miTu skillologies

Datasets and Data Repositories

Tushar B. Kute,
http://tusharkute.com

tusharkute .com

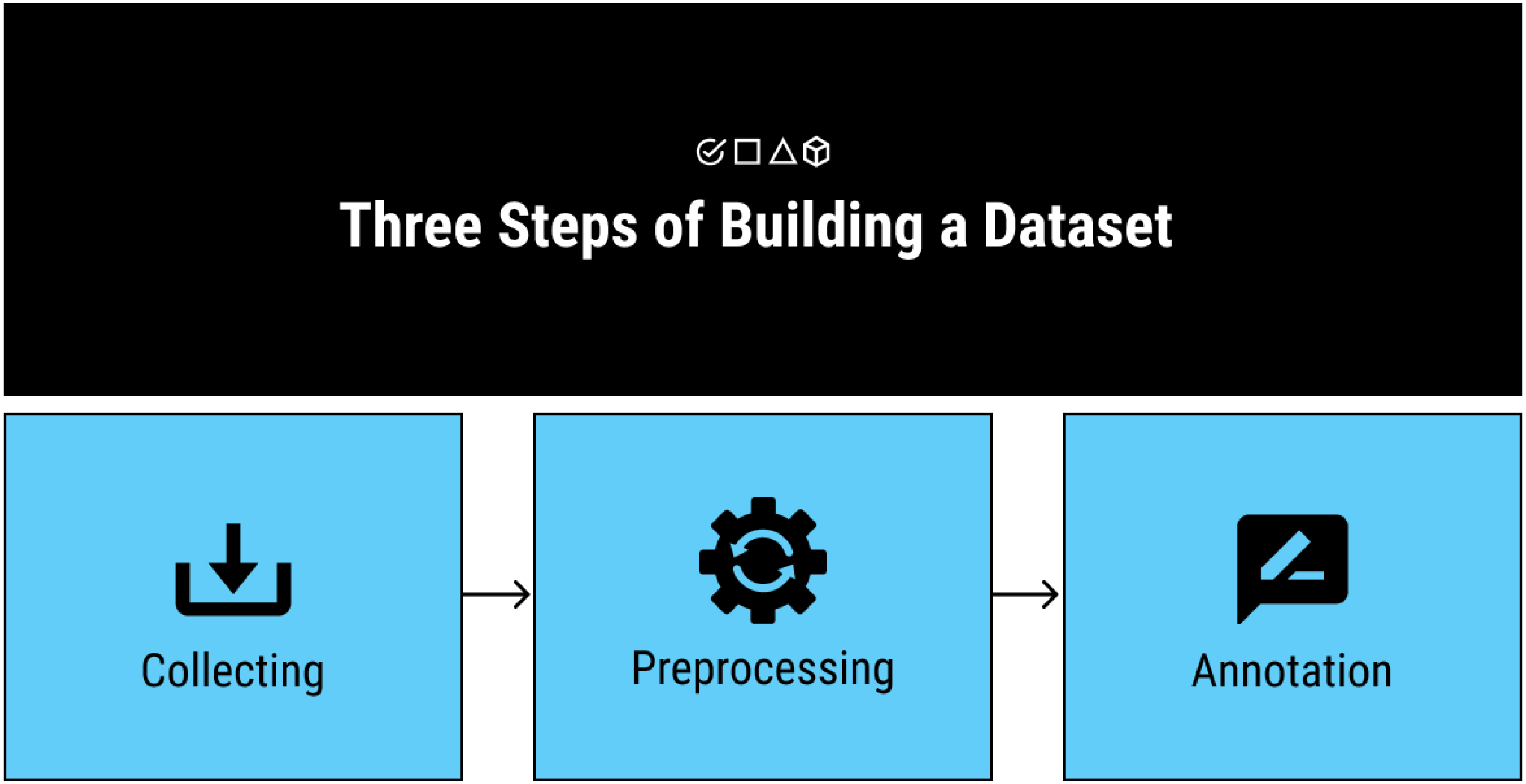
Dataset

- Oxford Dictionary defines a dataset as “a collection of data that is treated as a single unit by a computer”.
- This means that a dataset contains a lot of separate pieces of data but can be used to train an algorithm with the goal of finding predictable patterns inside the whole dataset.

