

# Machine learning Internship Day 1

Presented by Sakthiprasad.K.M

# Contents

---



DATA AND ITS  
PROCESSING



SUPERVISED  
LEARNING



UNSUPERVISED  
LEARNING



REINFORCEMENT  
LEARNING



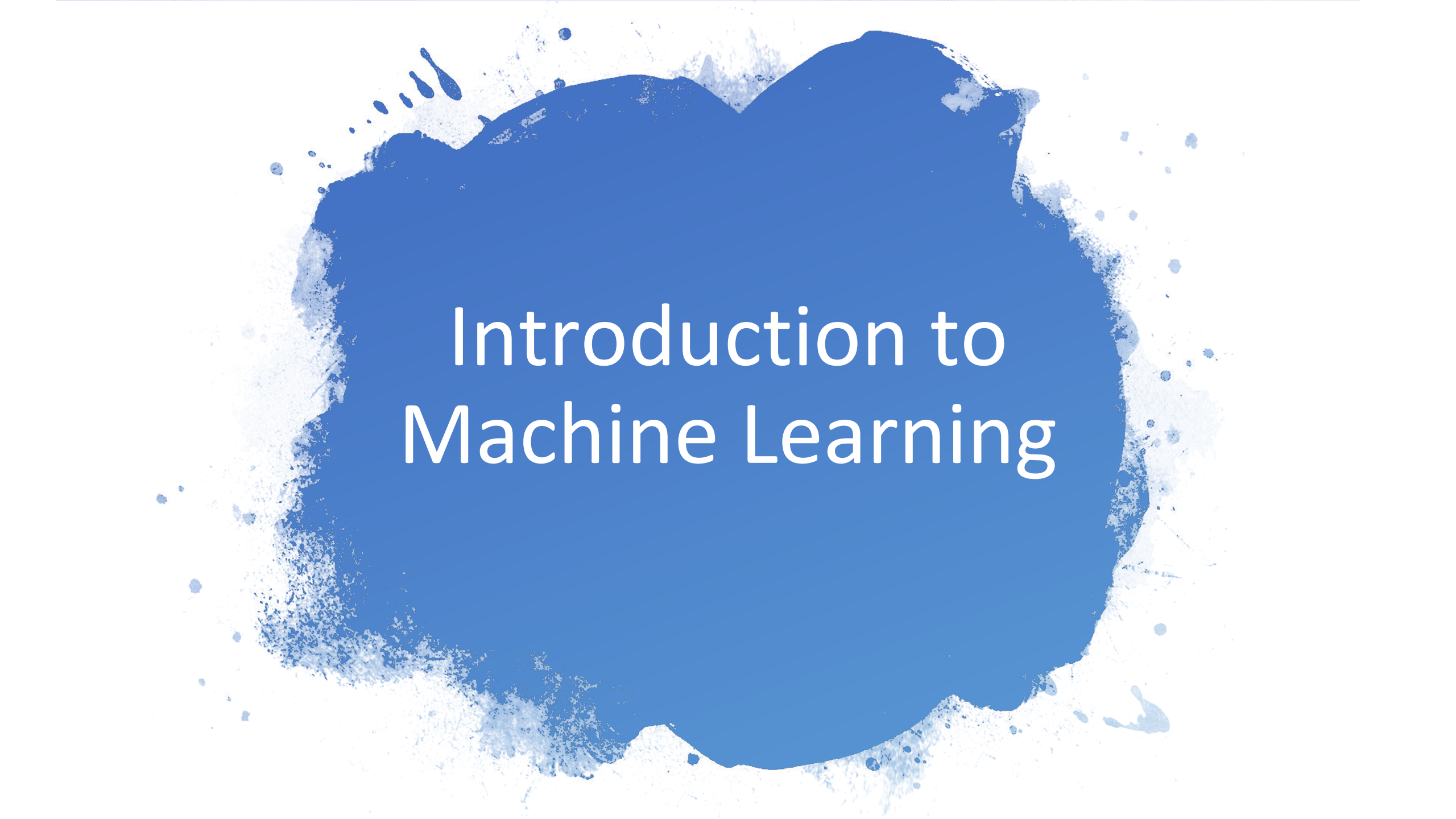
PYTHON FOR DATA  
SCIENCE



STRENGTH AND  
WEAKNESS OF  
PYTHON



PYTHON  
INSTALLATION

A large, irregular blue ink splash or watercolor blotch serves as the background for the text. The splash is centered and has a textured, painterly appearance with various shades of blue and white. The text is centered within the splash.

