Machine learning is a growing technology which enables computers to learn automatically from past data.

Machine learning uses various algorithms for **building mathematical models and making predictions using historical data or information**.

Currently, it is being used for various tasks such as **image recognition**, **speech recognition**, **email filtering**, **Facebook auto-tagging**, **recommender system**, and many more.

Jagendra Singh

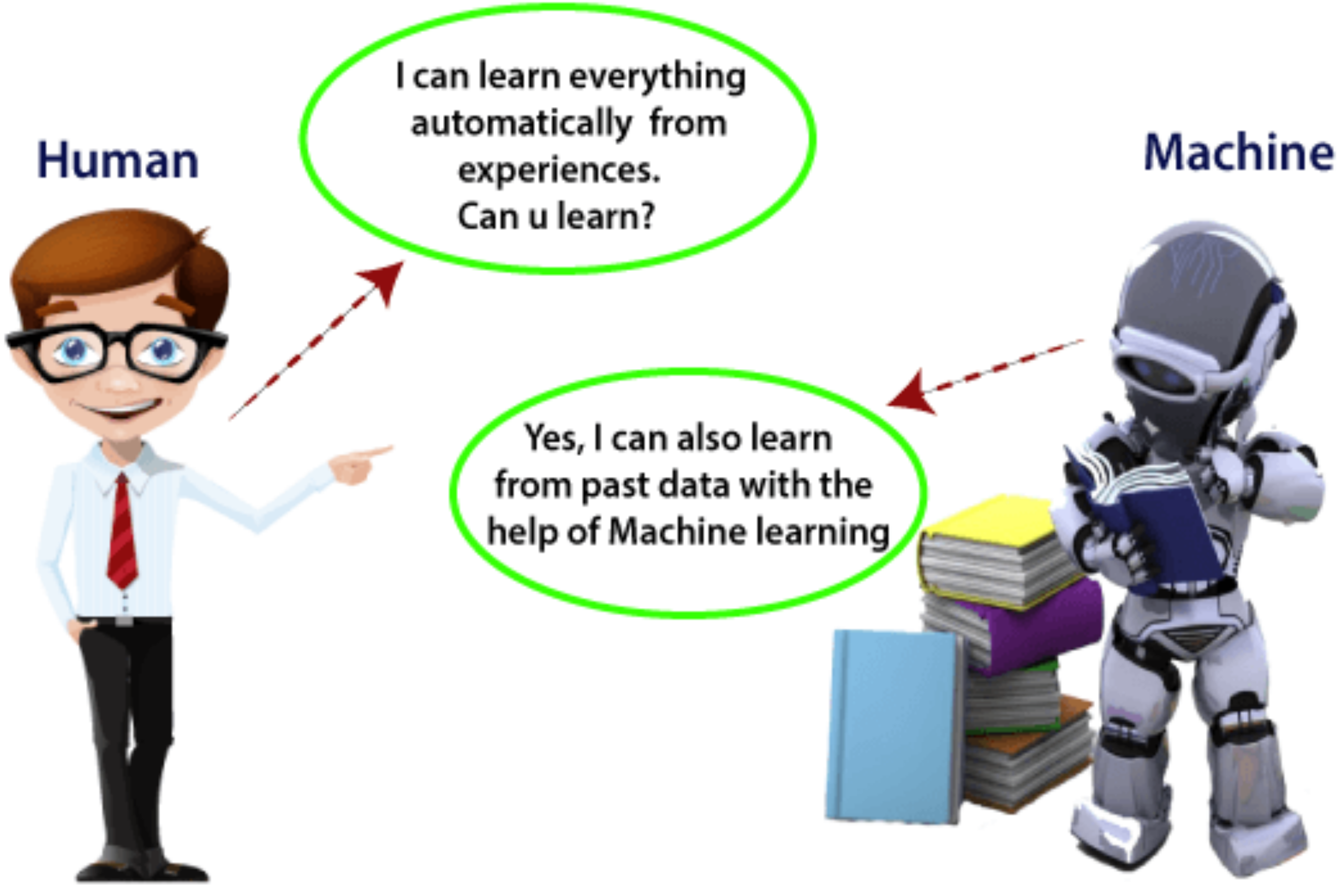
WHAT IS MACHINE LEARNING

In the real world, we are surrounded by humans who can learn everything from their experiences with their learning capability, and we have computers or machines which work on our instructions.

But can a machine also learn from experiences or past data like a human does? So here comes the role of Machine Learning.

Machine Learning is said as a subset of artificial intelligence that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own.

WHAT IS MACHINE LEARNING


Human
I can learn everything automatically from experiences. Can u learn?
Machine
Yes, I can also learn from past data with the help of Machine learning

WHAT IS MACHINE LEARNING

The term machine learning was first introduced by Arthur Samuel in 1959. We can define it in a summarized way as:

Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things

With the help of sample historical data, which is known as training data, machine learning algorithms build a mathematical model that helps in making predictions or decisions

Machine learning brings computer science and statistics together for creating predictive models.

Machine learning constructs or uses the algorithms that learn from historical data.

HOW DOES MACHINE LEARNING WORK

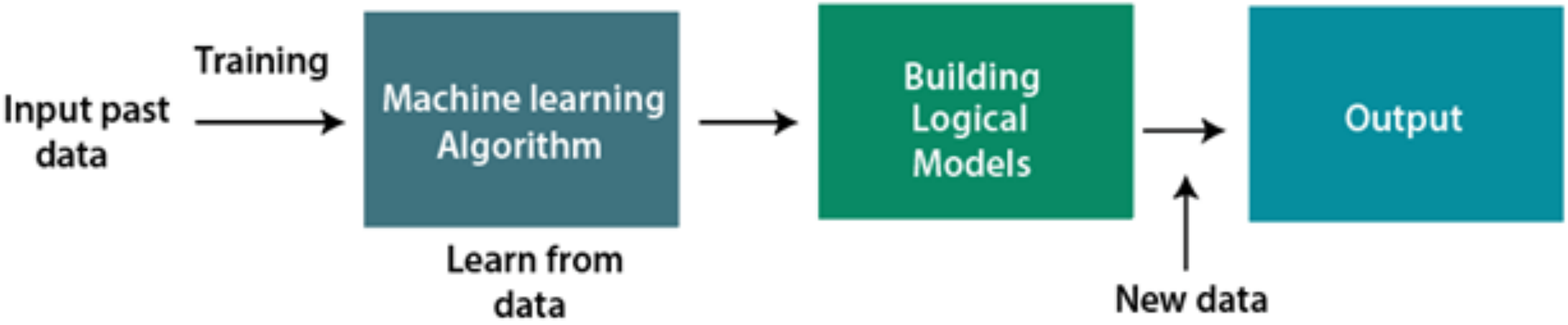
A Machine Learning system learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it.

The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.

Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output.

Machine learning has changed our way of thinking about the problem. The below block diagram explains the working of Machine Learning algorithm:

HOW DOES MACHINE LEARNING WORK



FEATURES OF MACHINE LEARNING

Machine learning uses data to detect various patterns in a given dataset.

It can learn from past data and improve automatically.

It is a data-driven technology.

Machine learning is much similar to data mining as it also deals with the huge amount of the data.

NEED FOR MACHINE LEARNING

The need for machine learning is increasing day by day. The reason behind the need for machine learning is that it is capable of doing tasks that are too complex for a person to implement directly.

As a human, we have some limitations as we cannot access the huge amount of data manually, so for this, we need some computer systems and here comes the machine learning to make things easy for us.

We can train machine learning algorithms by providing them the huge amount of data and let them explore the data, construct the models, and predict the required output automatically.

The performance of the machine learning algorithm depends on the amount of data, and it can be determined by the cost function. With the help of machine learning, we can save both time and money.

NEED FOR MACHINE LEARNING

The importance of machine learning can be easily understood by its uses cases, Currently, machine learning is used in **self-driving cars, cyber fraud detection, face recognition**, and **friend suggestion by Facebook**, etc.

Various top companies such as Netflix and Amazon have build machine learning models that are using a vast amount of data to analyze the user interest and recommend product accordingly.

SOME KEY POINTS WHICH SHOW THE IMPORTANCE OF MACHINE LEARNING

Rapid increment in the production of data

Solving complex problems, which are difficult for a human

Decision making in various sector including finance

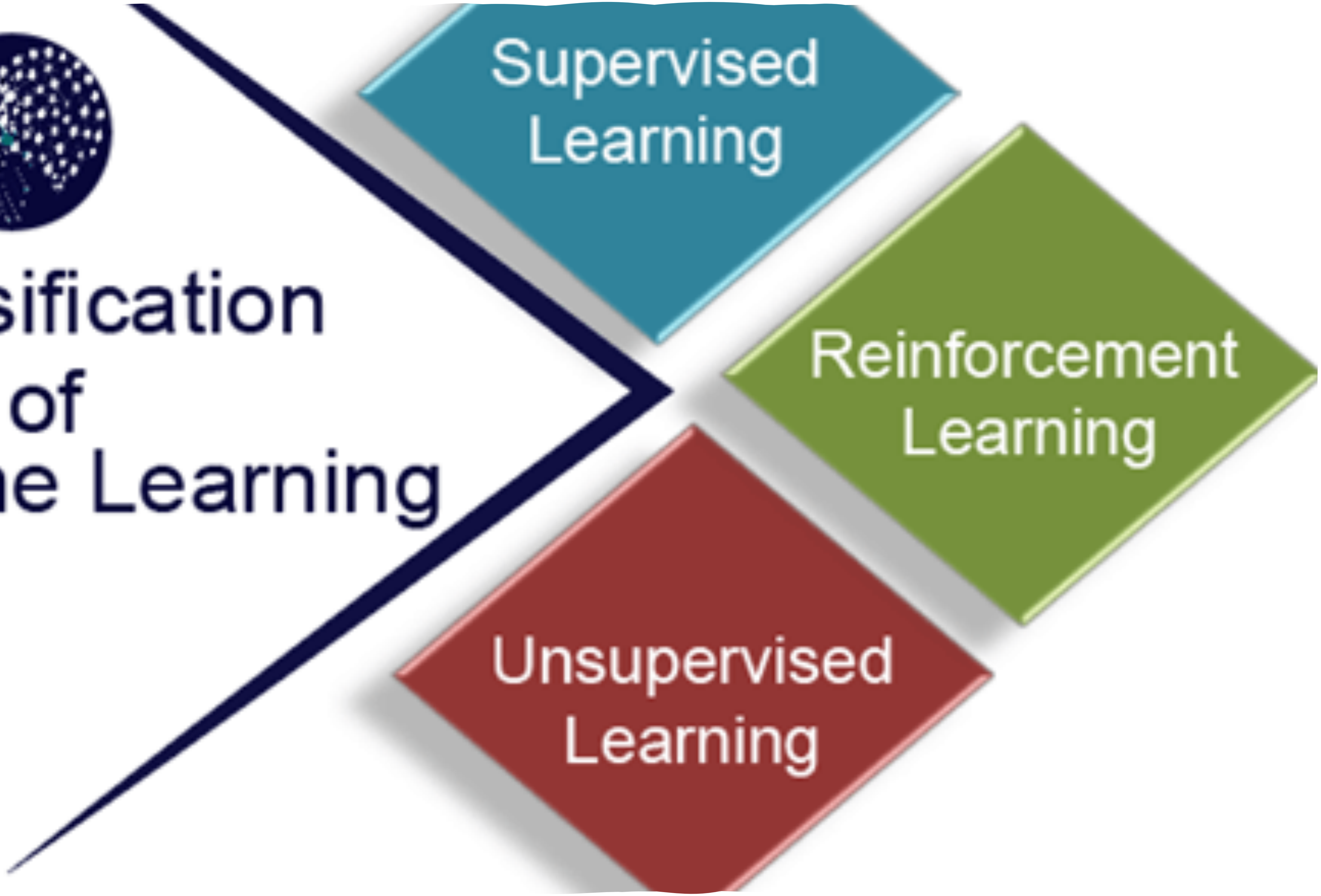
Finding hidden patterns and extracting useful information from data.

CLASSIFICATION OF MACHINE LEARNING

At a broad level, machine learning can be classified into three types:

- Supervised learning
- Unsupervised learning
- Reinforcement learning

Classification of Machine Learning



Supervised Learning

Reinforcement Learning

Unsupervised Learning

SUPERVISED LEARNING

Supervised learning is a type of machine learning method in which we provide sample labeled data to the machine learning system in order to train it, and on that basis, it predicts the output.

The system creates a model using labeled data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providing a sample data to check whether it is predicting the exact output or not.

The goal of supervised learning is to map input data with the output data.

SUPERVISED LEARNING

The supervised learning is based on supervision, and it is the same as when a student learns things in the supervision of the teacher. The example of supervised learning is spam filtering.

Supervised learning can be grouped further in two categories of algorithms:

Classification

Regression

UNSUPERVISED LEARNING

Unsupervised learning is a learning method in which a machine learns without any supervision.

The training is provided to the machine with the set of data that has not been labeled, classified, or categorized, and the algorithm needs to act on that data without any supervision.

The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.

UNSUPERVISED LEARNING

In unsupervised learning, we don't have a predetermined result. The machine tries to find useful insights from the huge amount of data.

It can be further classifieds into two categories of algorithms:

Clustering

Association

REINFORCEMENT LEARNING

Reinforcement learning is a feedback-based learning method, in which a learning agent gets a reward for each right action and gets a penalty for each wrong action.

The agent learns automatically with these feedbacks and improves its performance. In reinforcement learning, the agent interacts with the environment and explores it.

The goal of an agent is to get the most reward points, and hence, it improves its performance.

The robotic dog, which automatically learns the movement of his arms, is an example of Reinforcement learning.

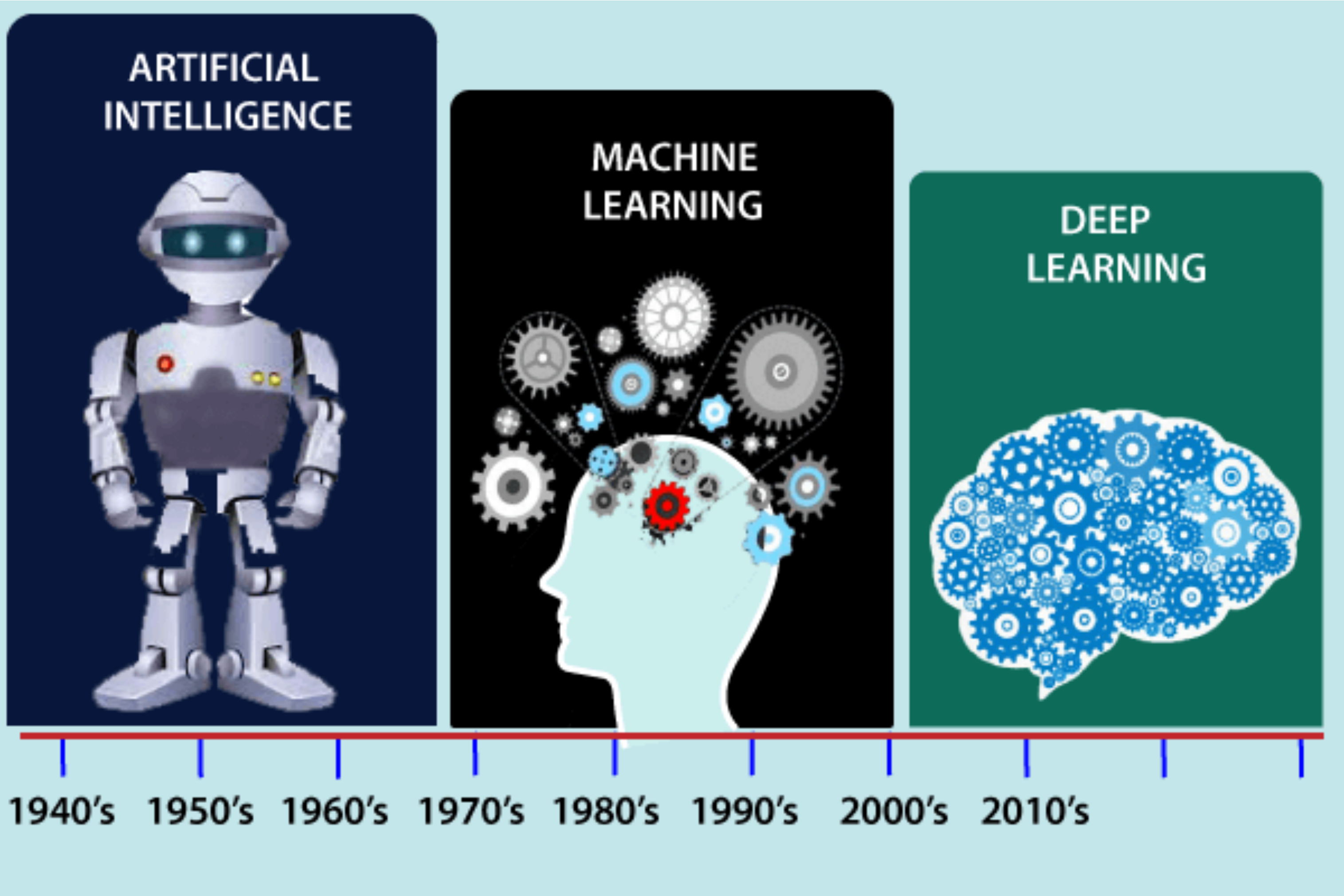
HISTORY OF MACHINE LEARNING

Before some years (about 40-50 years), machine learning was science fiction, but today it is the part of our daily life.

