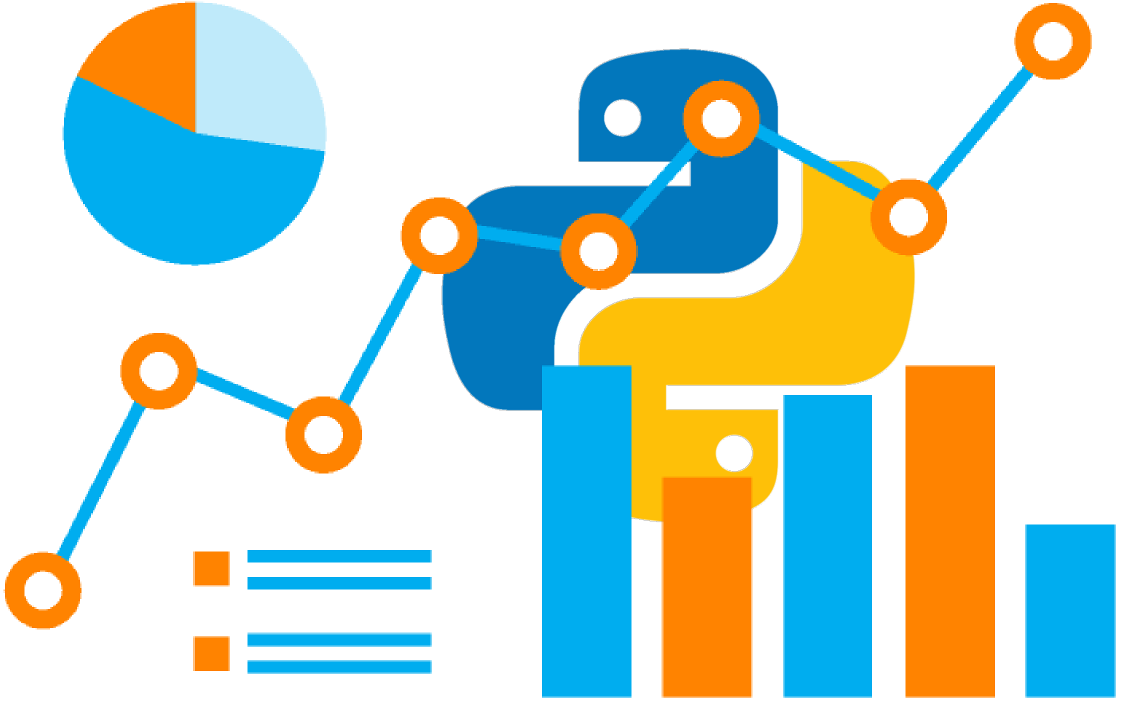
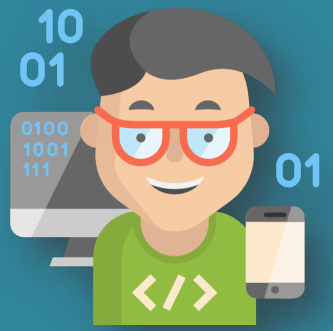
greatlearning
Learning for Life





Statistics For Data Science

Sampriti Chatterjee (Great Learning)