Dataset

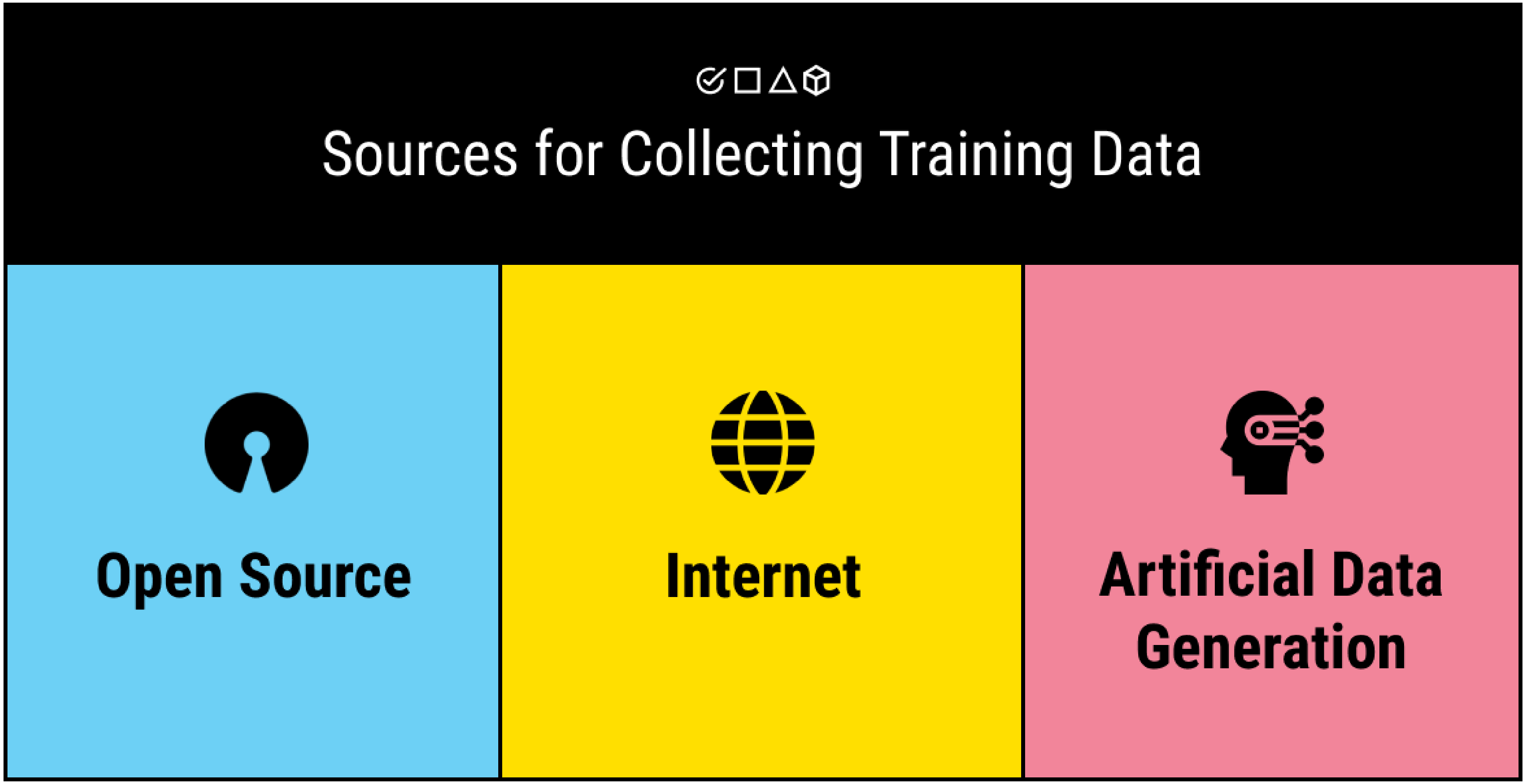
- Dataset contains a lot of separate pieces of data but can be used to train an algorithm with the goal of finding predictable patterns inside the whole dataset.
- Data is an essential component of any AI model and, basically, the sole reason for the spike in popularity of machine learning that we witness today.
- Due to the availability of data, scalable ML algorithms became viable as actual products that can bring value to a business, rather than being a by-product of its main processes.

Dataset



https://labelyourdata.com

Dataset



Sources for Collecting Training Data

Open Source

Internet

Artificial Data Generation

https://labelyourdata.com

Data Repositories

- A data repository is also known as a data library or data archive. This is a general term to refer to a data set isolated to be mined for data reporting and analysis.
- The data repository is a large database infrastructure — several databases — that collect, manage, and store data sets for data analysis, sharing and reporting.

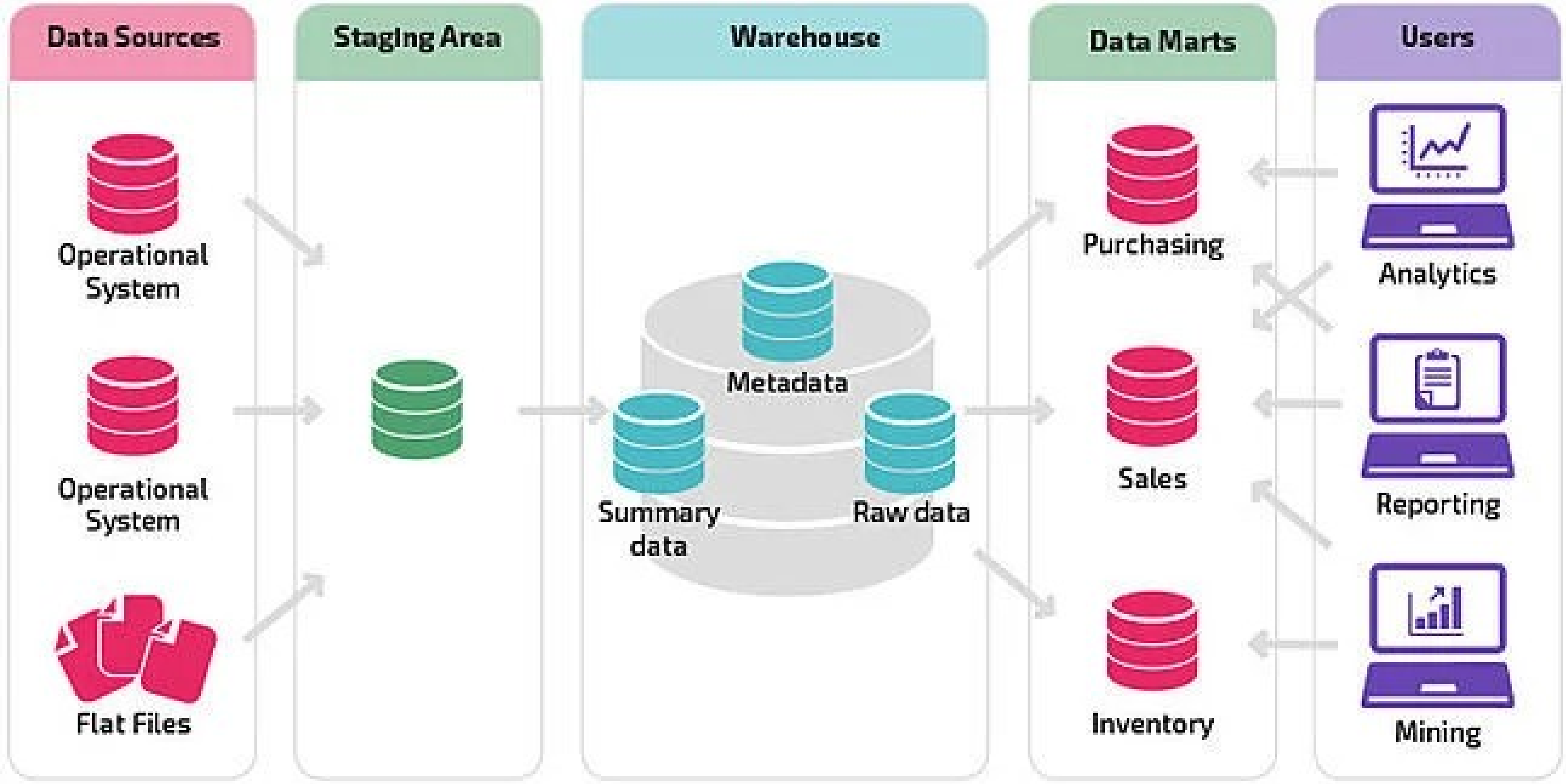
Data Repositories: Examples

- The term data repository can be used to describe several ways to collect and store data:
  - A data warehouse is a large data repository that aggregates data usually from multiple sources or segments of a business, without the data being necessarily related.
  - A data lake is a large data repository that stores unstructured data that is classified and tagged with metadata.