# Introduction to Machine Learning

# Machine learning

- Machine Learning (ML) is basically that field of **computer science** with the help of which computer systems can provide **sense to data** in much the same way as human beings do. In simple words
- ML is a type of **artificial intelligence** that extract patterns out of raw data by using an algorithm or method.
- The key focus of ML is to allow computer systems to **learn from experience without being explicitly programmed or human intervention**.
- When you tag a face in a **Facebook** photo, it is AI that is running behind the scenes and identifying faces in a picture. Face tagging is now omnipresent in several applications that display pictures with human faces
- There are several applications of AI that we use practically today. In fact, each one of us use AI in many parts of our lives, even without our knowledge. Today's AI can perform **extremely complex jobs** with a great accuracy and speed.

# Machine learning

- Some of the examples of statistical techniques that are used for developing AI applications **in those days and are still in practice** are listed here –
- **Regression** -Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the **relationship between one dependent variable** (usually denoted by Y) **and a series of other variables** (known as independent variables) Ex: amount of rain with respect to wind and humidity
- **Classification** - is an activity that consists of putting things into categories based on their **similarities or common criteria**. It allows humans to **organize things, objects, and ideas** that exist around them and simplify their understanding of the world Ex: Classification of the places based on the rainfall
- **Clustering** - Clustering is the task of dividing the population or data points into a number of groups such that data points in the **same groups are more similar** to other data points in the same group and dissimilar to the data points in other groups. Ex: divide the places based on the rainfall

# Machine Learning

- **classification** uses **predefined classes** in which objects are assigned, while **clustering** identifies **similarities between** objects, which it groups according to those characteristics in common and which differentiate them from other
- **Probability Theories** - Probability theory is the branch of mathematics concerned with probability. probability theory is the mathematical **study of uncertainty**. It plays a central role in machine learning, as the design of learning algorithms often relies on **proba-bilistic assumption** of the data. Ex : classification of the places based on the probability of raining in the next monsoon
- **Decision Trees** - A decision tree typically starts with a **single node**, which **branches into possible outcomes**. Each of those outcomes leads to additional nodes, which branch off into other possibilities Ex : prediction of rain fall based on dependent parameters.

# When & Why we need machine learning

## Lack of human expertise

- The very first scenario in which we want a machine to learn and take data-driven decisions, can be the domain where there is a lack of human expertise. The examples can be navigations in unknown territories or spatial planets.