Agenda

greatlearning
Learning for Life

1 Why do we need data science?

2 What is Data science?

3 Life cycle of Data science

4 Important statistics terms in data science

5 Install python

6 Python Library: Numpy and Pandas

7 Data manipulation using Numpy and Pandas

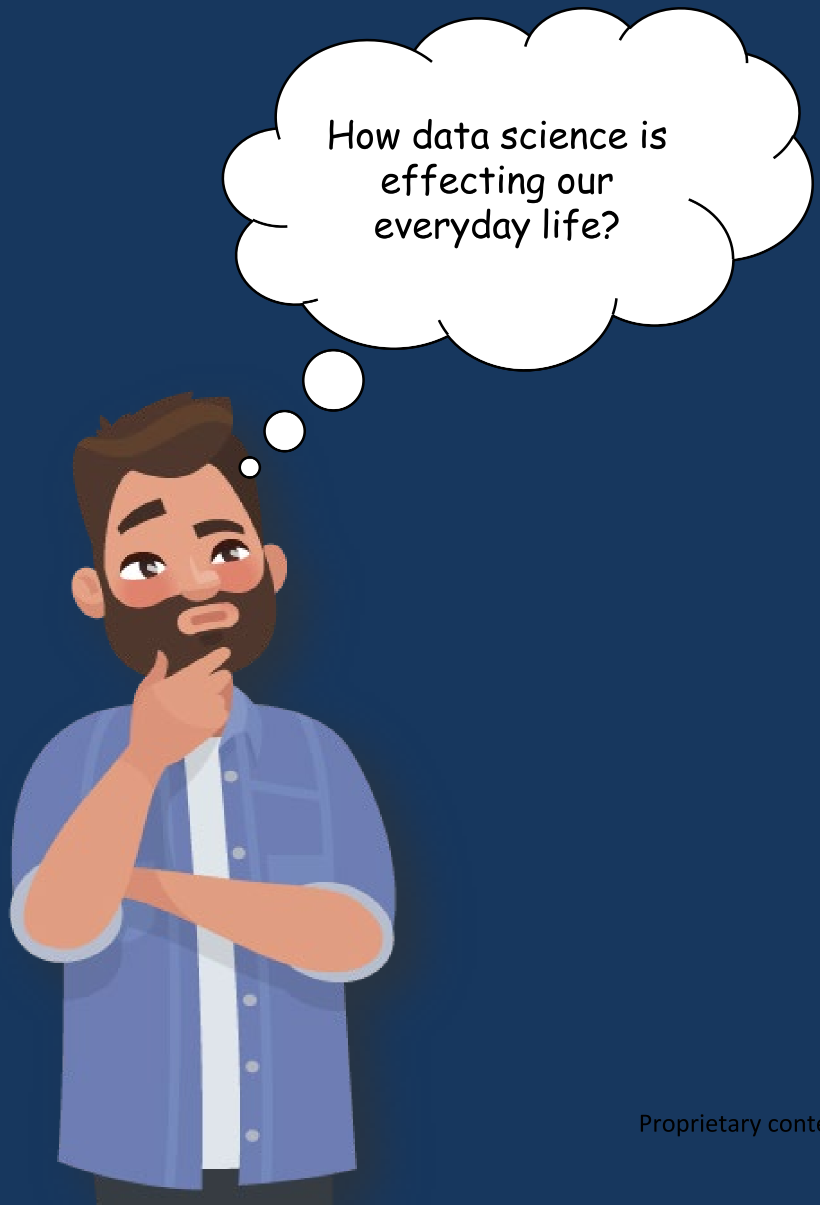
8 Data visualization with seaborn and Matplotlib

9 What is machine Learning?

10 Supervised Learning: Logistic Regression

11 Diabetes prediction using Python

Why do we need Data Science?