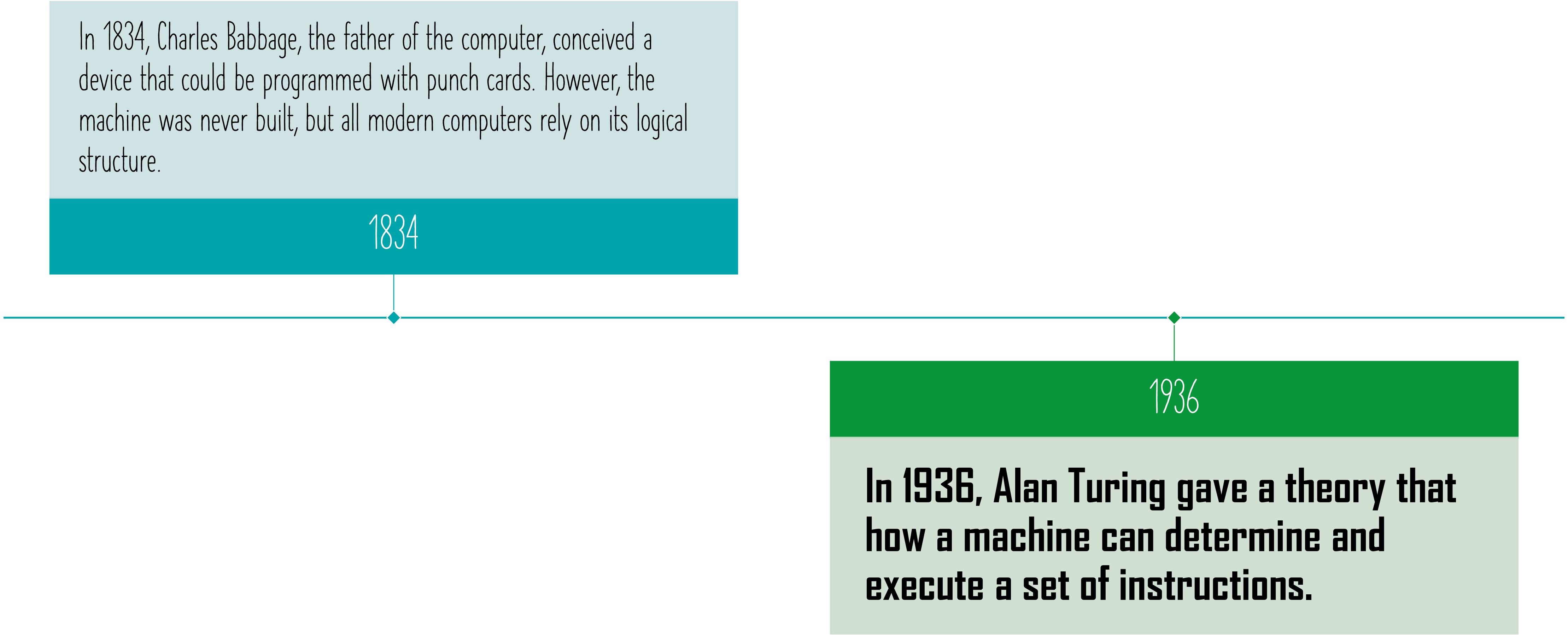
Machine learning is making our day to day life easy from self-driving cars to Amazon virtual assistant "Alexa".

However, the idea behind machine learning is so old and has a long history. Below some milestones are given which have occurred in the history of machine learning:

HISTORY OF MACHINE LEARNING



THE EARLY HISTORY OF MACHINE LEARNING (PRE-1940)



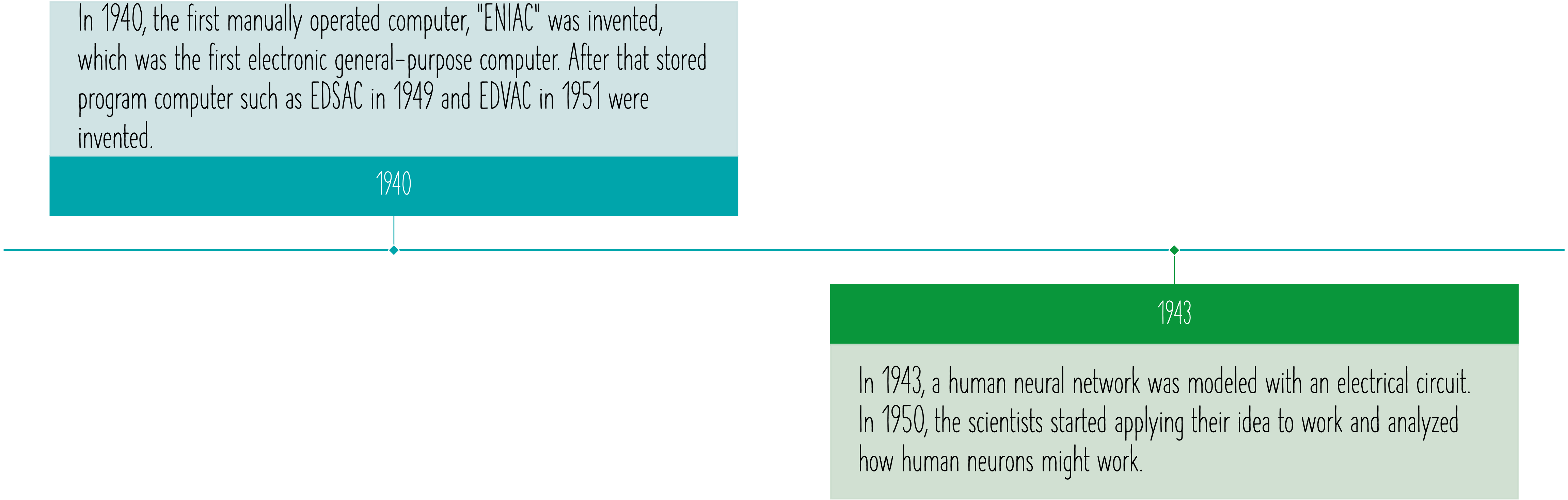
In 1834, Charles Babbage, the father of the computer, conceived a device that could be programmed with punch cards. However, the machine was never built, but all modern computers rely on its logical structure.

1834

1936

In 1936, Alan Turing gave a theory that how a machine can determine and execute a set of instructions.

THE ERA OF STORED PROGRAM COMPUTERS



In 1940, the first manually operated computer, "ENIAC" was invented, which was the first electronic general-purpose computer. After that stored program computer such as EDSAC in 1949 and EDVAC in 1951 were invented.

1940

1943

In 1943, a human neural network was modeled with an electrical circuit. In 1950, the scientists started applying their idea to work and analyzed how human neurons might work.

COMPUTER MACHINERY AND INTELLIGENCE

- **1950:** In 1950, Alan Turing published a seminal paper, "**Computer Machinery and Intelligence**," on the topic of artificial intelligence. **In his paper, he asked, "Can machines think?"**

MACHINE INTELLIGENCE IN GAMES

1952: Arthur Samuel, who was the pioneer of machine learning, created a program that helped an IBM computer to play a checkers game. It performed better more it played.

1959: In 1959, the term "Machine Learning" was first coined by Arthur Samuel.

THE FIRST "AI" WINTER

The duration of 1974 to 1980 was the tough time for AI and ML researchers, and this duration was called as AI winter.

In this duration, failure of machine translation occurred, and people had reduced their interest from AI, which led to reduced funding by the government to the researches.

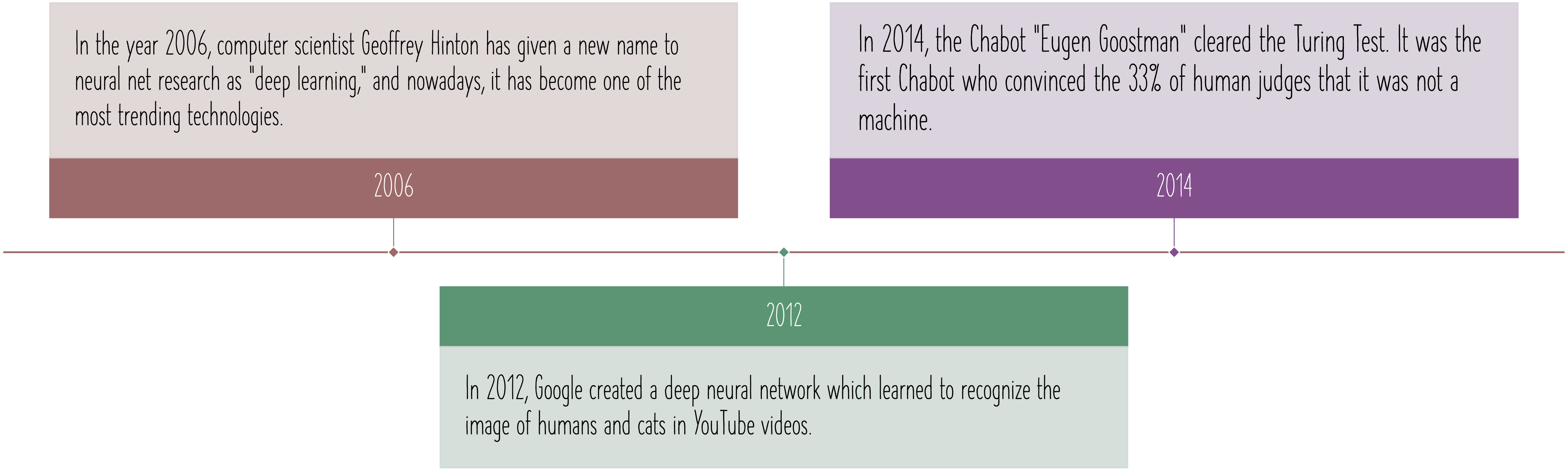
MACHINE LEARNING FROM THEORY TO REALITY

1959: In 1959, the first neural network was applied to a real-world problem to remove echoes over phone lines using an adaptive filter.

1985: In 1985, Terry Sejnowski and Charles Rosenberg invented a neural network NETtalk, which was able to teach itself how to correctly pronounce 20,000 words in one week.

1997: The IBM's Deep blue intelligent computer won the chess game against the chess expert Garry Kasparov, and it became the first computer which had beaten a human chess expert.

MACHINE LEARNING AT 21ST CENTURY



2006

In the year 2006, computer scientist Geoffrey Hinton has given a new name to neural net research as "deep learning," and nowadays, it has become one of the most trending technologies.

2012

In 2012, Google created a deep neural network which learned to recognize the image of humans and cats in YouTube videos.

2014

In 2014, the Chabot "Eugen Goostman" cleared the Turing Test. It was the first Chabot who convinced the 33% of human judges that it was not a machine.

MACHINE LEARNING AT 21ST CENTURY

2014

DeepFace was a deep neural network created by Facebook, and they claimed that it could recognize a person with the same precision as a human can do.

2016

AlphaGo beat the world's number second player Lee sedol at Go game. In 2017 it beat the number one player of this game Ke Jie.

2017

In 2017, the Alphabet's Jigsaw team built an intelligent system that was able to learn the online trolling. It used to read millions of comments of different websites to learn to stop online trolling.

MACHINE LEARNING AT PRESENT

Now machine learning has got a great advancement in its research, and it is present everywhere around us, such as self-driving cars, Amazon Alexa, Catboats, recommender system, and many more.

It includes Supervised, unsupervised, and reinforcement learning with clustering, classification, decision tree, SVM algorithms, etc.

Modern machine learning models can be used for making various predictions, including weather prediction, disease prediction, stock market analysis, etc.

THANK YOU

- **Machine Learning Applications**
- **Machine Learning Life Cycle**

APPLICATIONS OF MACHINE LEARNING

Machine learning is a buzzword for today's technology, and it is growing very rapidly day by day.

We are using machine learning in our daily life even without knowing it such as Google Maps, Google assistant, Alexa, etc.

Below are some most trending real-world applications of Machine Learning:

APPLICATIONS OF MACHINE LEARNING

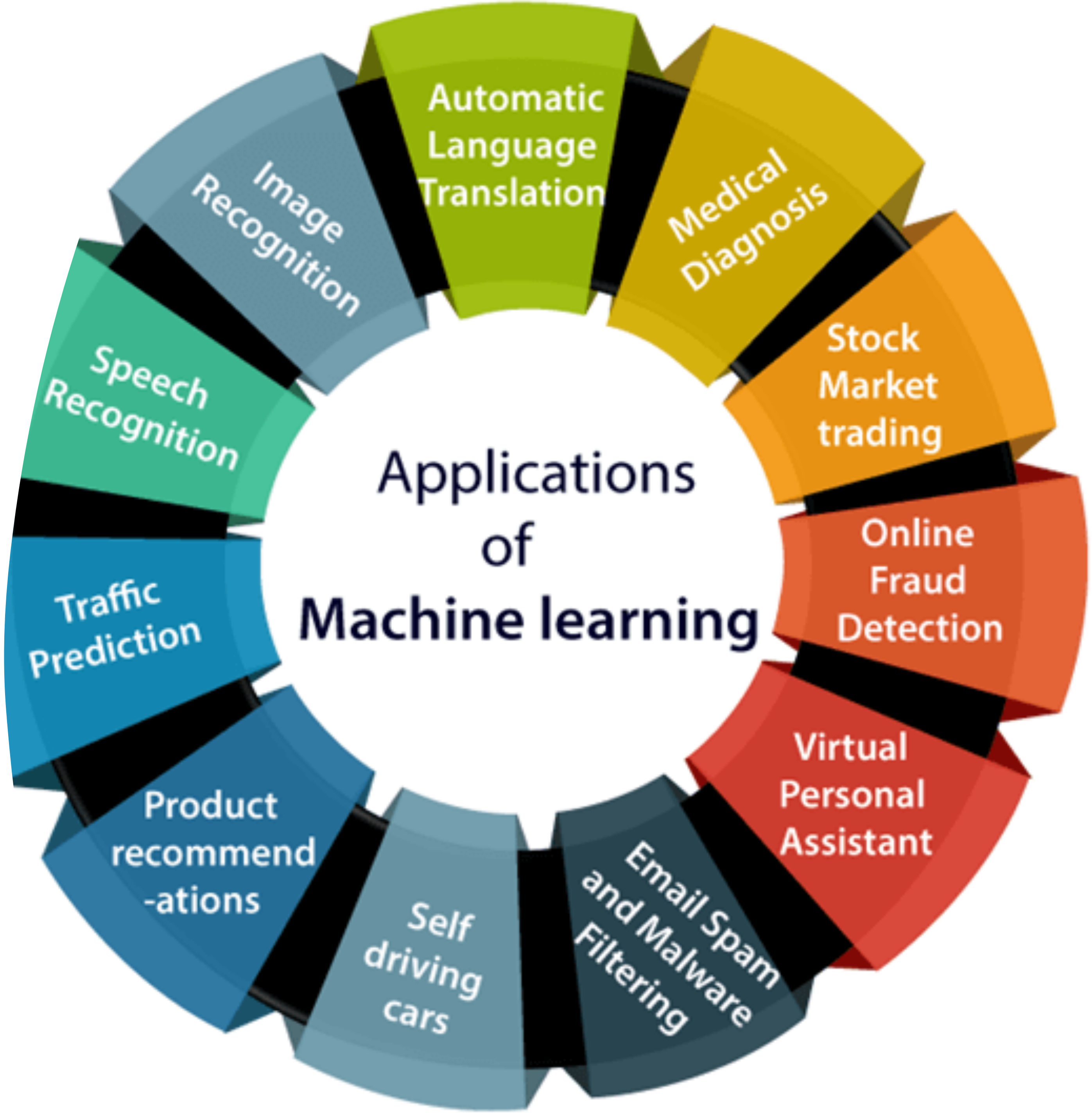

Applications of Machine learning
Automatic Language Translation
Medical Diagnosis
Stock Market trading
Online Fraud Detection
Virtual Personal Assistant
Email Spam and Malware Filtering
Self driving cars
Product recommend -ations
Traffic Prediction
Speech Recognition
Image Recognition

IMAGE RECOGNITION

Image recognition is one of the most common applications of machine learning. It is used to identify objects, persons, places, digital images, etc.