## Dynamic scenarios

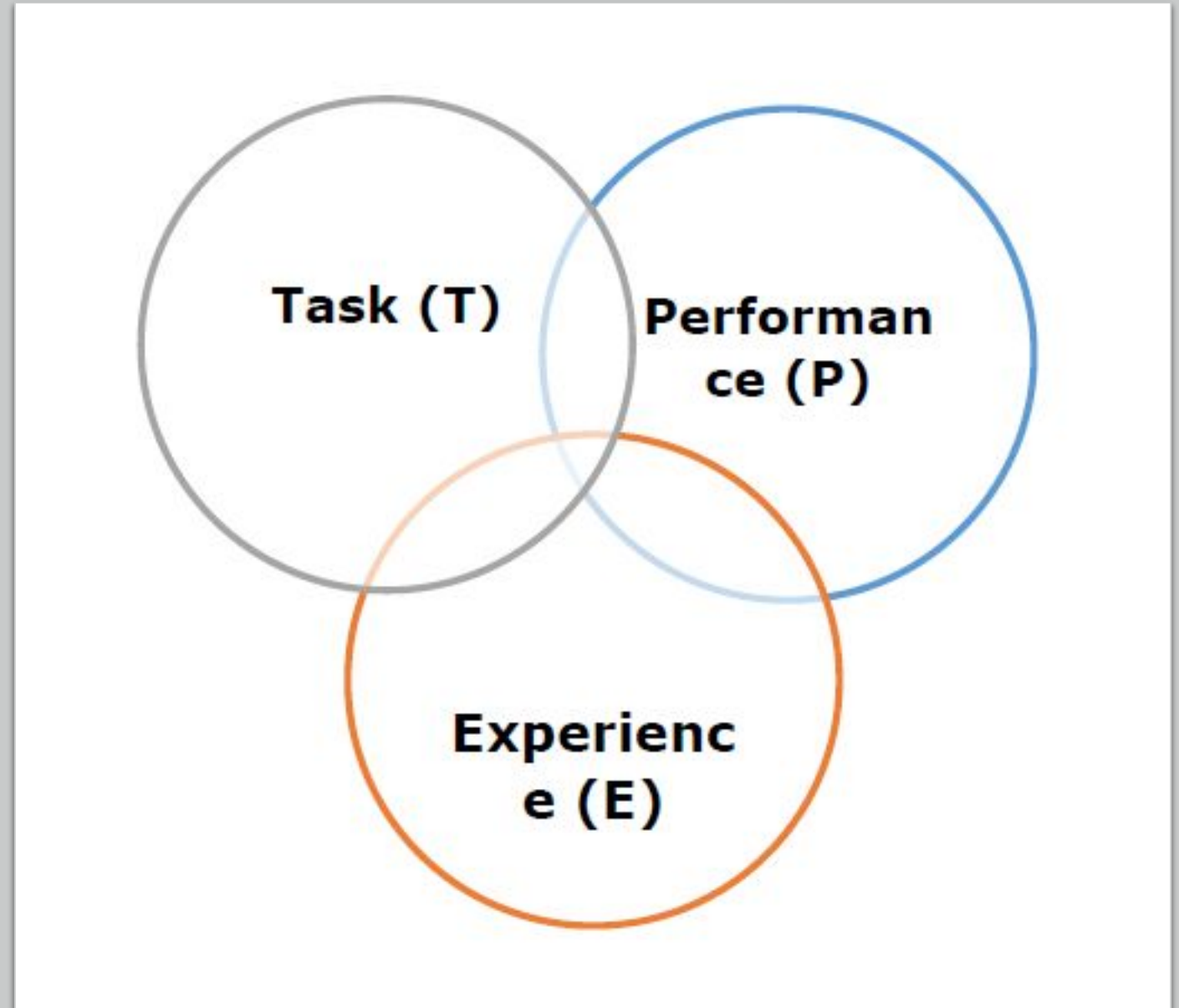
- There are some scenarios which are dynamic in nature i.e. **they keep changing over time**. In case of these scenarios and behaviors, we want a machine to learn and take data-driven decisions. Some of the examples can be network connectivity and availability of infrastructure in an organization.

## Difficulty in translating expertise into computational tasks

- There can be various domains in which humans have their expertise; however, they are unable to translate this expertise into computational tasks. In such circumstances we want machine learning. The examples can be the domains of speech recognition, cognitive tasks etc.

# Machine Learning Model

- ML is a field of AI consisting of learning algorithms that –
- Improve their performance (P)
- At executing some task (T)
- Over time with experience (E)





# Challenges in Machines Learning

- **Quality of data** – Having good-quality data for ML algorithms is one of the biggest challenges. Use of low-quality data leads to the problems related to data preprocessing and feature extraction.
- **Time-Consuming task** – Another challenge faced by ML models is the consumption of time especially for data acquisition, feature extraction and retrieval.
- **Lack of specialist persons** – As ML technology is still in its infancy stage, availability of expert resources is a tough job.
- **No clear objective for formulating business problems** – Having no clear objective and well-defined goal for business problems is another key challenge for ML because this technology is not that mature yet.
- **Issue of overfitting & underfitting** – If the model is overfitting or underfitting, it cannot be represented well for the problem.
- **Curse of dimensionality** – Another challenge ML model faces is too many features of data points. This can be a real hindrance.
- **Difficulty in deployment** – Complexity of the ML model makes it quite difficult to be deployed in real life.

# Applications of Machines Learning

- Emotion analysis
- Sentiment analysis
- Error detection and prevention
- Weather forecasting and prediction
- Stock market analysis and forecasting
- Speech synthesis
- Speech recognition
- Customer segmentation
- Object recognition
- Fraud detection
- Fraud prevention
- Recommendation of products to customer in online shopping



# Data and its processing

# Data and its processing

## Raw Data

The very first recipe is for looking at your raw data. It is important to look at raw data because the insight we will get after looking at raw data will boost our chances to better pre-processing as well as handling of data for ML projects.

## Checking Dimensions of Data

- It is always a good practice to know **how much data**, in terms of **rows and columns**, we are having for our ML project. The reasons behind are –
  - Suppose if we have too many rows and columns then it would take **long time to run** the algorithm and train the model.
  - Suppose if we have too less rows and columns then it we would **not have enough data** to well train the model.