Data Repositories: Examples

- Data marts are subsets of the data repository. These data marts are more targeted to what the data user needs and easier to use.
- Data marts also are more secure because they limit authorized users to isolated data sets. Those users cannot access all the data in the data repository.
- Metadata repositories store data about data and databases. The metadata explains where the data source, how it was captured, and what it represents.

Data Repositories



Ref: hubspot.com

Data Repositories: Why?

- There is value to storing and analyzing data. Businesses can make decisions based upon more than anecdote and instinct. However, using data repositories as part of data management is another level of investment that can improve business decisions, such as:
  - Isolation allows for easier and faster data reporting or analysis because the data is clustered together.
  - Database administrators have easier time tracking problems because data repositories are compartmentalized
  - Data is preserved and archived.

Data Repositories: Best Practices

- Enlist a high-level business champion to engage all stakeholders during the project development and during its use. This is not a developer but someone who can work across departments, engaging people who will use the data repository.
- The data repository will need to grow. Treat it as an ongoing system.
- Hire experts who can build and maintain the data repository that is needed.

Data Repositories: Best Practices

- In the beginning, keep the scope of the data repository modest. Collect smaller sets of data and restrict the number of data subjects. Build upon the complexity as the data users learn the system and discover return on investment.
- Use Extract-Transformation-Load (ETL) tools to migrate data to the data repository. These tools ensure data quality in the transfer.
- Build your data warehouse first, then build the data marts.

Data Repositories: Best Practices

- Decide how often the data warehouse will load new data. This often depends on the volume of data.
- Metadata is necessary for quality data analysis and reporting.
- Data users need to have access to education and support.
- The data repository will need to evolve. The types of data it collects and the uses for it will change. Therefore, having flexible plans will allow for changes in technology.

Ref: Chris Brook, editor of Data Insider.

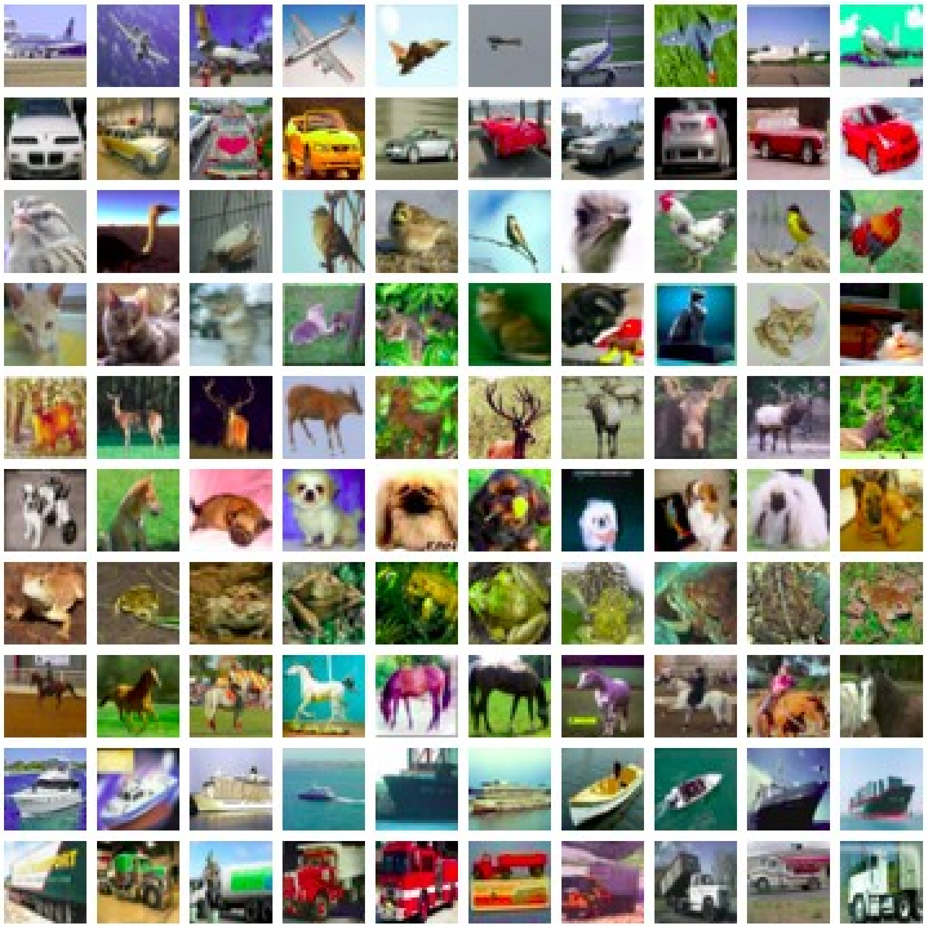
Computer Vision

- Computer vision is a field of artificial intelligence (AI) that enables computers and systems to derive meaningful information from digital images, videos and other visual inputs — and take actions or make recommendations based on that information.
- If AI enables computers to think, computer vision enables them to see, observe and understand.
- Computer vision works much the same as human vision, except humans have a head start.

