data Science Virtual Machine for a seamless development experience.' The bottom section of the screenshot has a white background with the heading 'Dataset Categories' in bold black text.

Microsoft | Microsoft Research Open Data Categories About FAQs Feedback Login

# Microsoft Research Open Data BETA

Search datasets

A collection of free datasets from Microsoft Research to advance state-of-the-art research in areas such as natural language processing, computer vision, and

domain specific sciences. Download or copy directly to a cloud-based Data Science Virtual Machine for a seamless development experience.

## Dataset Categories

The link to download or use the dataset from this resource is <https://msropendata.com/>.

# Awesome Public Datasets

## 🔗 Awesome Public Datasets



**NOTICE:** This repo is automatically generated by [apd-core](#). Please **DO NOT** modify this file directly. We have provided [a new way](#) to contribute to Awesome Public Datasets. The original PR entrance directly on repo is closed forever.

- 🟢 I am well.
- 🟡 ? Please fix me.

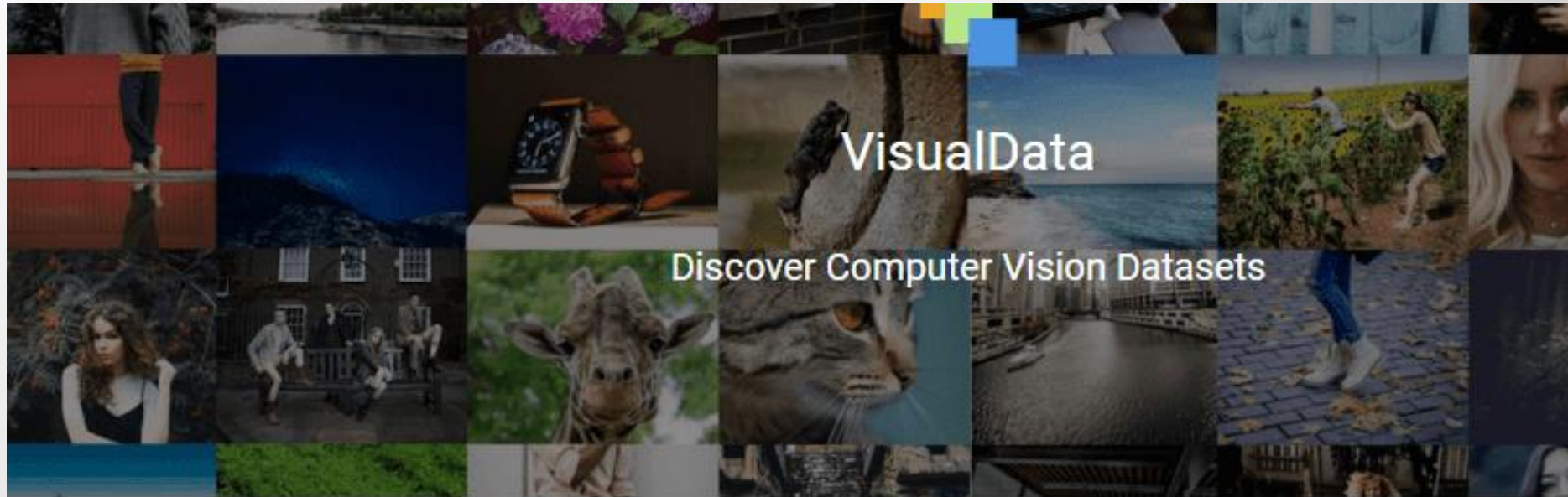
This list of [a topic-centric public data sources](#) in high quality. They are collected and tidied from blogs, answers, and user responses. Most of the data sets listed below are free, however, some are not. Other amazingly awesome lists can be found in [sindresorhus's awesome list](#).

### Table of Contents

- [Agriculture](#)
- [Biology](#)
- [Climate+Weather](#)
- [ComplexNetworks](#)
- [ComputerNetworks](#)

<https://github.com/awesomedata/awesome-public-datasets>.

# Computer Vision Datasets



Get notified for new dataset, code release and more

**Subscribe**



Select topics

Search



Sort by

Recently Added



The link for downloading the dataset from this source is <https://www.visualdata.io/>.

# Scikit-Learn Datasets

[Install](#) [User Guide](#) [API](#) [Examples](#) [More](#) 

[Prev](#) [Up](#) [Next](#)

**scikit-learn 0.22.1**  
[Other versions](#)

Please [cite us](#) if you use the software.

`sklearn.datasets.load_boston`  
Examples using  
`sklearn.datasets.load_boston`

## sklearn.datasets.load\_boston

```
sklearn.datasets.load_boston(return_X_y=False)
```

[\[source\]](#)

Load and return the boston house-prices dataset (regression).

Samples total	506
Dimensionality	13
Features	real, positive
Targets	real 5. - 50.

Read more in the [User Guide](#).

**Parameters:** **return\_X\_y : boolean, default=False.**  
If True, returns (data, target) instead of a Bunch object. See below for more information about the data and target object.  
*New in version 0.18.*

**Returns:** **data : Bunch**

The link to download datasets from this source is <https://scikit-learn.org/stable/datasets/index.html>.