The popular use case of image recognition and face detection is, Automatic friend tagging suggestion:

Facebook provides us a feature of auto friend tagging suggestion. Whenever we upload a photo with our Facebook friends, then we automatically get a tagging suggestion with name, and the technology behind this is machine learning's face detection and recognition algorithm.

It is based on the Facebook project named "Deep Face," which is responsible for face recognition and person identification in the picture.

SPEECH RECOGNITION

While using Google, we get an option of "Search by voice," it comes under speech recognition, and it's a popular application of machine learning.

Speech recognition is a process of converting voice instructions into text, and it is also known as "Speech to text", or "Computer speech recognition."

At present, machine learning algorithms are widely used by various applications of speech recognition.

Google assistant, Siri, Cortana, and Alexa are using speech recognition technology to follow the voice instructions.

TRAFFIC PREDICTION

If we want to visit a new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions.

It predicts the traffic conditions such as whether traffic is cleared, slow-moving, or heavily congested with the help of two ways:

Real Time location of the vehicle form Google Map app and sensors

Average time has taken on past days at the same time.

Everyone who is using Google Map is helping this app to make it better. It takes information from the user and sends back to its database to improve the performance.

PRODUCT RECOMMENDATIONS

Machine learning is widely used by various e-commerce and entertainment companies such as Amazon, Netflix, etc., for product recommendation to the user.

Whenever we search for some product on Amazon, then we started getting an advertisement for the same product while internet surfing on the same browser and this is because of machine learning.

Google understands the user interest using various machine learning algorithms and suggests the product as per customer interest.

As similar, when we use Netflix, we find some recommendations for entertainment series, movies, etc., and this is also done with the help of machine learning.

SELF-DRIVING CARS

One of the most exciting applications of machine learning is self-driving cars. Machine learning plays a significant role in self-driving cars.

Tesla, the most popular car manufacturing company is working on self-driving car. It is using unsupervised learning method to train the car models to detect people and objects while driving.

EMAIL SPAM AND MALWARE FILTERING

Whenever we receive a new email, it is filtered automatically as important, normal, and spam. We always receive an important mail in our inbox with the important symbol and spam emails in our spam box, and the technology behind this is Machine learning.

Following are some spam filters used by Gmail:

Content Filter

Header filter

General blacklists filter

Rules-based filters

Permission filters

Some machine learning algorithms such as Multi-Layer Perceptron, Decision tree, and Naïve Bayes classifier are used for email spam filtering and malware detection.

VIRTUAL PERSONAL ASSISTANT

We have various virtual personal assistants such as Google assistant, Alexa, Cortana, Siri. As the name suggests, they help us in finding the information using our voice instruction.

These assistants can help us in various ways just by our voice instructions such as Play music, call someone, Open an email, Scheduling an appointment, etc.

These virtual assistants use machine learning algorithms as an important part.

These assistant record our voice instructions, send it over the server on a cloud, and decode it using ML algorithms and act accordingly.

ONLINE FRAUD DETECTION

Machine learning is making our online transaction safe and secure by detecting fraud transaction.

Whenever we perform some online transaction, there may be various ways that a fraudulent transaction can take place such as fake accounts, fake ids, and steal money in the middle of a transaction.

So to detect this, Feed Forward Neural network helps us by checking whether it is a genuine transaction or a fraud transaction.

For each genuine transaction, the output is converted into some hash values, and these values become the input for the next round.

For each genuine transaction, there is a specific pattern which gets change for the fraud transaction hence, it detects it and makes our online transactions more secure.

STOCK MARKET TRADING

Machine learning is widely used in stock market trading.

In the stock market, there is always a risk of up and downs in shares, so for this machine learning's long short-term memory neural network is used for the prediction of stock market trends.

MEDICAL DIAGNOSIS

In medical science, machine learning is used for diseases diagnoses.

With this, medical technology is growing very fast and able to build 3D models that can predict the exact position of lesions in the brain.

It helps in finding brain tumors and other brain-related diseases easily.

AUTOMATIC LANGUAGE TRANSLATION

1

Nowadays, if we visit a new place and we are not aware of the language then it is not a problem at all, as for this also machine learning helps us by converting the text into our known languages.

2

Google's GNMT (Google Neural Machine Translation) provide this feature, which is a Neural Machine Learning that translates the text into our familiar language, and it called as automatic translation.

3

The technology behind the automatic translation is a sequence-to-sequence learning algorithm, which is used with image recognition and translates the text from one language to another language.

THANK YOU

- **Machine Learning Applications**
- **Machine Learning Life Cycle**

Jagendra Singh

MACHINE LEARNING LIFE CYCLE

Machine learning has given the computer systems the abilities to automatically learn without being explicitly programmed. But how does a machine learning system work?

So, it can be described using the life cycle of machine learning. Machine learning life cycle is a cyclic process to build an efficient machine learning project.

The main purpose of the life cycle is to find a solution to the problem or project.

MACHINE LEARNING LIFE CYCLE

Machine learning life cycle involves seven major steps, which are given below:

Gathering Data

Data preparation

Data Wrangling

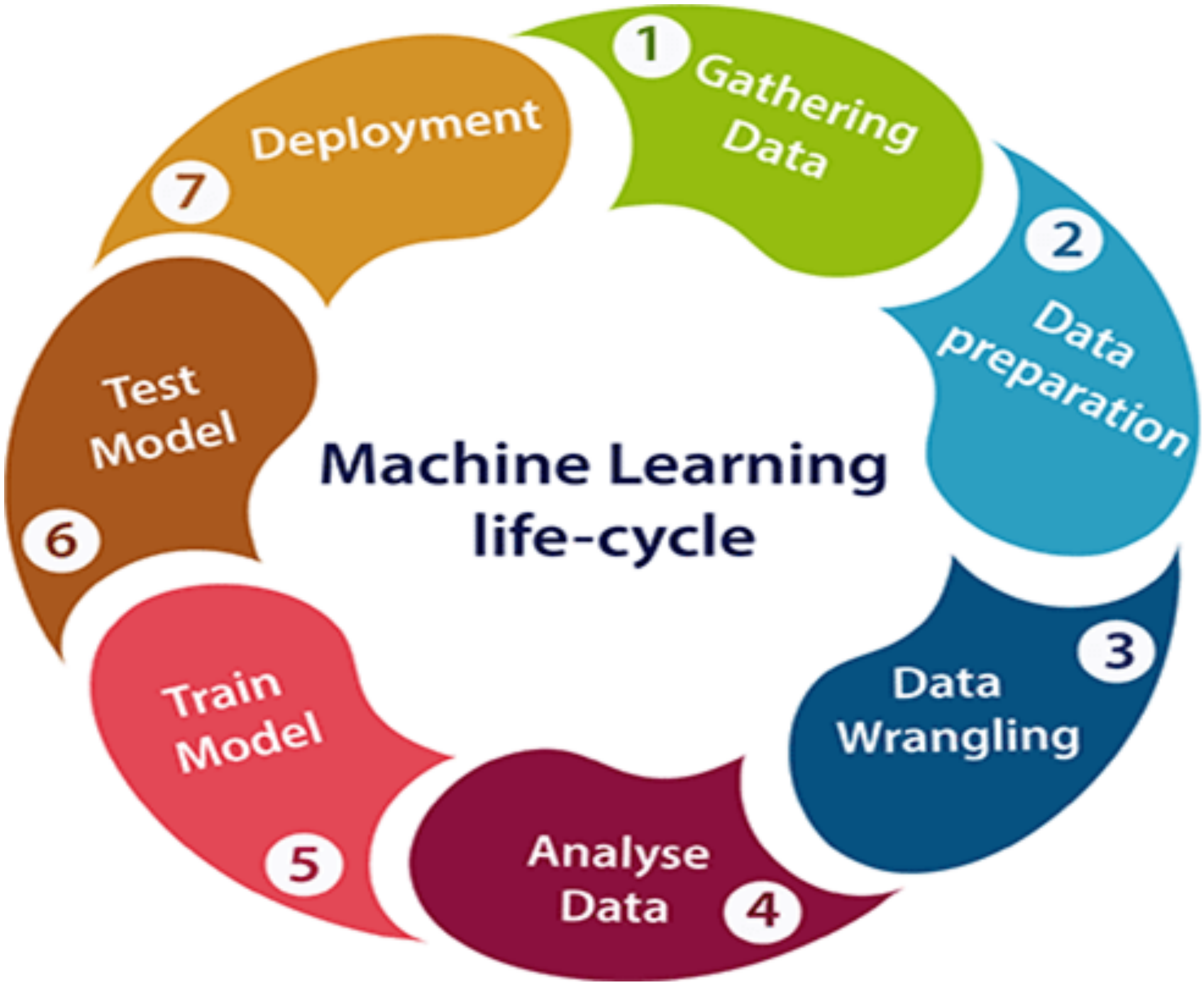
Analyse Data

Train the model

Test the model

Deployment

ML LIFE CYCLE



MACHINE LEARNING LIFE CYCLE

The most important thing in the complete process is to understand the problem and to know the purpose of the problem.

Therefore, before starting the life cycle, we need to understand the problem because the good result depends on the better understanding of the problem.

In the complete life cycle process, to solve a problem, we create a machine learning system called "model", and this model is created by providing "training".

But to train a model, we need data, hence, life cycle starts by collecting data.

GATHERING DATA

Data Gathering is the first step of the machine learning life cycle. The goal of this step is to identify and obtain all data-related problems.

In this step, we need to identify the different data sources, as data can be collected from various sources such as files, database, internet, or mobile devices.

It is one of the most important steps of the life cycle. The quantity and quality of the collected data will determine the efficiency of the output. The more will be the data, the more accurate will be the prediction.

GATHERING DATA

This step includes the below tasks:

Identify various data sources

Collect data

Integrate the data obtained from different sources

By performing the above task, we get a coherent set of data, also called as a dataset. It will be used in further steps.

DATA PREPARATION

After collecting the data, we need to prepare it for further steps. Data preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training.

In this step, first, we put all data together, and then randomize the ordering of data.

This step can be further divided into two processes:

Data exploration:
It is used to understand the nature of data that we have to work with. We need to understand the characteristics, format, and quality of data. A better understanding of data leads to an effective outcome. In this, we find Correlations, general trends, and outliers.

Data pre-processing:
Now the next step is pre-processing of data for its analysis.

DATA WRANGLING

Data wrangling is the process of cleaning and converting raw data into a useable format.

It is the process of cleaning the data, selecting the variable to use, and transforming the data in a proper format to make it more suitable for analysis in the next step.

It is one of the most important steps of the complete process. Cleaning of data is required to address the quality issues.

It is not necessary that data we have collected is always of our use as some of the data may not be useful.

DATA WRANGLING

In real-world applications, collected data may have various issues, including:

Missing Values

Duplicate data

Invalid data

Noise

So, we use various filtering techniques to clean the data.

It is mandatory to detect and remove the above issues because it can negatively affect the quality of the outcome.

DATA ANALYSIS

Now the cleaned and prepared data is passed on to the analysis step. This step involves:

Selection of analytical techniques

Building models

✓ Review the result

The aim of this step is to build a machine learning model to analyze the data using various analytical techniques and review the outcome.

It starts with the determination of the type of the problems, where we select the machine learning techniques such as Classification, Regression, Cluster analysis, Association, etc. then build the model using prepared data, and evaluate the model.

Hence, in this step, we take the data and use machine learning algorithms to build the model.

TRAIN MODEL

Now the next step is to train the model, in this step we train our model to improve its performance for better outcome of the problem.

We use datasets to train the model using various machine learning algorithms. Training a model is required so that it can understand the various patterns, rules, and, features.

TEST MODEL

Once our machine learning model has been trained on a given dataset, then we test the model. In this step, we check for the accuracy of our model by providing a test dataset to it.

Testing the model determines the percentage accuracy of the model as per the requirement of project or problem.

DEPLOYMENT

The last step of machine learning life cycle is deployment, where we deploy the model in the real-world system.

If the above-prepared model is producing an accurate result as per our requirement with acceptable speed, then we deploy the model in the real system.

But before deploying the project, we will check whether it is improving its performance using available data or not. The deployment phase is similar to making the final report for a project.

THANK YOU

- **Artificial intelligence vs Machine learning**
- **Datasets for Machine Learning**

Jagendra Singh

DIFFERENCE BETWEEN AI AND ML

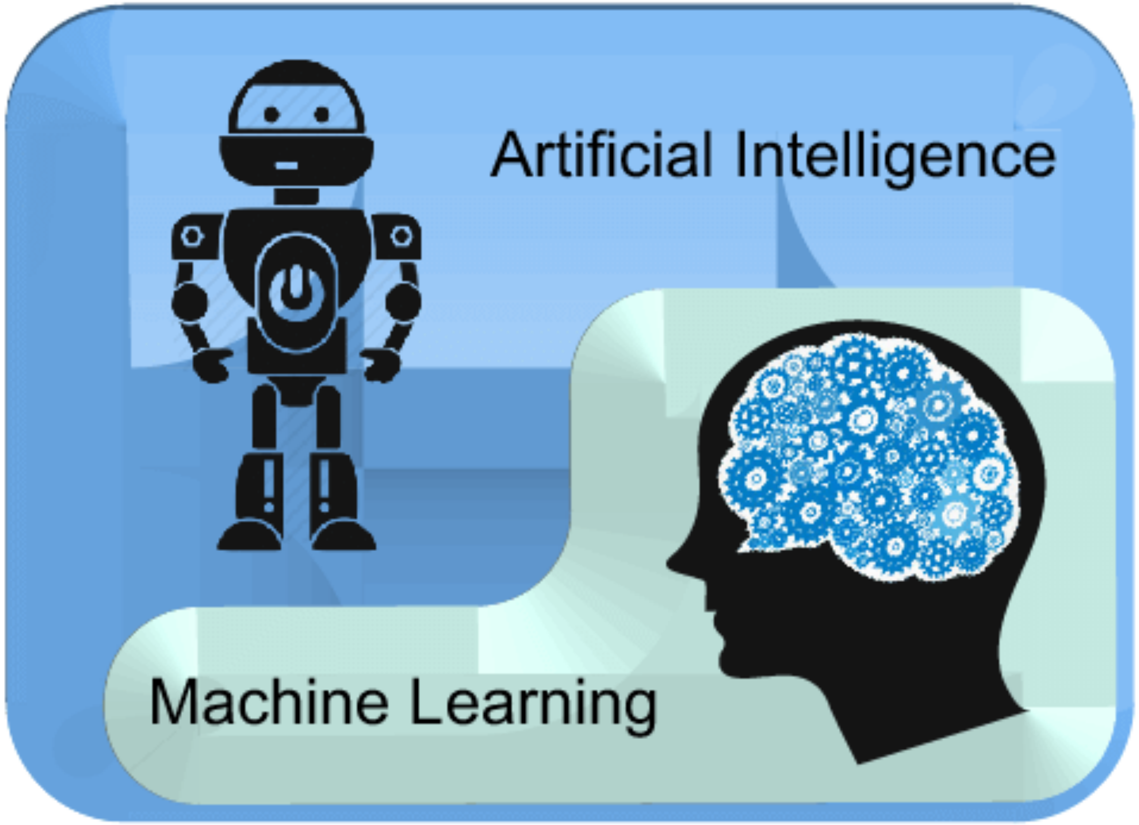
Artificial intelligence and machine learning are the part of computer science that are correlated with each other. These two technologies are the most trending technologies which are used for creating intelligent systems.