Computer Vision Datasets - CIFAR-10

- http://www.cs.toronto.edu/~kriz/cifar.html
- CIFAR-10 is a popular computer-vision dataset collected by Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton.
- This dataset is used for object recognition and it consists of 60,000 32×32 colour images in 10 classes, with 6,000 images per class.
- It is divided into five training batches and one test batch, each with 10,000 images which means there are 50,000 training images and 10,000 test images.

Computer Vision – CIFAR-10



Computer Vision Dataset - Cityscapes

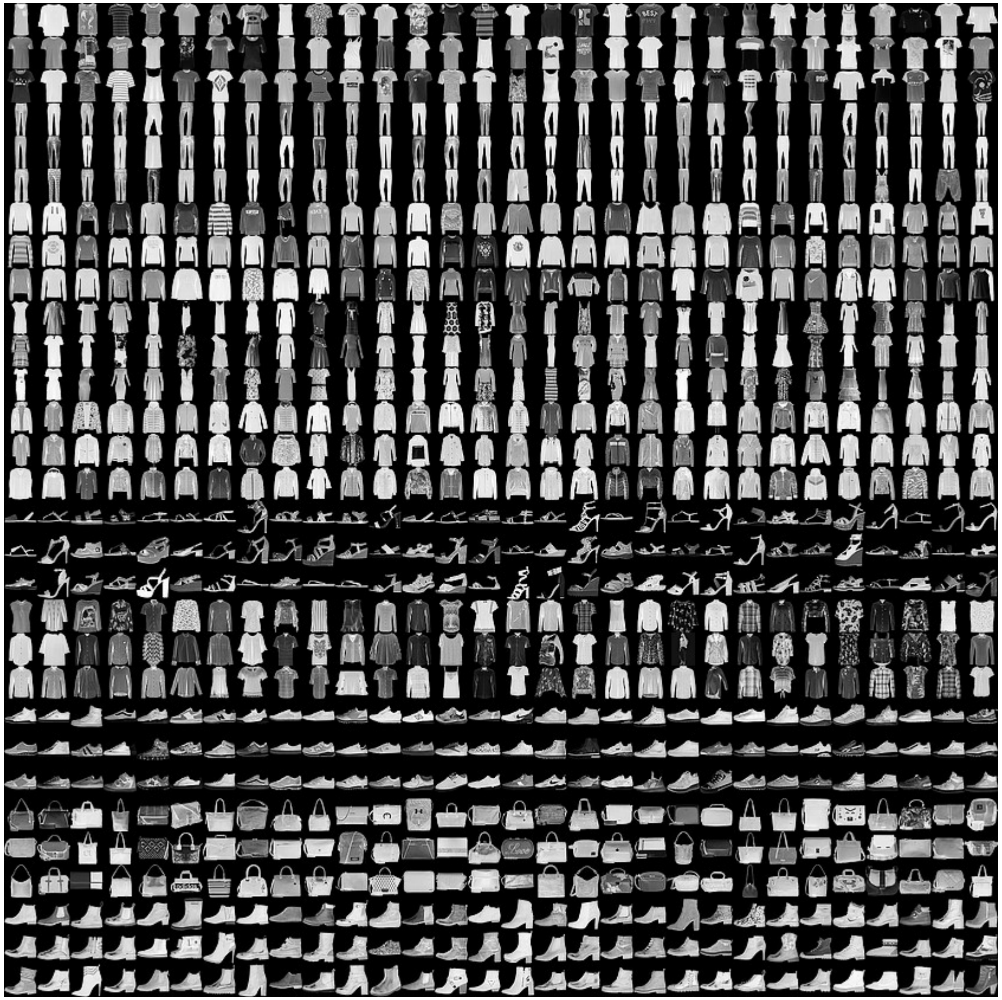
- Cityscapes is an open-sourced large-scale dataset for Computer Vision projects which contains a diverse set of stereo video sequences recorded in street scenes from 50 different cities.
- It includes high-quality pixel-level annotations of 5,000 frames in addition to a larger set of 20,000 weakly annotated frames.
- This dataset is mainly used for training deep neural networks and assessing the performance of vision algorithms for major tasks of semantic urban scene understanding.
- https://www.cityscapes-dataset.com

Computer Vision Dataset - Cityscapes

Computer Vision Dataset - Fashion MNIST

- Fashion-MNIST is an image dataset for Computer Vision which consists of a training set of 60,000 examples and a test set of 10,000 examples.
- In this dataset, each example is a 28×28 grayscale image, associated with a label from 10 classes.
- There is an automatic benchmarking system based on Scikit-learn that covers 129 classifiers with different parameters.
- https://github.com/zalandoresearch/fashion-mnist