## Getting Each Attribute's Data Type

- It is another good practice to know data type of each attribute. The reason behind is that, as per to the requirement, sometimes we may need **to convert one data type to another**.

# Data and its processing

## Statistical details

- Count
- Mean
- Standard Deviation
- Minimum Value
- Maximum value
- 25%
- Median i.e. 50%
- 75%

# Data and its processing

## Reviewing Class Distribution

- Class distribution statistics is useful in classification problems where we need to know the balance of class values. It is important to know class value distribution because if we have highly imbalanced class distribution

## Reviewing Correlation between Attributes

- The **relationship between two variables** is called correlation. In statistics, the most common method for calculating correlation is Pearson's Correlation Coefficient. It can have three values as follows –
  - Coefficient value = 1 – It represents full positive correlation between variables.
  - Coefficient value = -1 – It represents full negative correlation between variables.
  - Coefficient value = 0 – It represents no correlation at all between variables.

## Reviewing Skew of Attribute Distribution

- Skewness may be defined as the **distribution that is assumed to be Gaussian** but appears distorted or shifted in one direction or another, or either to the left or right. Reviewing the skewness of attributes is one of the important tasks due to following reasons –
- Presence of skewness in data requires the correction at data preparation stage so that we can get more accuracy from our model.
- Most of the ML algorithms assumes that data has a Gaussian distribution i.e. either normal or bell curved data.

# Data pre-processing

- data preprocessing will convert the **selected data into a form we can work with or can feed to ML algorithms**. We always need to preprocess our data so that it can be as per the expectation of machine learning algorithm

## Scaling

- Most probably our dataset comprises of the attributes with varying scale, but we cannot provide such data to ML algorithm hence it requires rescaling. Data rescaling makes sure that attributes are at **same scale**. Generally, attributes are rescaled into the **range of 0 and 1**.

## Normalization

- Another useful data preprocessing technique is Normalization. This is used to **rescale each row of data to have a length of 1**.

# Data pre-processing

## Binarization

- As the name suggests, this is the technique with the help of which we can **make our data binary**. We can use a binary threshold for making our data binary. The values above that threshold value will be converted to 1 and below that threshold will be converted to 0.

## Standardization

- Another useful data preprocessing technique which is basically used to transform the data attributes with a **Gaussian distribution**. It differs the mean and SD (Standard Deviation) to a standard Gaussian distribution with a mean of 0 and a SD of 1

## Data Labeling

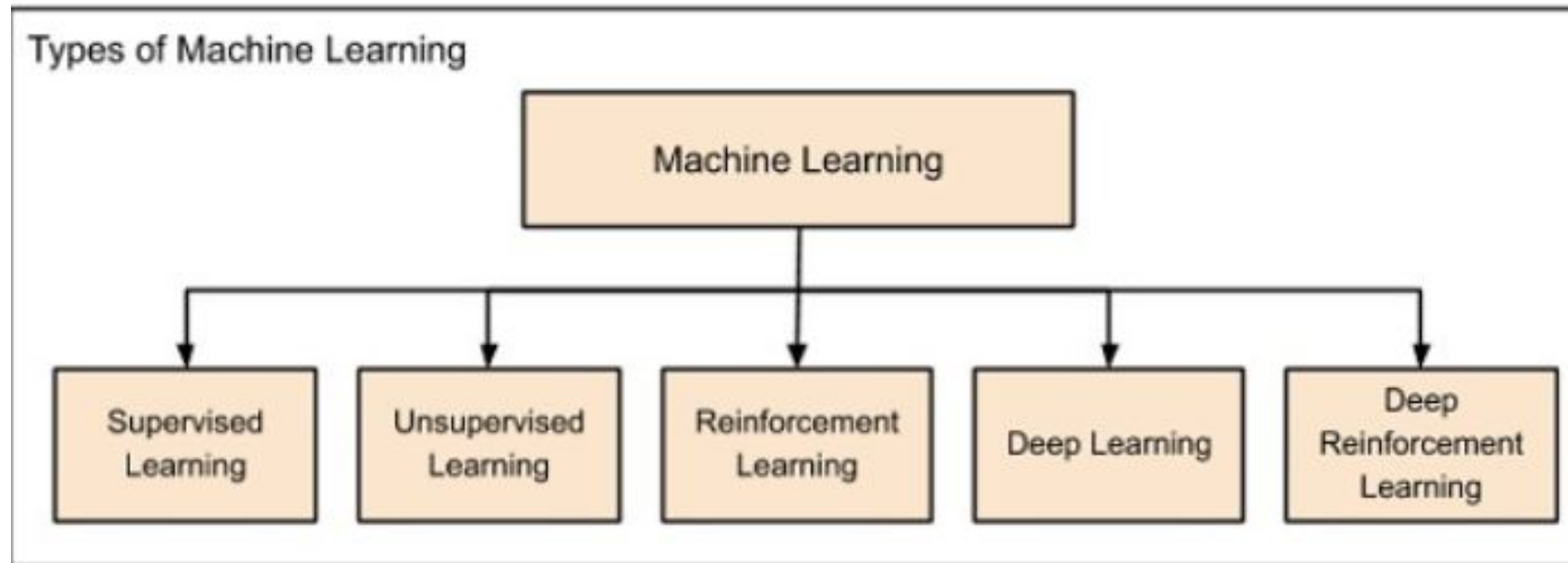
- We discussed the importance of good data for ML algorithms as well as some techniques to pre-process the data before sending it to ML algorithms. One more aspect in this regard is data labeling. It is also very important to send the data to ML algorithms having **proper labeling**