Although these are two related technologies and sometimes people use them as a synonym for each other, but still both are the two different terms in various cases.

DIFFERENCE

- On a broad level, we can differentiate both AI and ML as:

➢ AI is a bigger concept to create intelligent machines that can simulate human thinking capability and behaviour, whereas, machine learning is an application or subset of AI that allows machines to learn from data without being programmed explicitly.



ARTIFICIAL INTELLIGENCE

Artificial intelligence is a field of computer science which makes a computer system that can mimic human intelligence.

It is comprised of two words "**Artificial**" and "**intelligence**", which means "a human-made thinking power." Hence we can define it as,

Artificial intelligence is a technology using which we can create intelligent systems that can simulate human intelligence.

The Artificial intelligence system does not require to be pre-programmed, instead of that, they use such algorithms which can work with their own intelligence.

ARTIFICIAL INTELLIGENCE

It involves machine learning algorithms such as Reinforcement learning algorithm and deep learning neural networks.

AI is being used in multiple places such as Siri, Google?s AlphaGo, AI in Chess playing, etc.

Based on capabilities, AI can be classified into three types:

Weak AI

General AI

Strong AI

Currently, we are working with weak AI and general AI. The future of AI is Strong AI for which it is said that it will be intelligent than humans.

MACHINE LEARNING

Machine learning is about extracting knowledge from the data. It can be defined as,

Machine learning is a subfield of artificial intelligence, which enables machines to learn from past data or experiences without being explicitly programmed.

Machine learning enables a computer system to make predictions or take some decisions using historical data without being explicitly programmed.

Machine learning uses a massive amount of structured and semi-structured data so that a machine learning model can generate accurate result or give predictions based on that data.

MACHINE LEARNING

Machine learning works on algorithm which learn by it?s own using historical data. It works only for specific domains such as if we are creating a machine learning model to detect pictures of dogs, it will only give result for dog images, but if we provide a new data like cat image then it will become unresponsive

Machine learning is being used in various places such as for online recommender system, for Google search algorithms, Email spam filter, Facebook Auto friend tagging suggestion, etc.

It can be divided into three types:

Supervised learning

Reinforcement learning

Unsupervised learning

KEY DIFFERENCE BETWEEN AI AND ML

AI is working to create an intelligent system which can perform various complex tasks.	Machine learning is working to create machines that can perform only those specific tasks for which they are trained.
AI system is concerned about maximizing the chances of success.	Machine learning is mainly concerned about accuracy and patterns.
The main applications of AI are Siri, customer support using catboats, Expert System, Online game playing, intelligent humanoid robot, etc.	The main applications of machine learning are Online recommender system, Google search algorithms, Facebook auto friend tagging suggestions, etc.
On the basis of capabilities, AI can be divided into three types, which are, Weak AI, General AI, and Strong AI.	Machine learning can also be divided into mainly three types that are Supervised learning, Unsupervised learning, and Reinforcement learning.
It includes learning, reasoning, and self-correction.	It includes learning and self-correction when introduced with new data.

THANK YOU

- **Artificial intelligence vs Machine learning**
- **Datasets for Machine Learning**

Jagendra Singh

DATASET FOR MACHINE LEARNING

The key to success in the field of machine learning or to become a great data scientist is to practice with different types of datasets.

But discovering a suitable dataset for each kind of machine learning project is a difficult task. So, in this topic, we will provide the detail of the sources from where you can easily get the dataset according to your project.

Before knowing the sources of the machine learning dataset, let's discuss datasets.

WHAT IS A DATASET?

- **A dataset** is a collection of data in which data is arranged in some order. A dataset can contain any data from a series of an array to a database table. This table shows an example of the dataset:

Country	Age	Salary	Purchased
India	38	48000	No
France	43	45000	Yes
Germany	30	54000	No
France	48	65000	No
Germany	40		Yes
India	35	58000	Yes

WHAT IS A DATASET?

- A tabular dataset can be understood as a database table or matrix, where each column corresponds to a **particular variable,** and each row corresponds to the **fields of the dataset.**

- The most supported file type for a tabular dataset is **"Comma Separated File,"** or **CSV.** But to store a "tree-like data," we can use the JSON file more efficiently.

TYPES OF DATA IN DATASETS

- **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

To work with machine learning projects, we need a huge amount of data, because, without the data, one cannot train ML/AI models. Collecting and preparing the dataset is one of the most crucial parts while creating an ML/AI project.

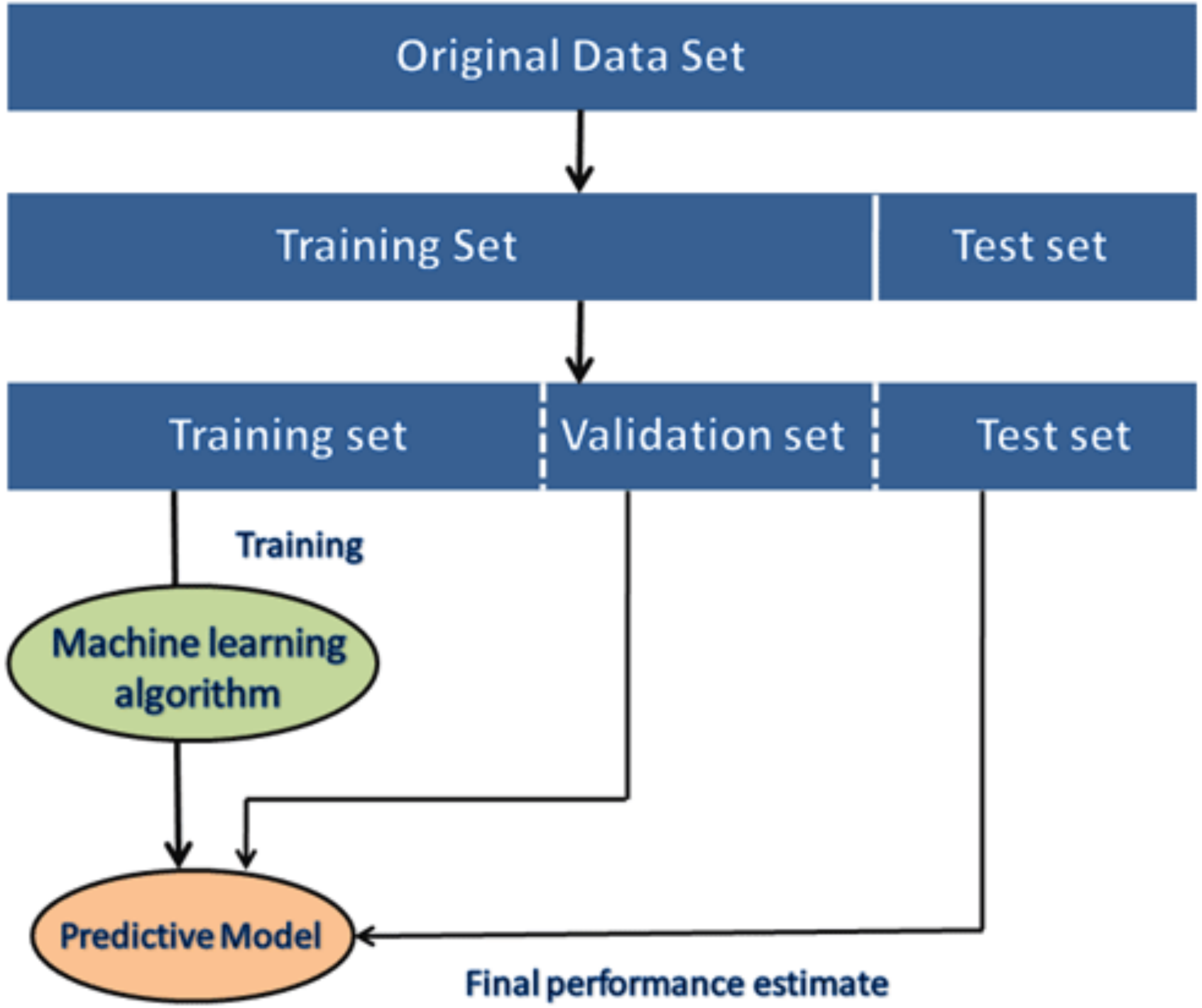
The technology applied behind any ML projects cannot work properly if the dataset is not well prepared and pre-processed.

During the development of the ML project, the developers completely rely on the datasets. In building ML applications, datasets are divided into two parts

Training dataset:

Test Dataset

NEED OF DATASET

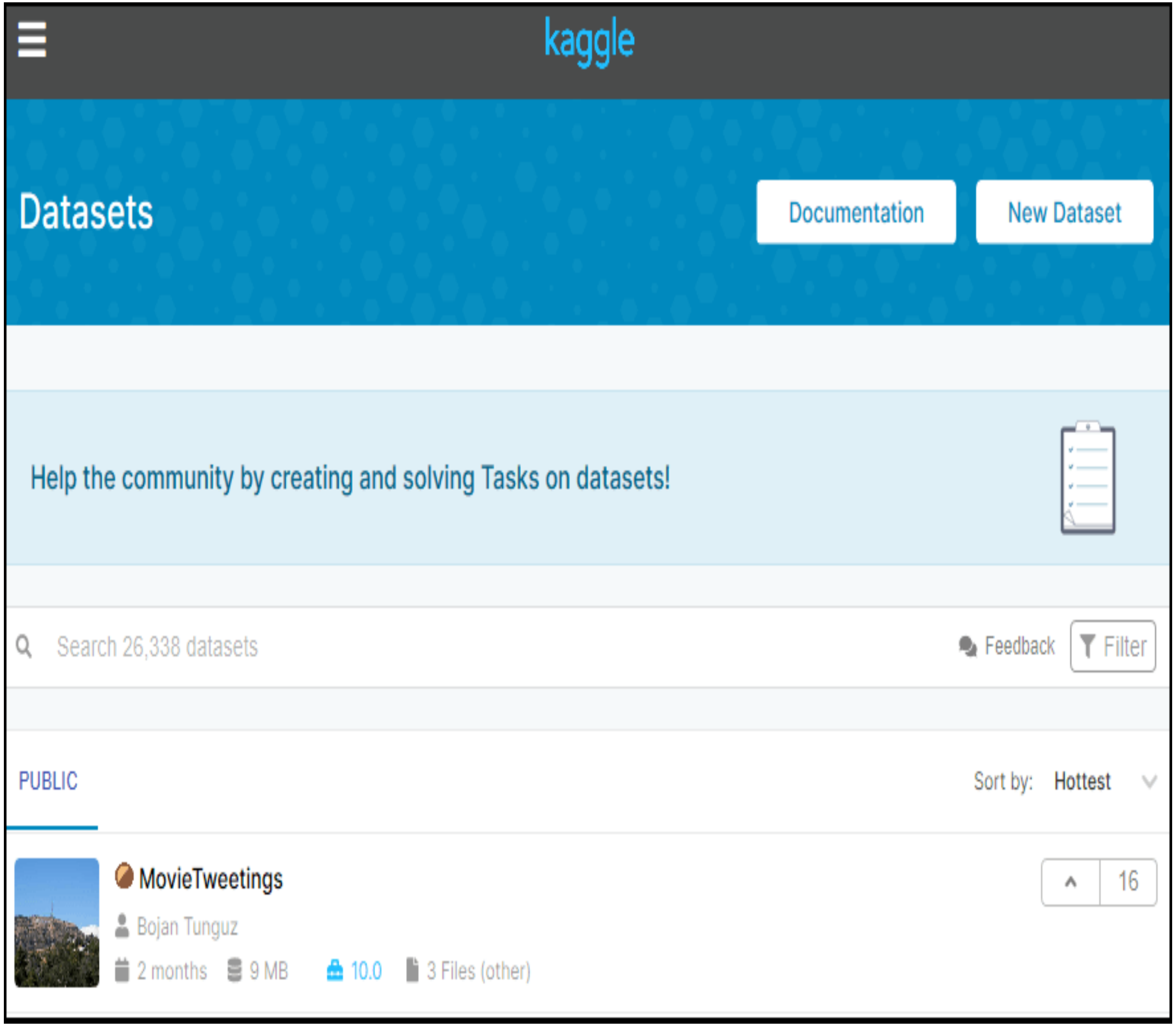


SOURCES FOR MACHINE LEARNING DATASETS

1 – Kaggle Datasets

- Kaggle is one of the best sources for providing datasets for Data Scientists and Machine Learners. It allows users to find, download, and publish datasets in an easy way.
- It also provides the opportunity to work with other machine learning engineers and solve difficult Data Science related tasks.
- Kaggle provides a high-quality dataset in different formats that we can easily find and download.
- The link for the Kaggle dataset is https://www.kaggle.com/datasets

KAGGLE DATASETS



Help the community by creating and solving Tasks on datasets!

MovieTweetings

Bojan Tunguz

2 months 9 MB 10.0 3 Files (other)

UCI MACHINE LEARNING REPOSITORY

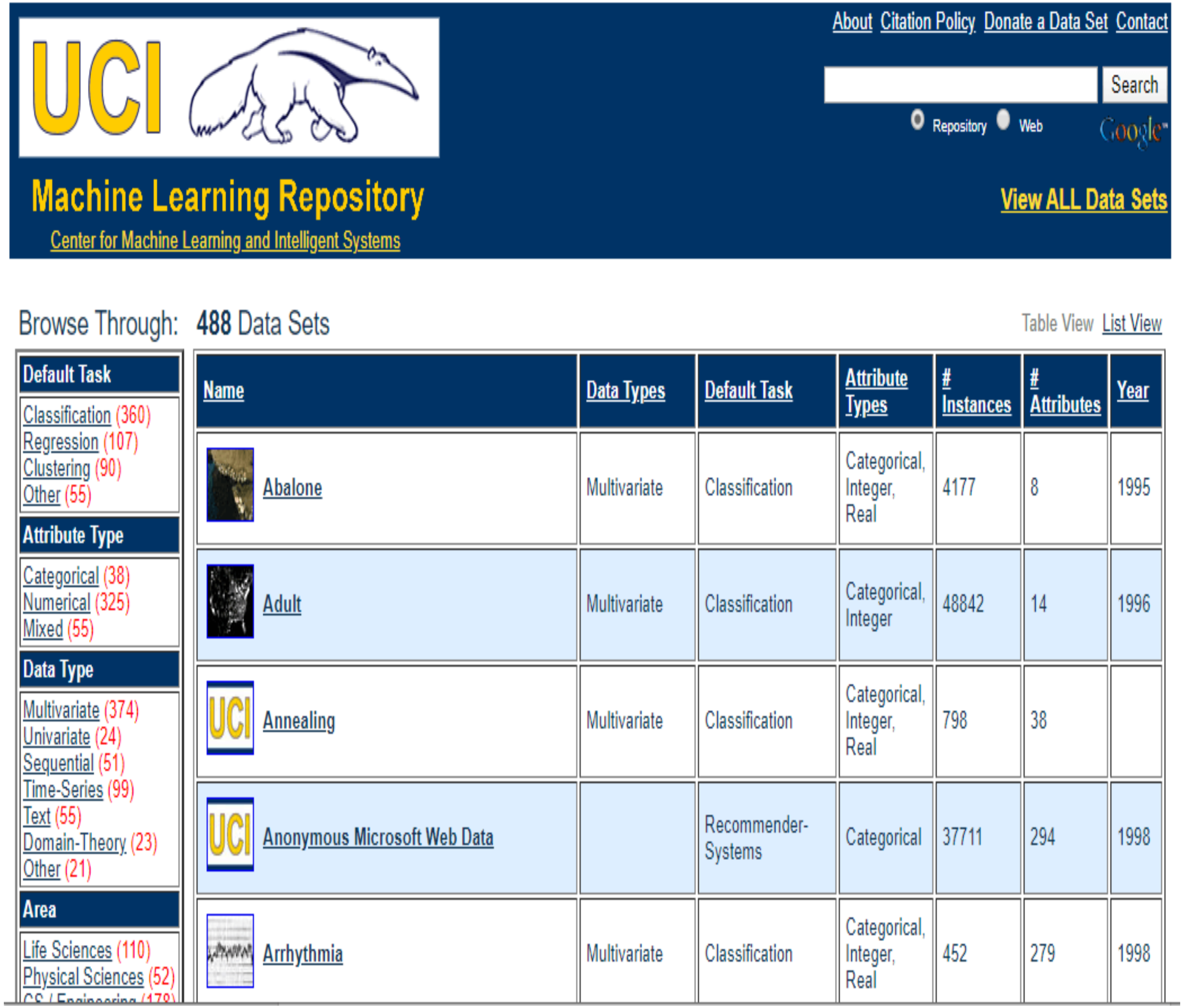
UCI Machine learning repository is one of the great sources of machine learning datasets. This repository contains databases, domain theories, and data generators that are widely used by the machine learning community for the analysis of ML algorithms.