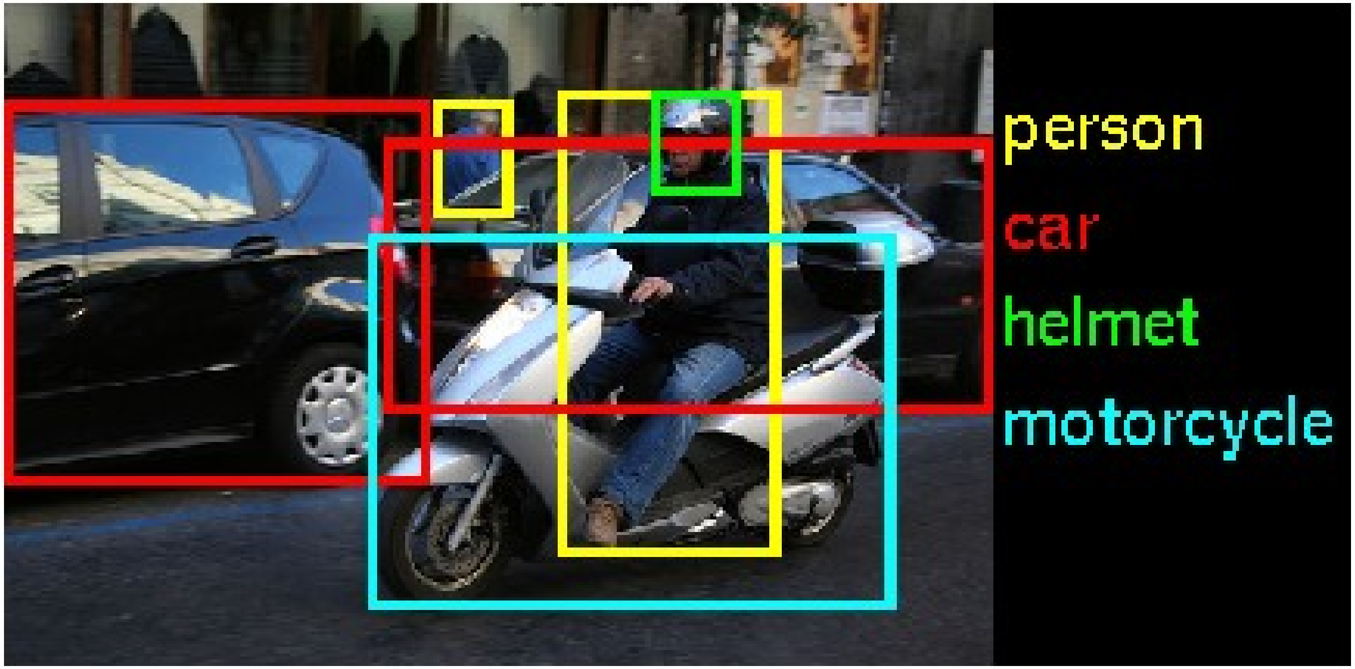
Computer Vision Dataset - Fashion MNIST



Computer Vision Dataset - ImageNet

- One of the popular datasets for Computer Vision projects, ImageNet provides an accessible image database which is organised according to the WordNet hierarchy.
- There are more than 100,000 synsets in WordNet where ImageNet provides an average of 1,000 images to illustrate each synset in the WordNet.
- It offers tens of millions of cleanly sorted images for most of the concepts in the WordNet hierarchy.
- http://www.image-net.org

Computer Vision Dataset - ImageNet



Computer Vision Dataset - IMDB-Wiki

- IMDB-Wiki dataset is one of the largest and open-sourced datasets of face images with gender and age labels for training.
- There is a total of 523,051 face images in this dataset where 460,723 face images are obtained from 20,284 celebrities from IMDB and 62,328 from Wikipedia.
- https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki

Computer Vision Dataset - IMDB-Wiki

IMDb

460,723 images

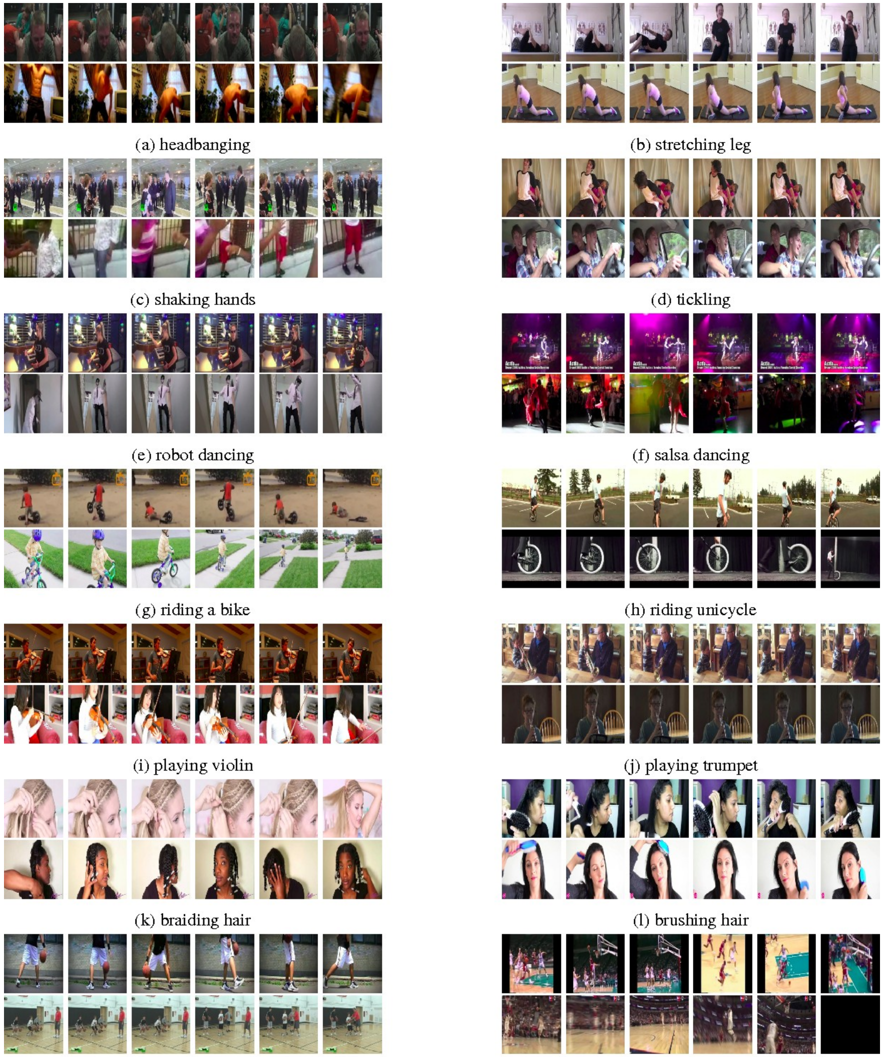
Wikipedia

62,328 images

Computer Vision Dataset - Kinetics-700

- Kinetics-700 is a large-scale, high-quality dataset of YouTube video URLs which include a diverse range of human-focused actions.
- The dataset consists of approximately 650,000 video clips and covers 700 human action classes with at least 600 video clips for each action class.
- Here, each clip lasts around 10 seconds and is labelled with a single class.
- https://deepmind.com/research/open-source/open-source-datasets/kinetics/