# Types of Machine Learning

# Types of Machine learning



# Supervised learning

- Supervised learning is the machine learning task of learning a function that maps an input to an output **based on example input-output pairs**. It infers a function from **labeled training data** consisting of a set of training examples

## Regression

- Similarly, in the case of supervised learning, you give **concrete known examples** to the computer. You say that for given feature value  $x_1$  the output is  $y_1$ , for  $x_2$  it is  $y_2$ , for  $x_3$  it is  $y_3$ , and so on. Based on this data, you let the computer figure out an **empirical relationship** between  $x$  and  $y$ .

## Classification

- You may also use machine learning techniques for classification problems. In classification problems, you **classify objects of similar nature** into a single group.

# Unsupervised Learning

- In unsupervised learning, we do not specify a target variable to the machine, rather we ask machine

## Clustering

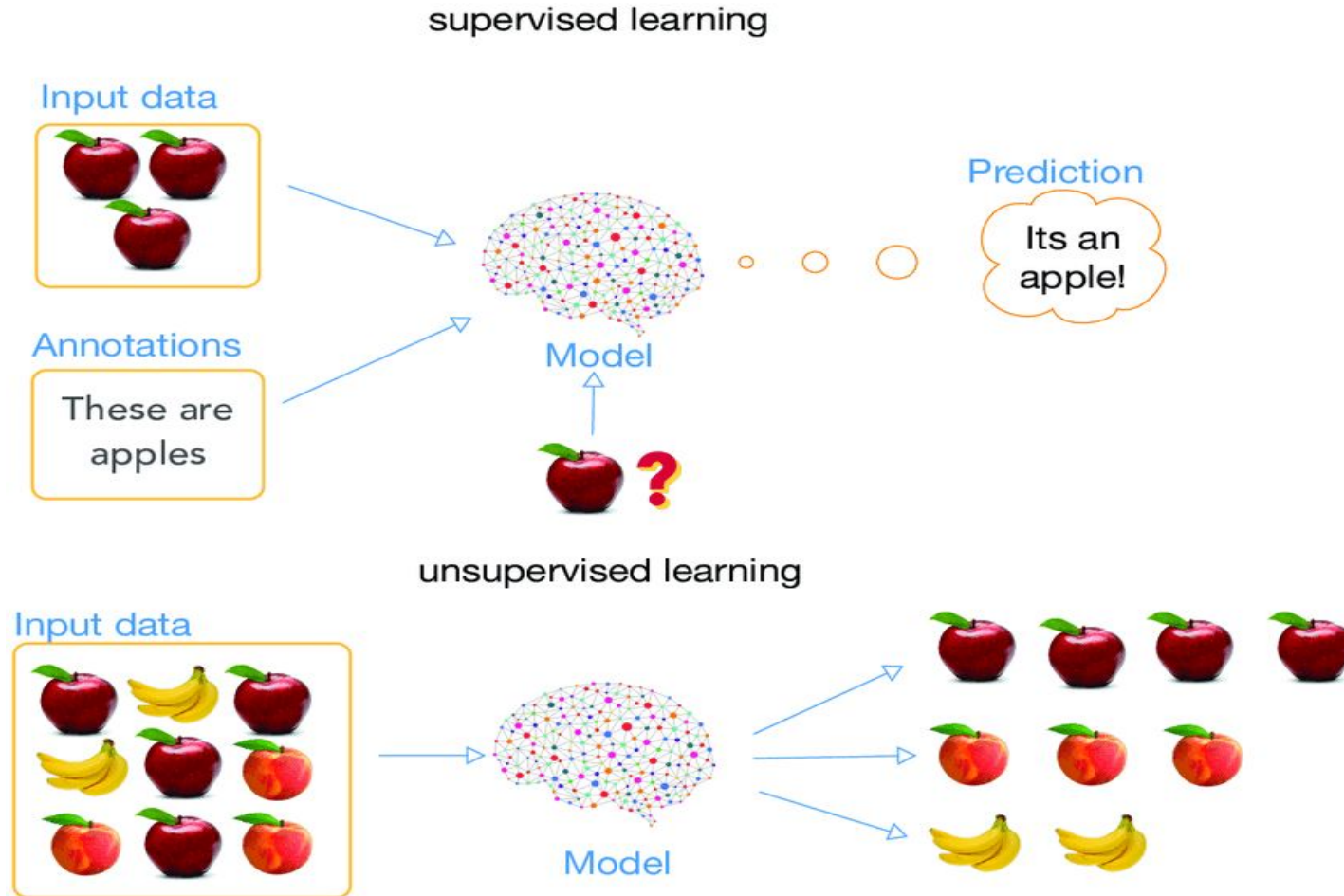
A clustering problem is where you want to **discover the inherent groupings** in the data, such as grouping customers by purchasing behavior.

