

Repository :

Peng Lab: (<https://data.mendeley.com/datasets/btchxktzyw/2>)

Scopus Datasets: (<https://data.mendeley.com/datasets/btchxktzyw/2>)

University of California ML dataset: (<http://archive.ics.uci.edu/ml/index.php>)
559 data sets as a service to the machine learning community.

Life Science, Healthcare and Medical Datasets:

(<https://lionbridge.ai/datasets/18-free-life-sciences-medical-datasets-for-machine-learning/>)

Video datasets , Image Classification datasets large scale ML repository

: (<https://www.datasetlist.com/>)

Yahoo News Feed Dataset: (

<https://webscope.sandbox.yahoo.com/catalog.php?datatype=r&did=75&guccounter=1>)

Time Series Analysis datasets, Airports, Traffic analysis : < GEO Dataset >

(https://geodata.lib.berkeley.edu/?f%5Bdc_creator_sm%5D%5B%5D=Research+and+Innovative+Technology+Administration%27s+Bureau+of+Transportation+Statistics+%28RITA%2FBTS%29&f%5Bdc_rights_s%5D%5B%5D=Public&f%5Bdc_spatial_sm%5D%5B%5D=United+States)

Prediction Analysis :

(<https://blog.bigml.com/2013/02/28/data-data-data-thousands-of-public-data-sources/#comment-7538>)

Cognitive and NLP datasets : (

<https://analyticsindiamag.com/10-open-source-datasets-for-text-classification/>)

Charitable Donation Dataset :

(<http://kdd.ics.uci.edu/databases/kddcup98/kddcup98.html>)

Network intrusion dataset: (<http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>)

CAN Dataset: <https://ocslab.hksecurity.net/Datasets/CAN-intrusion-dataset>
(CAN is a car network controller hub. This is a time series dataset analysing network traffic.)

ELVIRA Biomedical Data Set Repository: <http://leo.ugr.es/elvira/DBCRepository/>
(gene expression data, protein profiling data and genomic sequence data)

1000Genome dataset: <https://registry.opendata.aws/1000-genomes/>
(human genetic variation)

phenotype-associated cell dataset :
<https://www.kaggle.com/kmader/cellcnn-learning-disease-associated-cell-subsets>

List of large datasets:
<https://www.quora.com/Data/Where-can-I-find-large-datasets-open-to-the-public>
Or
<https://github.com/awesomedata/awesome-public-datasets>

Stanford large network datasets:
<http://snap.stanford.edu/data/index.html>

Colombia university:
<https://www.ee.columbia.edu/~cylin/course/bigdata/getdatasetinfo.html>