Computer Vision Dataset - Kinetics-700



(a) headbanging

(b) stretching leg

(c) shaking hands

(d) tickling

(e) robot dancing

(f) salsa dancing

(g) riding a bike

(h) riding unicycle

(i) playing violin

(j) playing trumpet

(k) braiding hair

(l) brushing hair

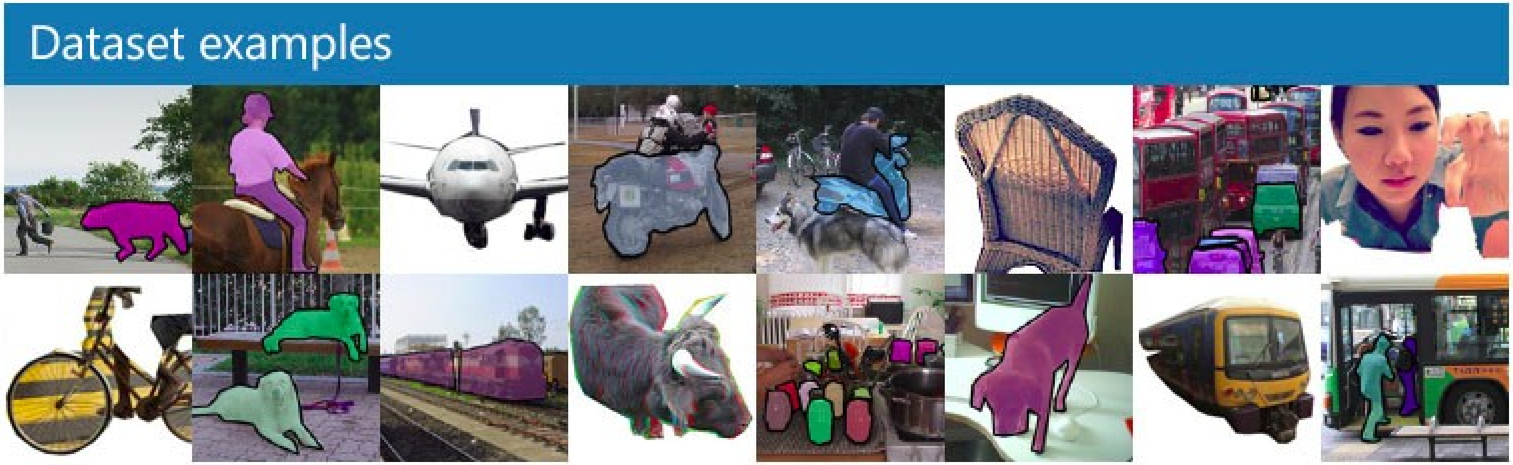
(m) dribbling basketball

(n) dunking basketball

Computer Vision Dataset - COCO

- COCO or Common Objects in COntext is large-scale object detection, segmentation, and captioning dataset.
- The dataset contains photos of 91 objects types which is easily recognisable and has a total of 2.5 million labelled instances in 328k images.
- http://cocodataset.org

Computer Vision Dataset - COCO

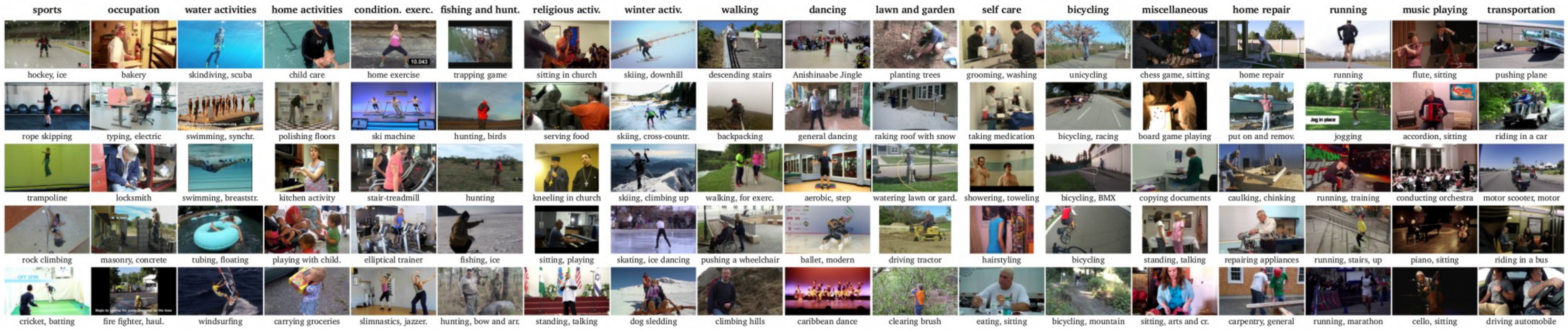


Dataset examples

Computer Vision Dataset - MPII Human Pose

- MPII Human Pose dataset is used for evaluation of articulated human pose estimation.
- The dataset includes around 25K images containing over 40K people with annotated body joints. Here, each image is extracted from a YouTube video and provided with preceding ann following un-annotated frames.
- Overall the dataset covers 410 human activities and each image is provided with an activity label.
- http://human-pose.mpi-inf.mpg.de