## Association

An association **rule learning** problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

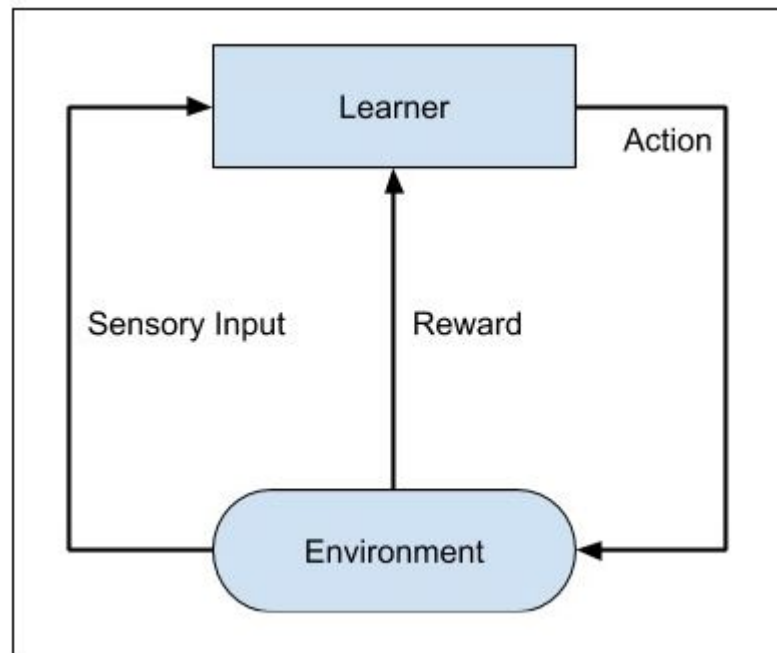
Association learning is a **rule based machine** learning and data mining technique that finds important relations between variables or features in a data set

# Supervised Vs Unsupervised



# Reinforcement learning

- Reinforcement learning (RL) is an area of machine learning concerned with how software agents ought to take actions in an environment in order to **maximize the notion of cumulative reward**.
- Reinforcement learning is one of three basic machine learning paradigms, alongside supervised learning and unsupervised learning.





# Python for Machine Learning

# Overview

- **Python is Interpreted** – Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.
- **Python is Interactive** – You can actually sit at a Python prompt and interact with the interpreter directly to write your programs.
- **Python is Object-Oriented** – Python supports Object-Oriented style or technique of programming that encapsulates code within objects.
  - OOP stands for Object-Oriented Programming. **Procedural programming** is about writing procedures or functions that **perform operations on the data**, while **object-oriented programming** is about **creating objects that contain both data and functions**.
- **Python is a Beginner's Language** – Python is a great language for the beginner-level programmers and **supports the development of a wide range of applications** from simple text processing to WWW browsers to games.