Since the year 1987, it has been widely used by students, professors, researchers as a primary source of machine learning dataset.

It classifies the datasets as per the problems and tasks of machine learning such as Regression, Classification, Clustering, etc. It also contains some of the popular datasets such as the Iris dataset, Car Evaluation dataset, Poker Hand dataset, etc.

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

UCI MACHINE LEARNING REPOSITORY



Machine Learning Repository

Center for Machine Learning and Intelligent Systems

View ALL Data Sets

Browse Through: 488 Data Sets

Table View List View

Default Task

- Classification (360)
- Regression (107)
- Clustering (90)
- Other (55)

Attribute Type

- Categorical (38)
- Numerical (325)
- Mixed (55)

Data Type

- Multivariate (374)
- Univariate (24)
- Sequential (51)
- Time-Series (99)
- Text (55)
- Domain-Theory (23)
- Other (21)

Area

- Life Sciences (110)
- Physical Sciences (52)
- CS / Engineering (178)

Name	Data Types	Default Task	Attribute Types	# Instances	# Attributes	Year
Abalone	Multivariate	Classification	Categorical, Integer, Real	4177	8	1995
Adult	Multivariate	Classification	Categorical, Integer	48842	14	1996
Annealing	Multivariate	Classification	Categorical, Integer, Real	798	38	
Anonymous Microsoft Web Data		Recommender-Systems	Categorical	37711	294	1998
Arrhythmia	Multivariate	Classification	Categorical, Integer, Real	452	279	1998

DATASETS VIA AWS

We can search, download, access, and share the datasets that are publicly available via AWS resources. These datasets can be accessed through AWS resources but provided and maintained by different government organizations, researches, businesses, or individuals.

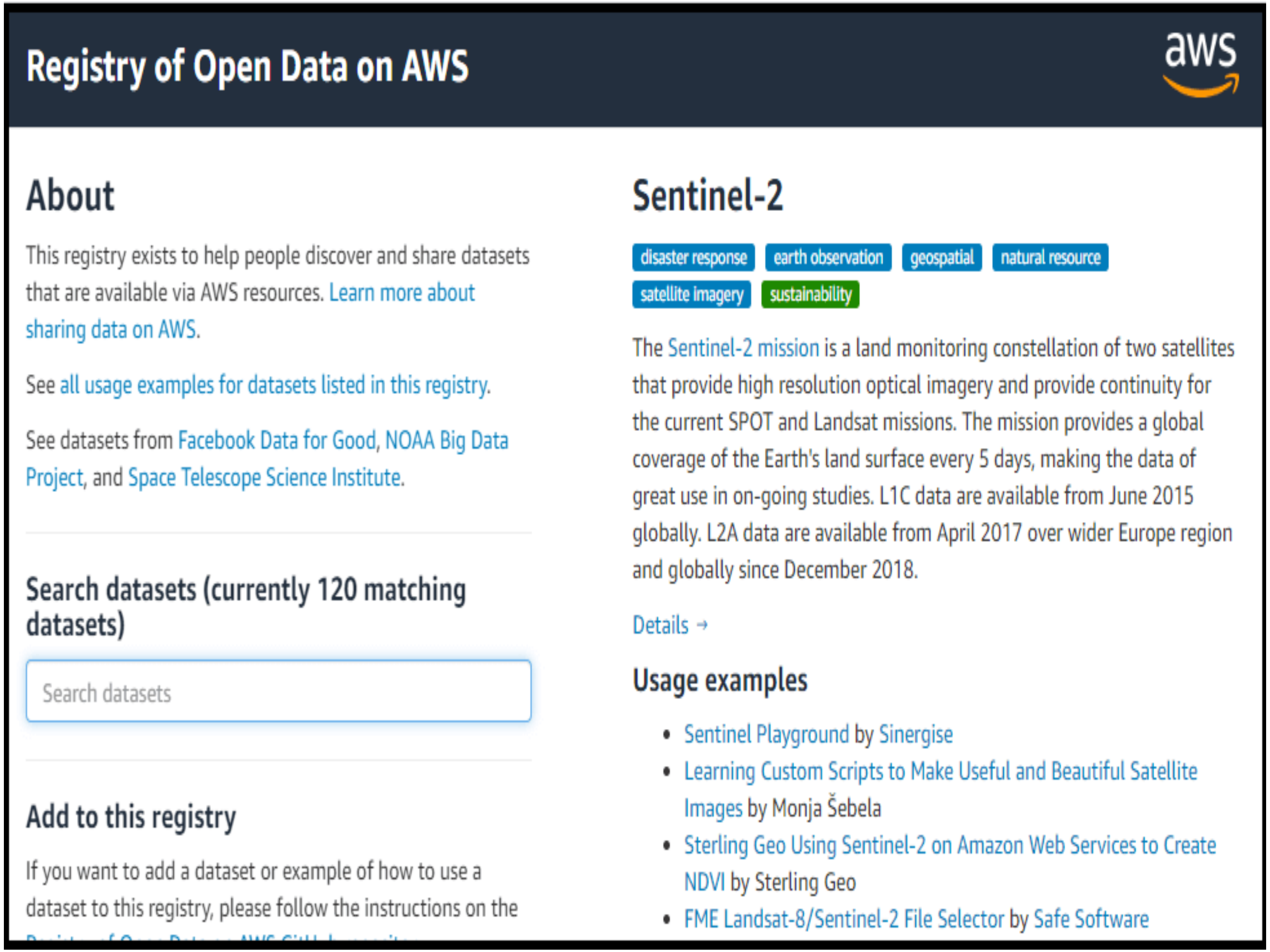
Anyone can analyze and build various services using shared data via AWS resources. The shared dataset on cloud helps users to spend more time on data analysis rather than on acquisitions of data

This source provides the various types of datasets with examples and ways to use the dataset. It also provides the search box using which we can search for the required dataset. Anyone can add any dataset or example to the Registry of Open Data on AWS

https://www.opendatani.gov.uk

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

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)

Search 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

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

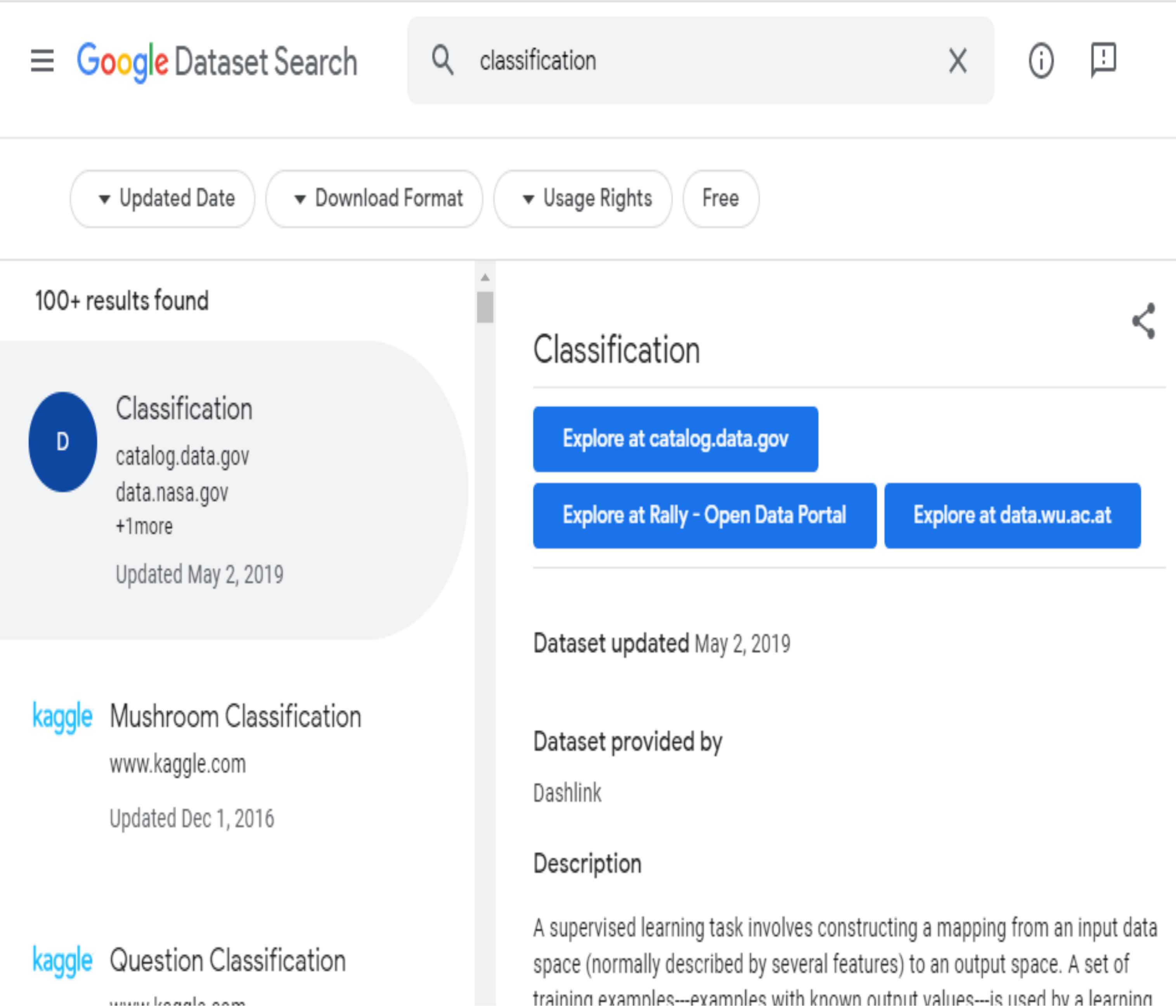
- Sentinel Playground by Sinergise
- Learning Custom Scripts to Make Useful and Beautiful Satellite Images by Monja Šebela
- Sterling Geo Using Sentinel-2 on Amazon Web Services to Create NDVI by Sterling Geo
- FME Landsat-8/Sentinel-2 File Selector by Safe Software

GOOGLE'S DATASET SEARCH ENGINE

Google dataset search engine is a search engine launched by Google on September 5, 2018. This source helps researchers to get online datasets that are freely available for use.

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

GOOGLE'S DATASET SEARCH ENGINE



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

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

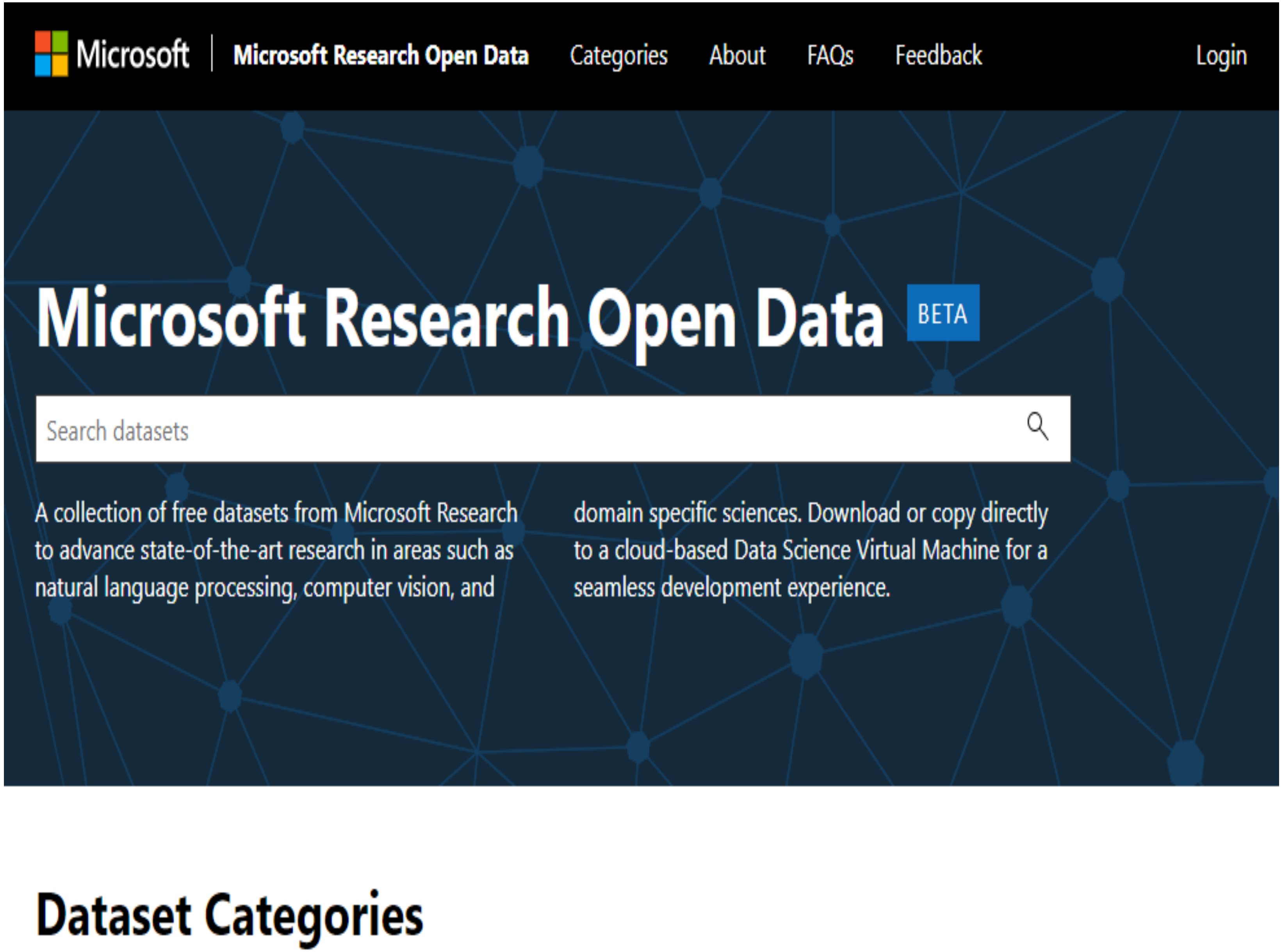
MICROSOFT DATASETS

The Microsoft has launched the **"Microsoft Research Open data"** repository with the collection of free datasets in various areas such as **natural language processing, computer vision, and domain-specific sciences.**

Using this resource, we can download the datasets to use on the current device, or we can also directly use it on the cloud infrastructure.