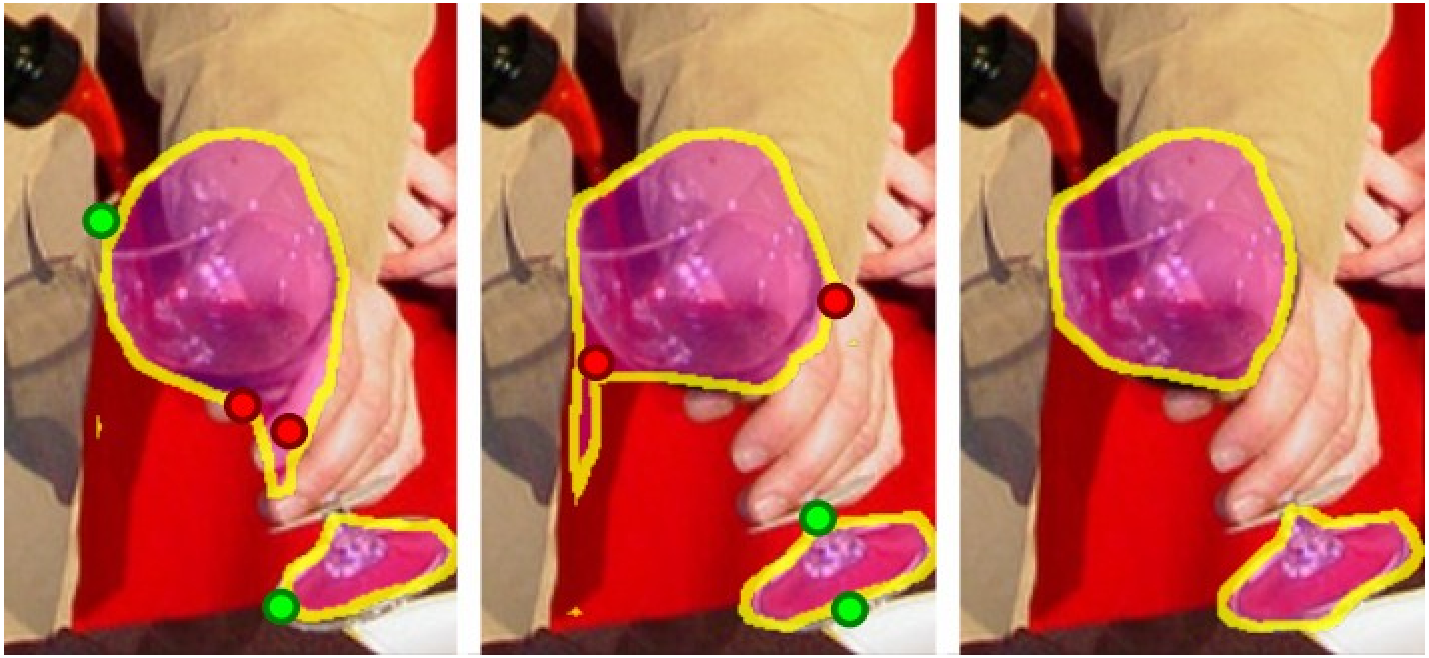
Computer Vision Dataset - MPII Human Pose



Computer Vision Dataset - Open Images

- This Open Images dataset is one of the largest existing datasets with object location annotations.
- It consists of around 9 million images annotated with image-level labels, object bounding boxes, object segmentation masks, and visual relationships.
- The dataset contains a total of 16 million bounding boxes for 600 object classes on 1.9 million images.
- https://storage.googleapis.com/openimages/web/factsfigures.html

Computer Vision Dataset - Open Images



The 20BN-something-something

- The 20BN-Something-Something dataset is a large collection of densely-labelled video clips that show humans performing pre-defined basic actions with everyday objects.
- It was created by a large number of crowd workers which allows ML models to develop a fine-grained understanding of basic actions that occur in the physical world.
- The total number of videos includes 220,847 where 168,913 is training set, 24,777 is validation set and 27,157 is the test set.
- https://20bn.com/datasets/something-something/v2

The 20BN-something-something

https://analyticsindiamag.com

Sentiment Analysis

- Sentiment analysis (or opinion mining) is a natural language processing technique used to determine whether data is positive, negative or neutral.
- Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs.

https://monkeylearn.com

Sentiment Analysis

- Fine-grained Sentiment Analysis
- If polarity precision is important to your business, you might consider expanding your polarity categories to include:
  - Very positive
  - Positive
  - Neutral
  - Negative
  - Very negative
- This is usually referred to as fine-grained sentiment analysis, and could be used to interpret 5-star ratings in a review, for example:
  - Very Positive = 5 stars
  - Very Negative = 1 star

https://monkeylearn.com

Sentiment Dataset – Amazon

- Amazon product data is a subset of a large 142.8 million Amazon review dataset that was made available by Stanford professor, Julian McAuley.
- This sentiment analysis dataset contains reviews from May 1996 to July 2014. The dataset reviews include ratings, text, helpfull votes, product description, category information, price, brand, and image features.
- http://jmcauley.ucsd.edu/data/amazon/

https://analyticsindiamag.com

Sentiment Dataset – Stanford

- This dataset contains just over 10,000 pieces of Stanford data from HTML files of Rotten Tomatoes.
- The sentiments are rated between 1 and 25, where one is the most negative and 25 is the most positive.
- The deep learning model by Stanford has been built on the representation of sentences based on the sentence structure instead just giving points based on the positive and negative words.
- https://nlp.stanford.edu/sentiment/code.html

Multi-Domain Sentiment Dataset

- This dataset contains positive and negative files for thousands of Amazon products.
- Although the reviews are for older products, this data set is excellent to use. The data derives from the Department of Computer Science at John Hopkins University.
- The reviews contain ratings from 1 to 5 stars that can be converted to binary as needed.
- http://www.cs.jhu.edu/~mdredze/datasets/sentiment/unprocessed.tar.gz

IMDB Movie Reviews Dataset

- This large movie dataset contains a collection of about 50,000 movie reviews from IMDB.
- In this dataset, only highly polarised reviews are being considered.
- The positive and negative reviews are even in number; however, the negative review has a score of ≤ 4 out of 10, and the positive review has a score of ≥ 7 out of 10.
- https://www.kaggle.com/iarunava/imdb-movie-reviews-dataset

IMDB Movie Reviews Dataset

- Sentiment140 is used to discover the sentiment of a brand or product or even a topic on the social media platform Twitter.
- Rather than working on keywords-based approach, which leverages high precision for lower recall, Sentiment140 works with classifiers built from machine learning algorithms.
- The Sentiment140 uses classification results for individual tweets along with the traditional surface that aggregated metrics. The Sentiment140 is used for brand management, polling, and planning a purchase.
- http://help.sentiment140.com/for-students

Twitter US Airline Sentiment

- This sentiment analysis dataset contains tweets since Feb 2015 about each of the major US airline. Each tweet is classified either positive, negative or neutral.
- The included features including Twitter ID, sentiment confidence score, sentiments, negative reasons, airline name, retweet count, name, tweet text, tweet coordinates, date and time of the tweet, and the location of the tweet.
- https://www.kaggle.com/crowdflower/twitter-airline-sentiment

Paper Reviews Data Set

- Paper Reviews Data Set contains reviews from English and Spanish languages on computing and informatics conferences.
- The algorithm used will predict the opinions of academic paper reviews. Most of the dataset for the sentiment analysis of this type is sent in Spanish. It has a total of instances of N=405 evaluated with a 5-point scale, -2: very negative, -1: neutral, 1: positive, 2: very positive.
- The distribution of the scores is uniform, and there exists a difference between the way the paper is evaluated and the review written by the original reviewer.