# Features

- **Easy-to-learn** – Python has **few keywords, simple structure**, and a **clearly defined syntax**. This allows the student to pick up the language quickly.
- **Easy-to-read** – Python code is more clearly defined and visible to the eyes.
- **Easy-to-maintain** – Python's source code is fairly easy-to-maintain.
- **A broad standard library** – Python's bulk of the library is very portable and **cross-platform compatible** on UNIX, Windows, and Macintosh.
- **Interactive Mode** – Python has support for an interactive mode which allows **interactive testing and debugging** of snippets of code.
- **Portable** – Python can run on a wide variety of **hardware platforms** and has the same interface on all platforms.
- **Extendable** – You can add low-level **modules** to the Python interpreter. These modules enable programmers to add to or customize their tools to be more efficient.
- **Databases** – Python **provides interfaces to all major commercial databases**.
- **GUI Programming** – Python supports **GUI applications** that can be created and ported to many system calls, libraries and windows systems, such as Windows MFC, Macintosh, and the X Window system of Unix.
- **Scalable** – Python provides a better structure and support for **large programs** than shell scripting.

# Strengths and Weaknesses of Python

- Every programming language has some strengths as well as weaknesses, so does Python too.

## Strengths

- According to studies and surveys, Python is the **fifth most important language as well as the most popular language for machine learning and data science**. It is because of the following strengths that Python has –
- **Easy to learn and understand** – The syntax of Python is simpler; hence it is relatively easy, even for beginners also, to learn and understand the language.
- **Multi-purpose language** – Python is a multi-purpose programming language because it **supports structured programming, object-oriented programming as well as functional programming**.
- **Huge number of modules** – Python has **huge number of modules for covering every aspect of programming**. These modules are easily available for use hence making Python an extensible language.
- **Support of open source community** – As being open source programming language, **Python is supported by a very large developer community**. Due to this, the bugs are easily fixed by the Python community. This characteristic makes Python very robust and adaptive.
- **Scalability** – Python is a scalable programming language because it provides an **improved structure for supporting large programs than shell-scripts**.

## Weakness

- Although Python is a popular and powerful programming language, it has its own weakness of **slow execution speed**.
- The execution speed of Python is slow **as compared to compiled languages** because Python is an interpreted language. This can be the major area of improvement for Python community.

# Python for data science

- There are two main factors that make Python a widely-used programming language in scientific computing, in particular:
- The stunning ecosystem
- a great number of **data-oriented feature packages** that can speed up and simplify data processing, making it time-saving.

## What Makes Python a Fantastic Option for Data Analysis?

### Easy to Learn

- Being involved in development for web services, mobile apps, or coding, you have a notion that Python is widely recognized thanks to its clear syntax and readability

### Well-Supported

- Having the experience of using some tools for free, you probably know that it is a challenge to get decent support.

### Flexibility

- The cool options don't end there. So, let's observe another reason why Python is really a fantastic option for data processing. Another strong feature of the language is the hyper flexibility that makes Python highly requested among data scientists and analysts.

### Scalability

- This Python's feature is described right after the flexibility, not by accident, but because it is closely connected with the previous option. Comparing with other languages like R, Go, and Rust, Python is much faster and more scalable

# Python for data science

## Huge Libraries Collection

- As we have already mentioned, Python is one of **the most supported languages** nowadays. It has a long list of totally **free libraries** available for all the users. That's a key factor that gives a strong push for Python at all, and in the data science, too

## Exceeding Python Community

- It's a kind of open-source language. That means you get at least two strong advantages. **Python is free**, plus **it employs a community-based model for development**.

## Graphics and Visualization Tools

- It's a well-known fact that **visual information is much easier to understand**, operate, and remember. There is a **pack of diverse visualization** options available. That makes Python a must-have tool not only for data analysis but for all data science.

## Extended Pack of Analytics Tools Available

- Straight after you gather data, you're **to handle** it. Python suits this purpose supremely well. So, seeking for the perfect tool for complex data processing or **self-service analytics**, **Python's built-in data analytics tools**. Dozens of data mining companies over the globe utilize Python to reduce data.

Python is the **internationally acclaimed programming language** to help in handling your data in a better manner for a variety of causes.



# Installation

# Google colab