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

MICROSOFT DATASETS



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

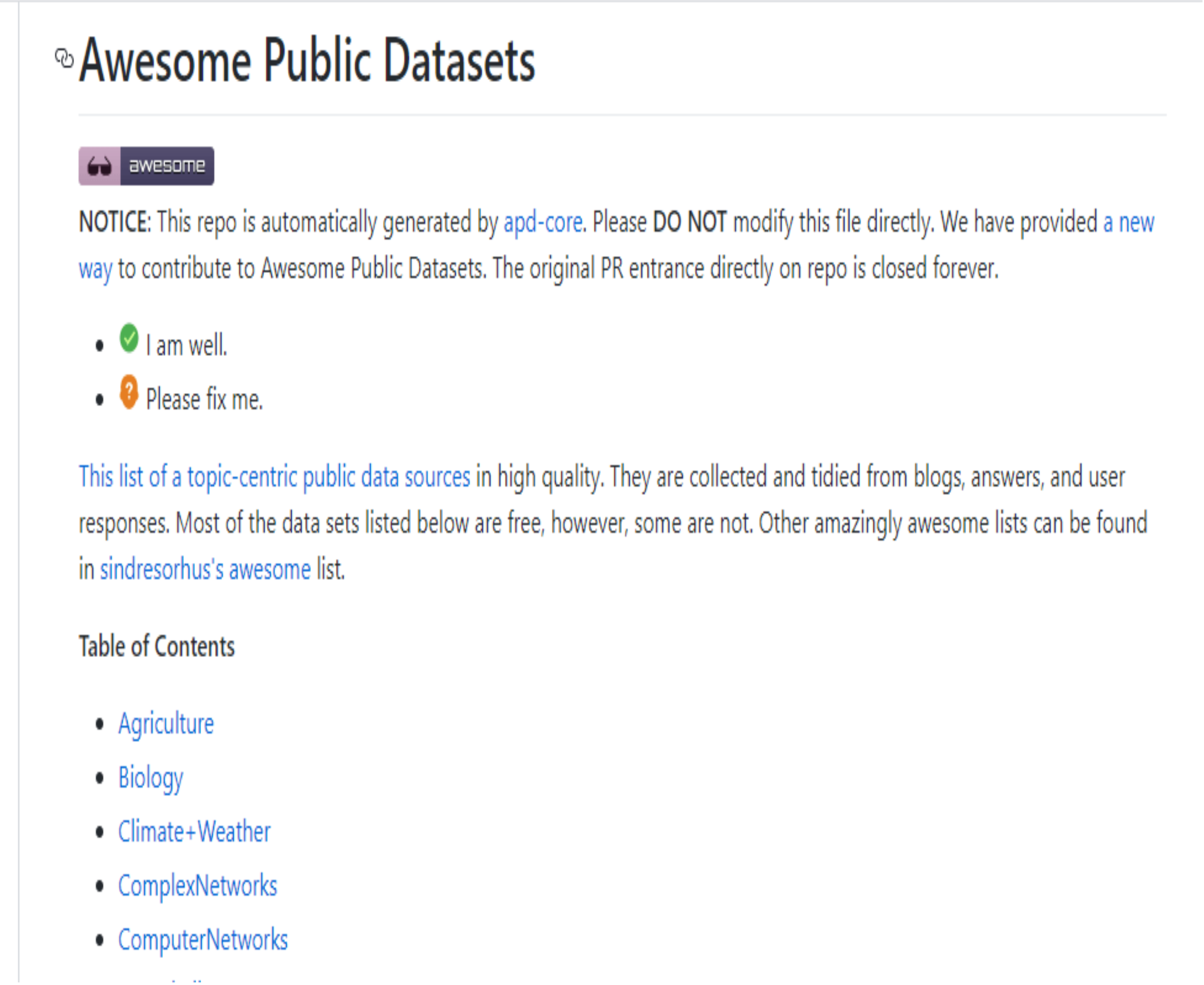
AWESOME PUBLIC DATASET COLLECTION

Awesome public dataset collection provides high-quality datasets that are arranged in a well-organized manner within a list according to topics such as Agriculture, Biology, Climate, Complex networks, etc.

Most of the datasets are available free, but some may not, so it is better to check the license before downloading the dataset.

The link to download the dataset from Awesome public dataset collection is https://github.com/awesomedata/awesome-public-datasets

AWESOME PUBLIC DATASET COLLECTION



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

GOVERNMENT DATASETS

There are different sources to get government-related data. Various countries publish government data for public use collected by them from different departments.

The goal of providing these datasets is to increase transparency of government work among the people and to use the data in an innovative approach. Below are some links of government datasets:

Indian Government dataset

US Government Dataset

Northern Ireland Public Sector Datasets

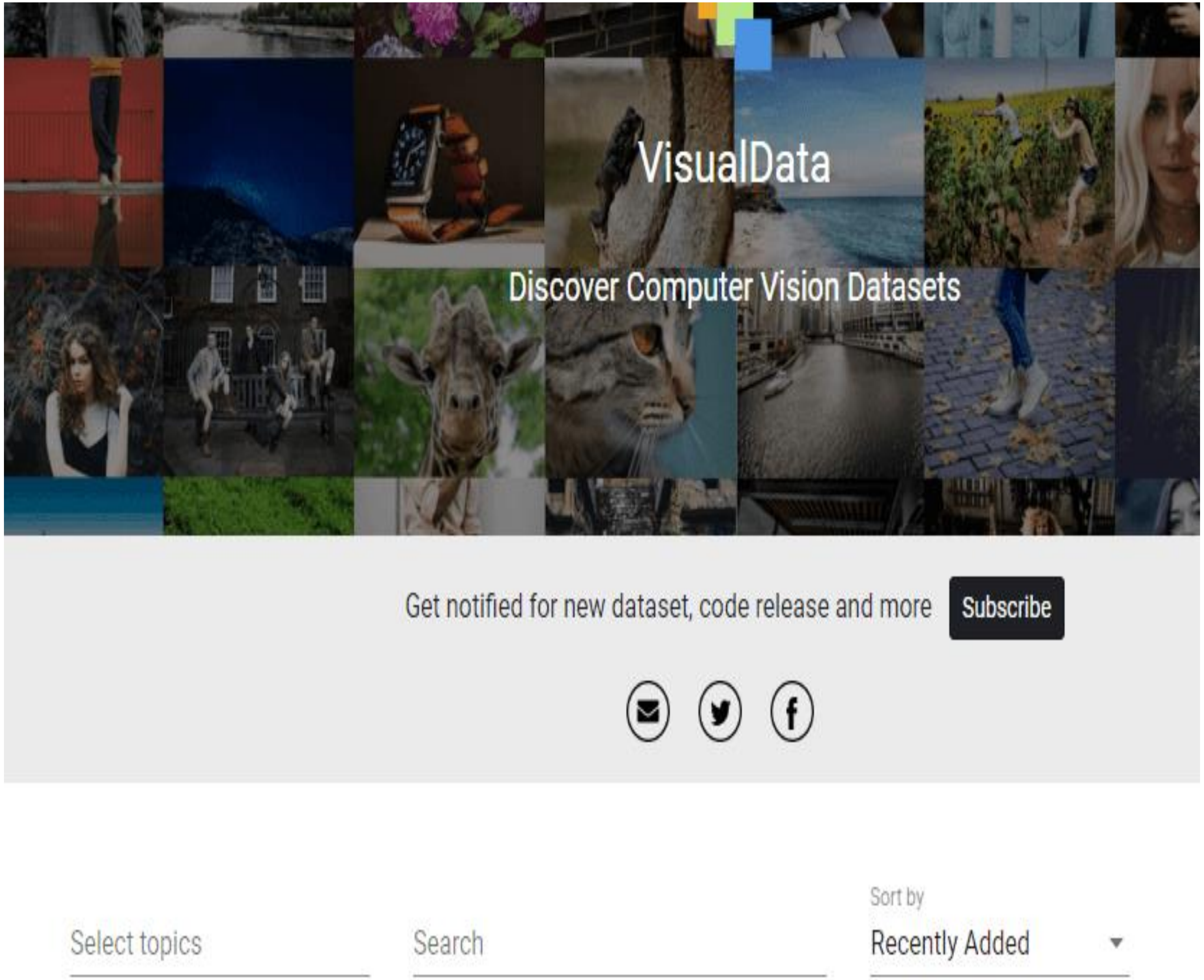
COMPUTER VISION DATASETS

Visual data provides multiple numbers of the great dataset that are specific to computer visions such as Image Classification, Video classification, Image Segmentation, etc.

Therefore, if you want to build a project on deep learning or image processing, then you can refer to this source.

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

COMPUTER VISION DATASETS



VisualData

Discover Computer Vision Datasets

Get notified for new dataset, code release and more

SCIKIT-LEARN DATASET

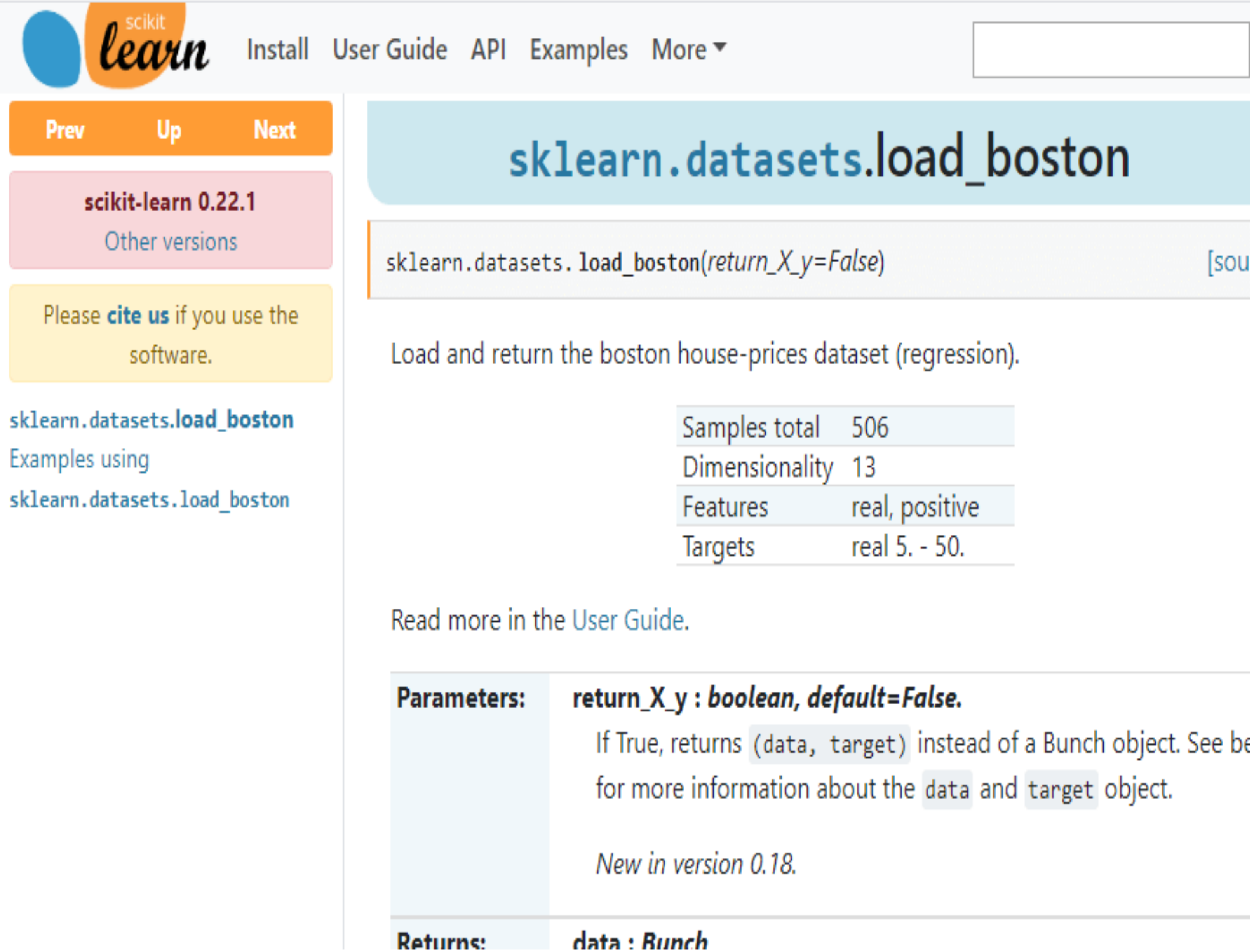
Scikit-learn is a great source for machine learning enthusiasts. This source provides both toy and real-world datasets.

These datasets can be obtained from sklearn.datasets package and using general dataset API.

The toy dataset available on scikit-learn can be loaded using some predefined functions such as, load_boston([return_X_y]), load_iris([return_X_y]), etc, rather than importing any file from external sources. But these datasets are not suitable for real-world projects.

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

SCIKIT-LEARN DATASET



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)` [sou

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 be for more information about the `data` and `target` object.

New in version 0.18.

Returns: **data : *Bunch***

THANK YOU

- **Data Preprocessing in Machine learning**

Dr. Jagendra Singh

DATA PREPROCESSING

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model.

It is the first and crucial step while creating a machine learning model.

When creating a machine learning project, it is not always a case that we come across the clean and formatted data.

And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.

WHY DO WE NEED DATA PREPROCESSING?

- A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models.
- Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.
- It involves following steps:
  - **Getting the dataset**
  - **Importing libraries**
  - **Importing datasets**
  - **Finding Missing Data**
  - **Encoding Categorical Data**
  - **Splitting dataset into training and test set**
  - **Feature scaling**

1. GET THE DATASET

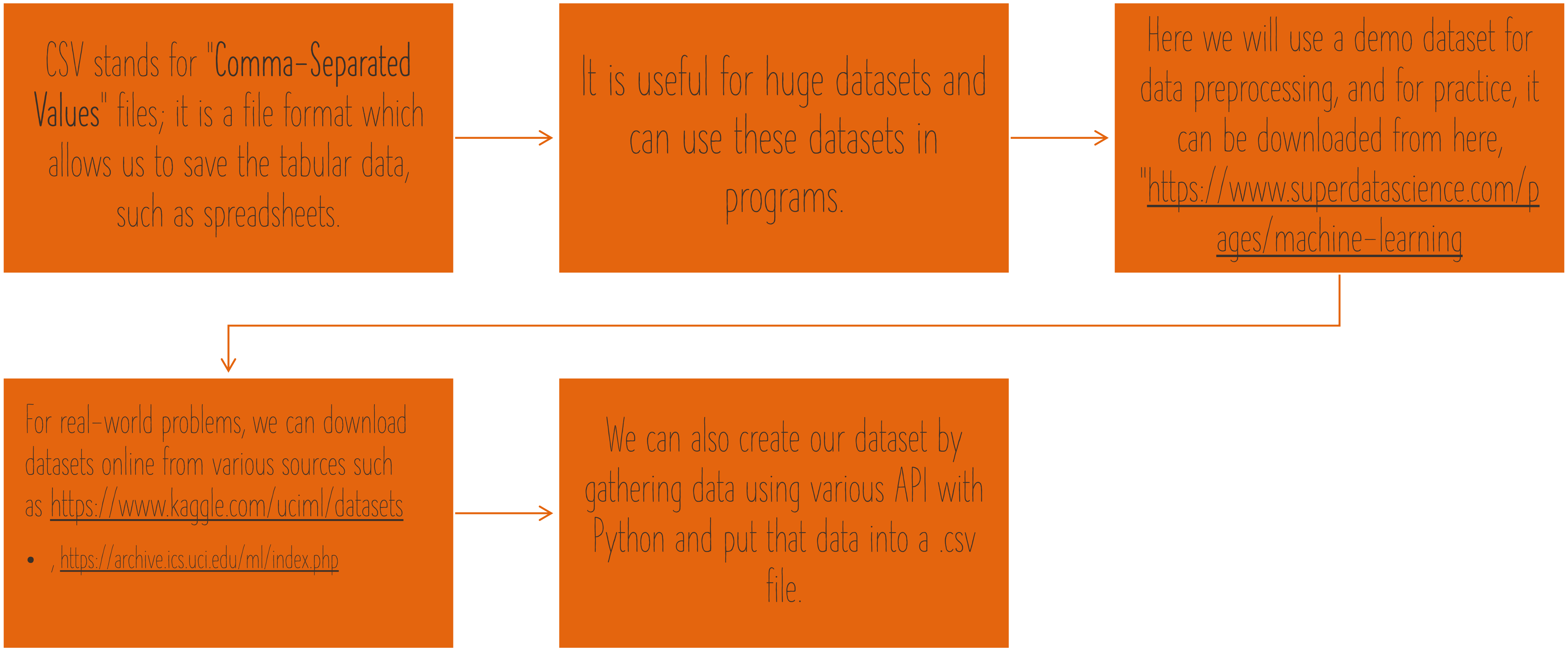
To create a machine learning model, the first thing we required is a dataset as a machine learning model completely works on data.