Sentiment Lexicons For 81 Languages

- Sentiment Lexicons for 81 Languages contains languages from Afrikaans to Yiddish. This data includes both positive and negative sentiment lexicons for a total of 81 languages.
- These lexica were generated via graph propagation for the sentiment analysis based on a knowledge graph which is a graphical representation of real-world objects and the relationship between them.
- The general idea is that words closely linked on a knowledge graph may have similar sentiment polarities. The sentiments were built based on English sentiment lexicons.

Lexicoder Sentiment Dictionary

- This dataset for the sentiment analysis is designed to be used within the Lexicoder, which performs the content analysis.
- This dictionary consists of 2,858 negative sentiment words and 1,709 positive sentiment words.
- In addition to that, 2,860 negations of negative and 1,721 positive words are also included.
- Anyone willing to test this is advised by the developers to subtract negated positive words from positive counts and subtract the negated negative words from the negative count.

Opin-Rank Review Dataset

- Opin-Rank Review Dataset contains full reviews on cars and hotels. This data set includes about 2,59,000 hotel reviews and 42,230 car reviews collected from TripAdvisor and Edmunds, respectively.
- The car dataset has the models from 2007, 2008, 2009 and has about 140-250 cars from each year. The fields include dates, favourites, author names, and full review in text.
- The dataset contains information from 10 different cities which include Dubai, Beijing, Las Vegas, San Fransisco, etc. There are reviews of about 80-700 hotels from each city. The fields include review, date, title and full-textual review.

NLP

- Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken and written -- referred to as natural language. It is a component of artificial intelligence (AI).
- NLP enables computers to understand natural language as humans do. Whether the language is spoken or written, natural language processing uses artificial intelligence to take real-world input, process it, and make sense of it in a way a computer can understand.

Self Driving Car

- nuScenes is a public large-scale dataset for autonomous driving. It enables researchers to study challenging urban driving situations using the full sensor suite of a real self-driving car.
- https://www.nuscenes.org/

Self Driving Car

- 93k video clips of 6s each (150h of driving)
- 93k annotated and 1.1M un-annotated images
- Two diverse cities: Boston and Singapore
- The same proven sensor suite as in nuScenes
- Images mined for diversity
- 800k annotated foreground objects with 2d bounding boxes and instance masks
- 100k 2d semantic segmentation masks for background classes
- Attributes such as rider, pose, activity, emergency lights and flying
- Free to use for non-commercial use

Clinical Dataset

- Clinical data is a staple resource for most health and medical research. Clinical data is either collected during the course of ongoing patient care or as part of a formal clinical trial program. Clinical data falls into six major types:
  - Electronic health records
  - Administrative data
  - Claims data
  - Patient / Disease registries
  - Health surveys
  - Clinical trials data
- https://guides.lib.uw.edu/hsl/data/findclin

Audio Dataset

- Audioset — a large scale dataset that consists of an expanding ontology of 632 audio event classes and a collection of 2,084,320 human-labeled 10-second sound clips drawn from YouTube videos.
- Link: https://research.google.com/audioset/index.html

Open Access Datasets

- In simple terms, Open Data means the kind of data which is open for anyone and everyone for access, modification, reuse, and sharing.
- Open Data derives its base from various “open movements” such as open source, open hardware, open government, open science etc.
- Governments, independent organizations, and agencies have come forward to open the floodgates of data to create more and more open data for free and easy access.

Open Access Datasets

- World Bank Open Data. ...
- WHO (World Health Organization) — Open data repository. ...
- Google Public Data Explorer. ...
- Registry of Open Data on AWS (RODA) ...
- European Union Open Data Portal. ...
- FiveThirtyEight. ...
- U.S. Census Bureau. ...
- Data.gov.

Google Dataset Search

- Google Dataset Search is a search engine from Google that helps researchers locate online data that is freely available for use.
- The company launched the service on September 5, 2018, and stated that the product was targeted at scientists and data journalists. The service was out of beta as of January 23rd, 2020.
- Google Dataset Search complements Google Scholar, the company's search engine for academic studies and reports
- https://datasetsearch.research.google.com/

Kaggle

- Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.
- https://www.kaggle.com

UCI ML Repository

- The UCI Machine Learning Repository is a database of machine learning problems that you can access for free.
- It is hosted and maintained by the Center for Machine Learning and Intelligent Systems at the University of California, Irvine. It was originally created by David Aha as a graduate student at UC Irvine.
- For more than 25 years it has been the go-to place for machine learning researchers and machine learning practitioners that need a dataset.
- Each dataset gets its own webpage that lists all the details known about it including any relevant publications that investigate it. The datasets themselves can be downloaded as ASCII files, often the useful CSV format.

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

Python Libraries

- TensorFlow Datasets — a collection of ready-to-use datasets. TensorFlow Datasets is a collection of datasets ready to use, with TensorFlow or other Python ML frameworks, such as Jax. All datasets are exposed as tf.data.Datasets , enabling easy-to-use and high-performance input pipelines. To get started see the guide and our list of datasets.
- Sklearn — Machine Learning package. This package also features helpers to fetch larger datasets commonly used by the machine learning community to benchmark algorithms on data that comes from the ‘real world’.

Python Libraries

- nltk: Natural Language Tool Kit package. Practical work in Natural Language Processing typically uses large bodies of linguistic data, or corpora.
- statsmodel: Statistical Model package. Provides data sets (i.e. data and meta-data) for use in examples, tutorials, model testing, etc.

Python Libraries

- pydataset — Dataset for educational purposes, mainly. It tries to help those approaching Data Science in Python for the first time, who must deal with common (and time-consuming) data preparation tasks.
- seaborn: Data Visualisation package where you can also load an example dataset from the online repository (requires internet).

Thank you

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com