The collected data for a particular problem in a proper format is known as the **dataset**.

Dataset may be of different formats for different purposes, such as, if we want to create a machine learning model for business purpose, then dataset will be different with the dataset required for a liver patient.

So each dataset is different from another dataset. To use the dataset in our code, we usually put it into a CSV **file**.

WHAT IS A CSV FILE?



CSV stands for "Comma-Separated Values" files; it is a file format which allows us to save the tabular data, such as spreadsheets.

It is useful for huge datasets and can use these datasets in programs.

Here we will use a demo dataset for data preprocessing, and for practice, it can be downloaded from here, "https://www.superdatascience.com/pages/machine-learning

For real-world problems, we can download datasets online from various sources such as https://www.kaggle.com/uciml/datasets

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

We can also create our dataset by gathering data using various API with Python and put that data into a .csv file.

2. IMPORTING LIBRARIES

There are three specific libraries that we will use for data preprocessing, which are:

Numpy: Numpy Python library is used for including any type of mathematical operation in the code. It is the fundamental package for scientific calculation in Python. It also supports to add large, multidimensional arrays and matrices. So, in Python, we can import it as:

- import numpy as nm

Matplotlib: The second library is matplotlib, which is a Python 2D plotting library, and with this library, we need to import a sub-library pyplot. This library is used to plot any type of charts in Python for the code. It will be imported as below:

import matplotlib.pyplot as mpt

Pandas: The last library is the Pandas library, which is one of the most famous Python libraries and used for importing and managing the datasets. It is an open-source data manipulation and analysis library.

3. IMPORTING THE DATASETS

Now we need to import the datasets which we have collected for our machine learning project.

read_csv() function:

- Now to import the dataset, we will use read_csv() function of pandas library, which is used to read a csv file and performs various operations on it. Using this function, we can read a csv file locally as well as through an URL.

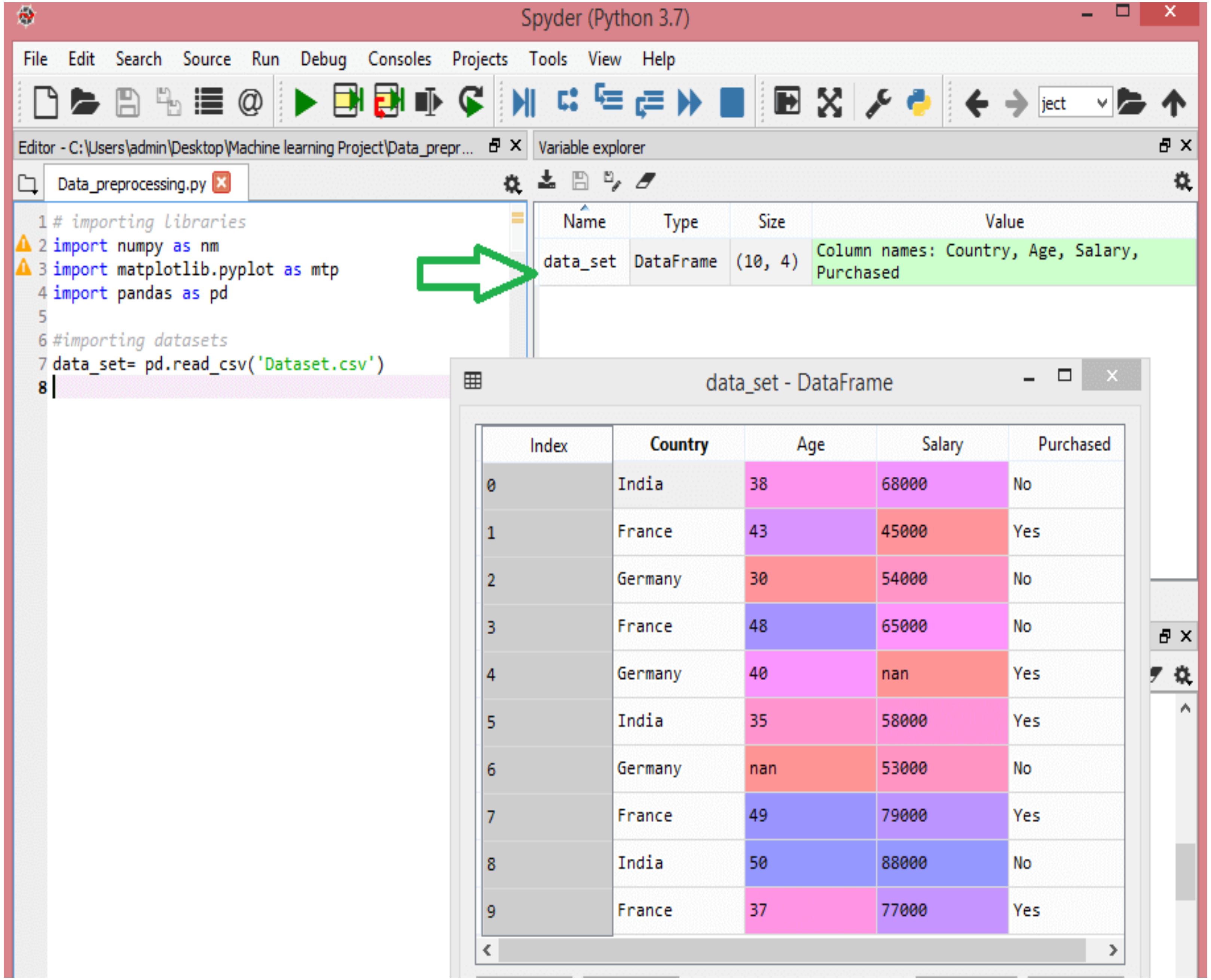
IMPORTING THE DATASETS

We can use read_csv function as below:

```
data_set= pd.read_csv('Dataset.csv')
```

- Here, **data_set** is a name of the variable to store our dataset, and inside the function, we have passed the name of our dataset.
- Once we execute the above line of code, it will successfully import the dataset in our code. We can also check the imported dataset by clicking on the section **variable explorer**, and then double click on **data_set**. Consider the below image:

IMPORTING THE DATASETS



Spyder (Python 3.7)
File Edit Search Source Run Debug Consoles Projects Tools View Help
Editor - C:\Users\admin\Desktop\Machine learning Project\Data_prepr...
Data_preprocessing.py
Variable explorer
data_set - DataFrame

```
# importing libraries
import numpy as nm
import matplotlib.pyplot as mtp
import pandas as pd

#importing datasets
data_set= pd.read_csv('Dataset.csv')
```

Name	Type	Size	Value
data_set	DataFrame	(10, 4)	Column names: Country, Age, Salary, Purchased

Index	Country	Age	Salary	Purchased
0	India	38	68000	No
1	France	43	45000	Yes
2	Germany	30	54000	No
3	France	48	65000	No
4	Germany	40	nan	Yes
5	India	35	58000	Yes
6	Germany	nan	53000	No
7	France	49	79000	Yes
8	India	50	88000	No
9	France	37	77000	Yes

4. HANDLING MISSING DATA

The next step of data preprocessing is to handle missing data in the datasets. If our dataset contains some missing data, then it may create a huge problem for our machine learning model. Hence it is necessary to handle missing values present in the dataset.

Ways to handle missing data

- There are mainly two ways to handle missing data, which are:
- By deleting the particular row: The first way is used to commonly deal with null values. In this way, we just delete the specific row or column which consists of null values. But this way is not so efficient and removing data may lead to loss of information which will not give the accurate output.
- By calculating the mean: In this way, we will calculate the mean of that column or row which contains any missing value and will put it on the place of missing value. This strategy is useful for the features which have numeric data such as age, salary, year, etc. Here, we will use this approach.

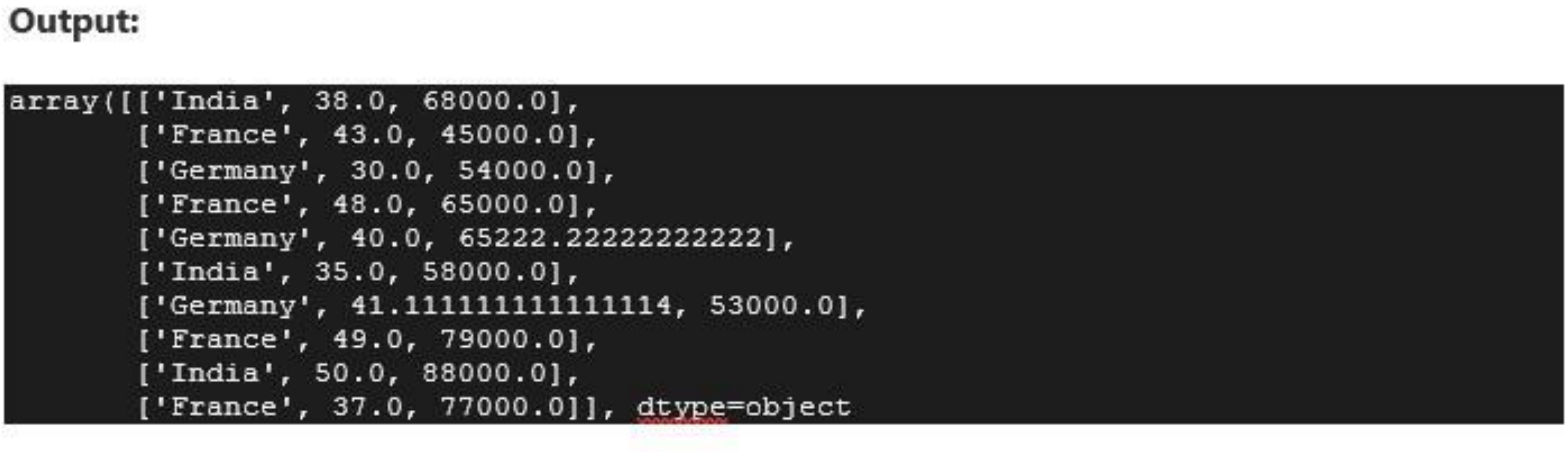
HANDLING MISSING DATA

To handle missing values, we will use **Scikit-learn** library in our code, which contains various libraries for building machine learning models. Here we will use **Imputer** class of **sklearn.preprocessing** library. Below is the code for it:

```
#handling missing data (Replacing missing data with the mean value)

#handling missing data (Replacing missing data with the mean value)

from sklearn.preprocessing import Imputer

imputer= Imputer(missing_values ='NaN', strategy='mean', axis = 0)

#Fitting imputer object to the independent variables x.

imputerimputer= imputer.fit(x[:, 1:3])

#Replacing missing data with the calculated mean value

x[:, 1:3]= imputer.transform(x[:, 1:3])
```

HANDLING MISSING DATA

- As we can see in the below output, the missing values have been replaced with the means of rest column values.



Output:

```
array([['India', 38.0, 68000.0],
       ['France', 43.0, 45000.0],
       ['Germany', 30.0, 54000.0],
       ['France', 48.0, 65000.0],
       ['Germany', 40.0, 65222.22222222222],
       ['India', 35.0, 58000.0],
       ['Germany', 41.111111111111114, 53000.0],
       ['France', 49.0, 79000.0],
       ['India', 50.0, 88000.0],
       ['France', 37.0, 77000.0]], dtype=object
```

5. ENCODING CATEGORICAL DATA

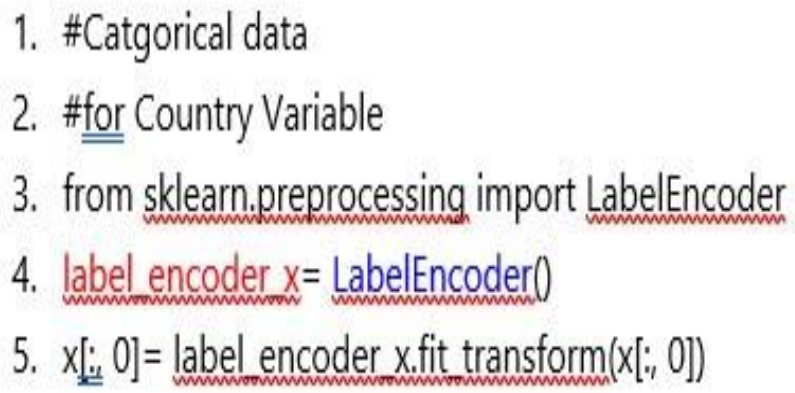
Categorical data is data which has some categories such as, in our dataset; there are two categorical variable, **Country**, and **Purchased**.

Since machine learning model completely works on mathematics and numbers, but if our dataset would have a categorical variable, then it may create trouble while building the model. So it is necessary to encode these categorical variables into numbers.

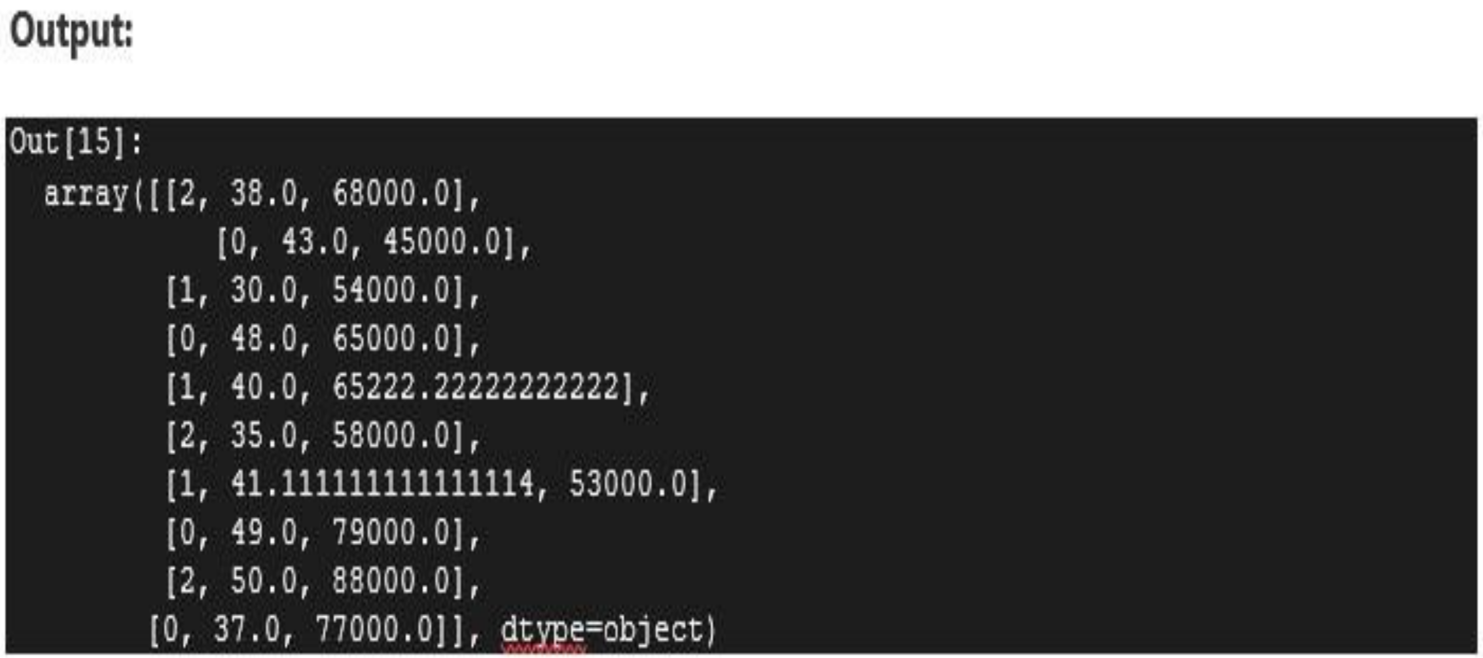
ENCODING CATEGORICAL DATA

- **For Country variable:**

- Firstly, we will convert the country variables into categorical data. So to do this, we will use **LabelEncoder()** class from **preprocessing** library.



```
#Catgorical data
#for Country Variable
from sklearn.preprocessing import LabelEncoder
label_encoder_x= LabelEncoder()
x[:, 0]= label_encoder_x.fit_transform(x[:, 0])
```



Output:

```
Out[15]:
  array([[2, 38.0, 68000.0],
             [0, 43.0, 45000.0],
          [1, 30.0, 54000.0],
          [0, 48.0, 65000.0],
          [1, 40.0, 65222.22222222222],
          [2, 35.0, 58000.0],
          [1, 41.111111111111114, 53000.0],
          [0, 49.0, 79000.0],
          [2, 50.0, 88000.0],
         [0, 37.0, 77000.0]], dtype=object)
```

ENCODING CATEGORICAL DATA

Explanation:

- In above code, we have imported LabelEncoder class of sklearn library. This class has successfully encoded the variables into digits.
- But in our case, there are three country variables, and as we can see in the above output, these variables are encoded into 0, 1, and 2. By these values, the machine learning model may assume that there is some correlation between these variables which will produce the wrong output. So to remove this issue, we will use dummy encoding.

ENCODING CATEGORICAL DATA

Dummy Variables:

Dummy variables are those variables which have values 0 or 1. The 1 value gives the presence of that variable in a particular column, and rest variables become 0. With dummy encoding, we will have a number of columns equal to the number of categories.

In our dataset, we have 3 categories so it will produce three columns having 0 and 1 values. For Dummy Encoding, we will use **OneHotEncoder** class of **preprocessing** library.

6. SPLITTING THE DATASET INTO THE TRAINING SET AND TEST SET

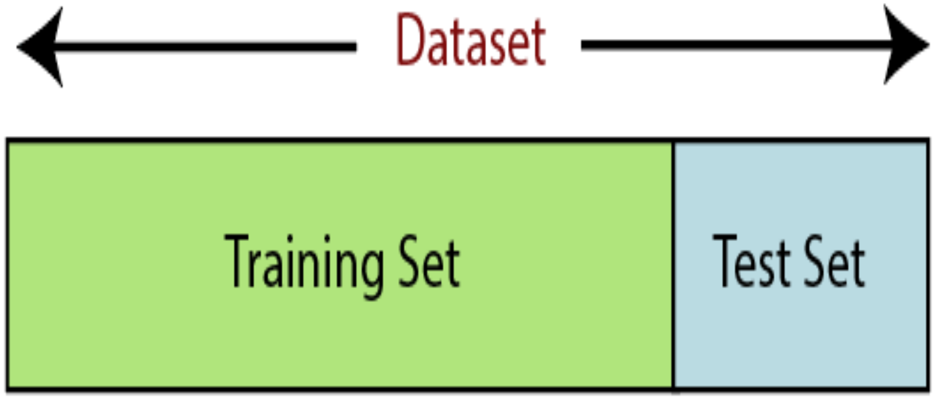
In machine learning data preprocessing, we divide our dataset into a training set and test set. This is one of the crucial steps of data preprocessing as by doing this, we can enhance the performance of our machine learning model.

Suppose, if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models.

If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance.

So we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:

TRAINING SET AND TEST SET



- **Training Set:** A subset of dataset to train the machine learning model, and we already know the output.

- **Test set:** A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.

- For splitting the dataset, we will use the below lines of code:

1. from sklearn.model_selection import train_test_split

2. x_train, x_test, y_train, y_test= train_test_split(x, y, test_siz e= 0.2, random_state=0)

TRAINING SET AND TEST SET

- **Explanation:**
  - In the above code, the first line is used for splitting arrays of the dataset into random train and test subsets.
  - In the second line, we have used four variables for our output that are
    - **x_train:** features for the training data
    - **x_test:** features for testing data
    - **y_train:** Dependent variables for training data
    - **y_test:** Independent variable for testing data
  - In **train_test_split() function**, we have passed four parameters in which first two are for arrays of data, and **test_size** is for specifying the size of the test set.
  - The test_size maybe .5, .3, or .2, which tells the dividing ratio of training and testing sets.
  - The last parameter **random_state** is used to set a seed for a random generator so that you always get the same result.

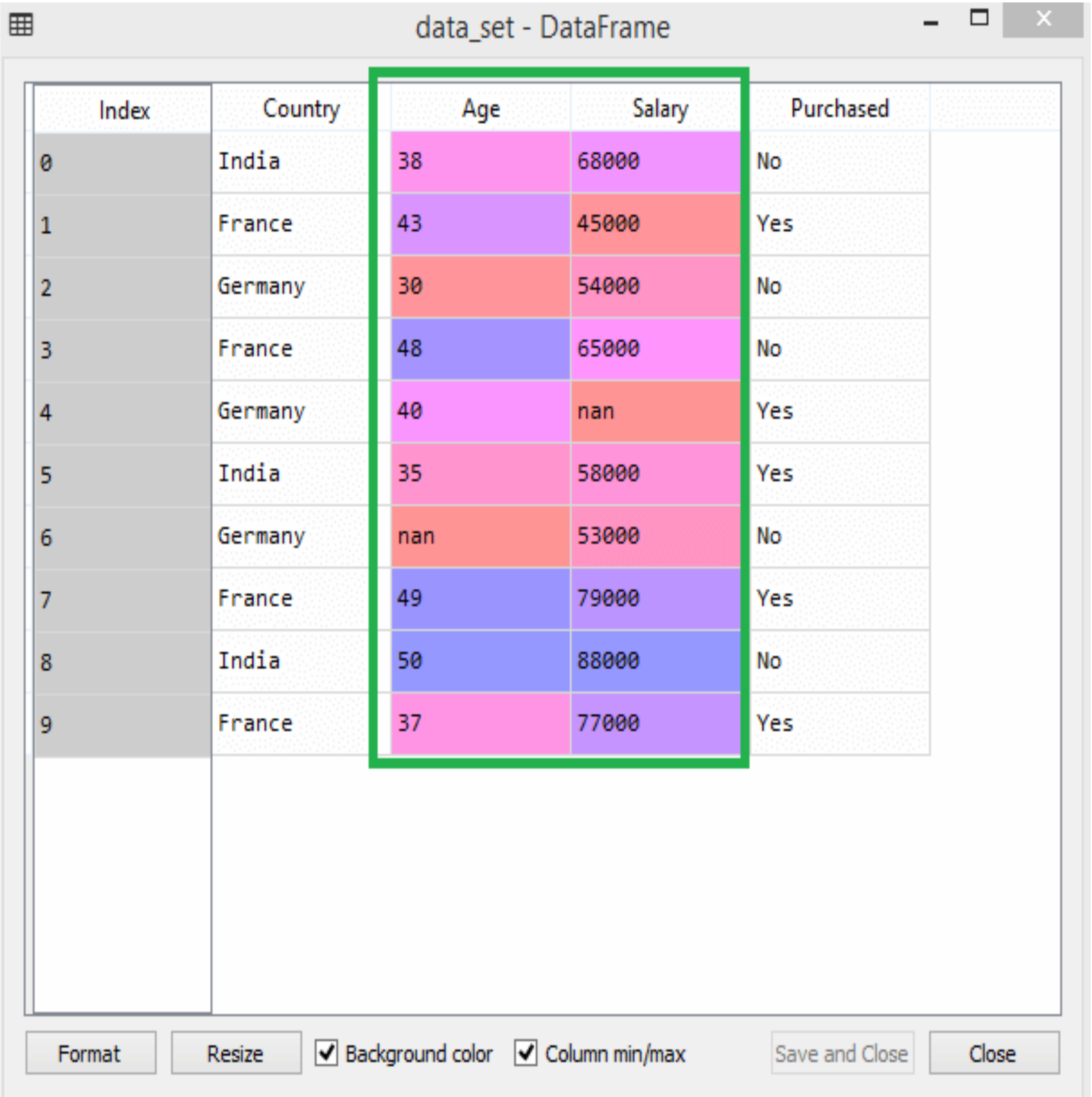
7. FEATURE SCALING

Feature scaling is the final step of data preprocessing in machine learning. It is a technique to standardize the independent variables of the dataset in a specific range.

In feature scaling, we put our variables in the same range and in the same scale so that no any variable dominate the other variable.

Consider the below dataset:

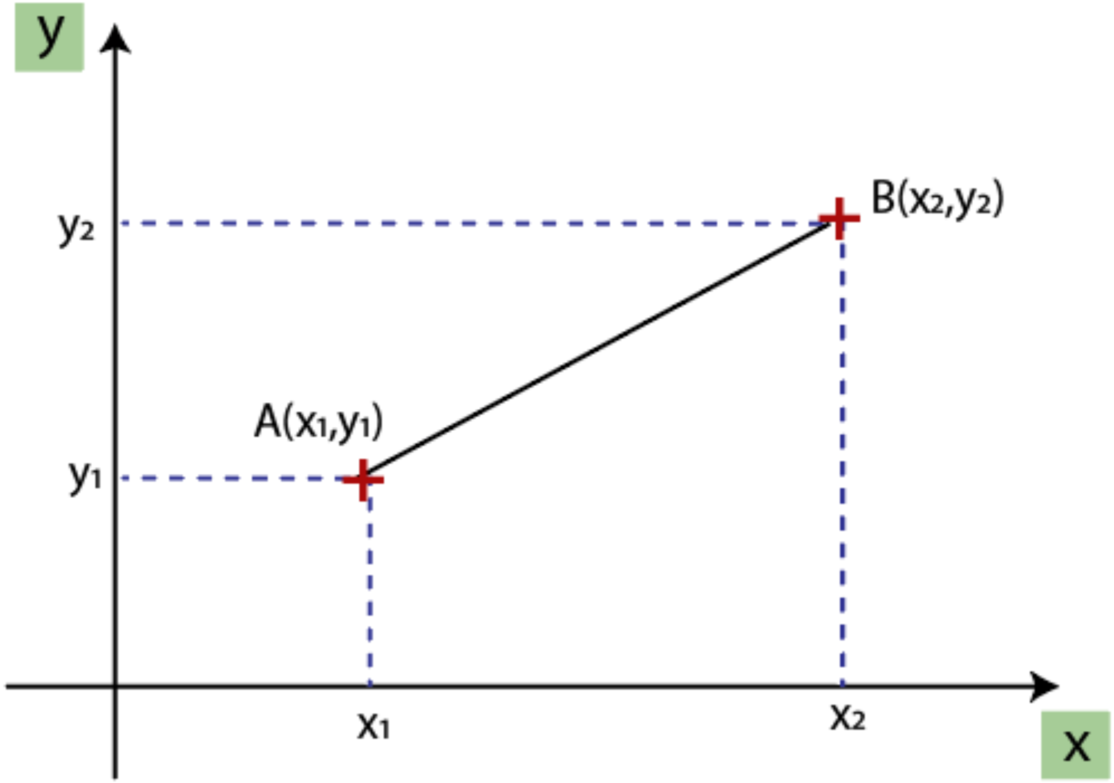
FEATURE SCALING



data_set - DataFrame

Index	Country	Age	Salary	Purchased
0	India	38	68000	No
1	France	43	45000	Yes
2	Germany	30	54000	No
3	France	48	65000	No
4	Germany	40	nan	Yes
5	India	35	58000	Yes
6	Germany	nan	53000	No
7	France	49	79000	Yes
8	India	50	88000	No
9	France	37	77000	Yes

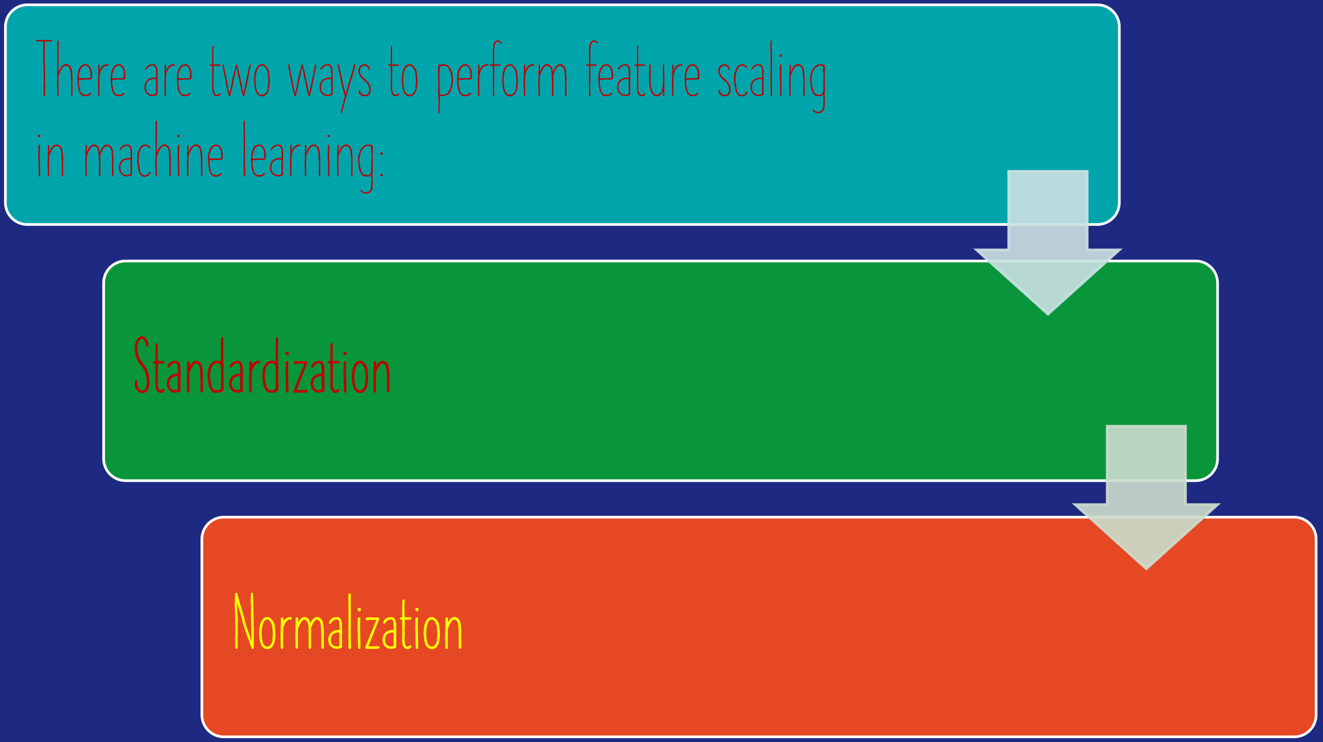
FEATURE SCALING



Euclidean Distance Between A and B=$\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

- As we can see, the age and salary column values are not on the same scale.
- A machine learning model is based on **Euclidean distance**, and if we do not scale the variable, then it will cause some issue in our machine learning model.
- Euclidean distance is given as:
- If we compute any two values from age and salary, then salary values will dominate the age values, and it will produce an incorrect result.
- So to remove this issue, we need to perform feature scaling for machine learning.

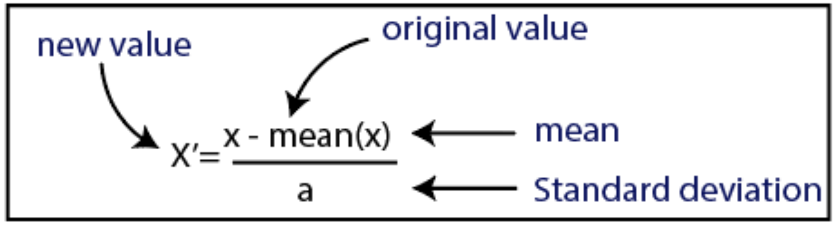
FEATURE SCALING



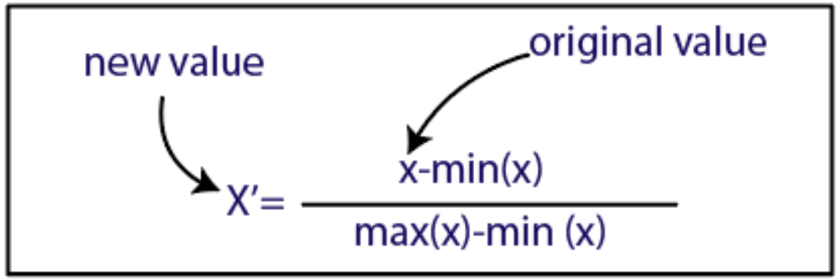
There are two ways to perform feature scaling in machine learning:

Standardization

Normalization



$$X' = \frac{x - \text{mean}(x)}{a}$$



$$X' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

THANK YOU