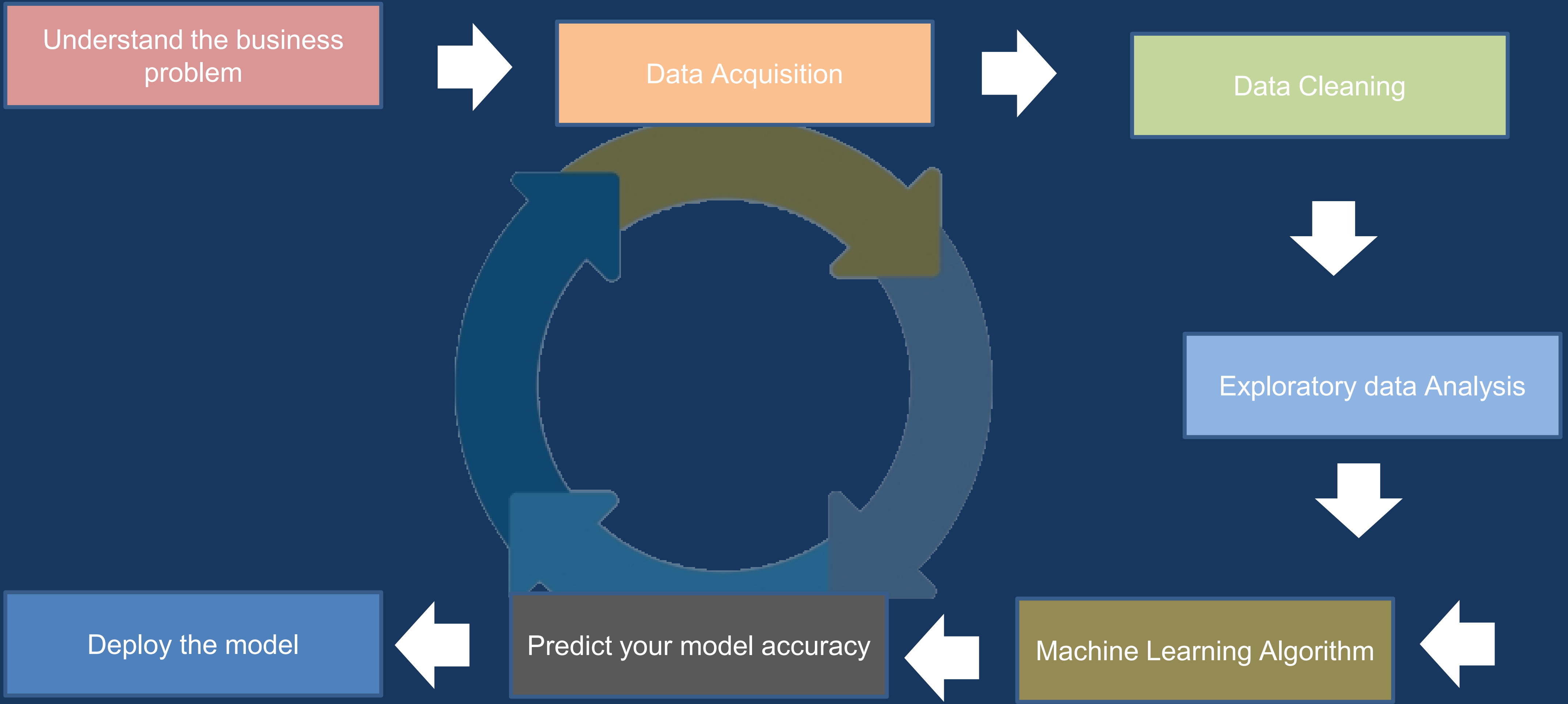
- In the past, we used to have data in a structured format but now as the volume of the data is increasing, so the number of structured data becomes very less, so to handle the massive amount of data we need data science techniques
- Those data can be used to get the proper business insights and the hidden trends from them.
- These insights helps the organization to predict the Future
- Using data science decision making can be faster and effective
- Helps to reduce the production cost
- Build model based on the data to give the ability to the machine to predicts on its own

What is Data Science?

greatlearning
Learning for Life

Data science is a process to get some meaningful information from the massive amount of data. In simple terms, read and study the data to get proper intuitive insights. Data Science is a mixture of various tools, algorithms, and machine learning and deep learning concepts to discover hidden patterns from the raw and unstructured data

Life cycle of Data Science?



Understand the business problem

Data Acquisition

Data Cleaning

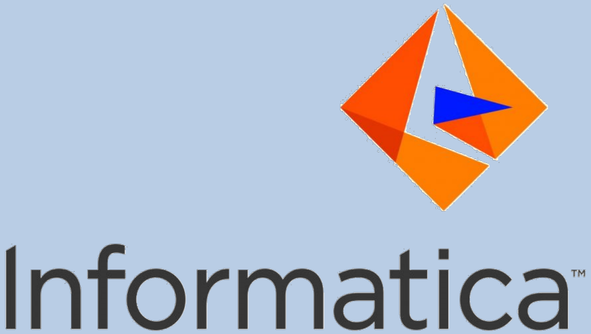
Exploratory data Analysis

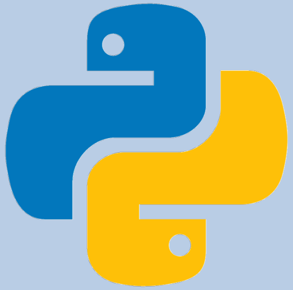
Machine Learning Algorithm

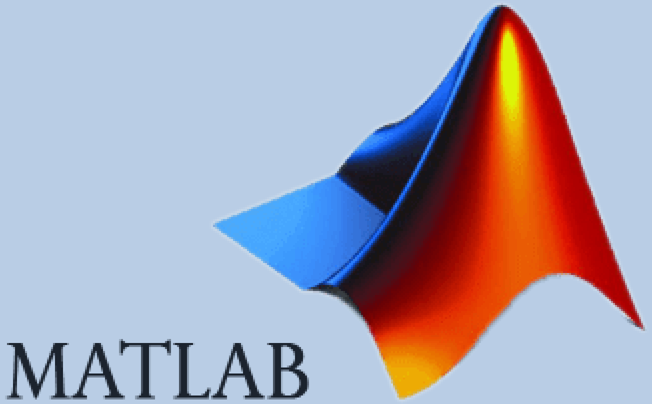
Predict your model accuracy

Deploy the model

Most Popular Programming Languages For Data Science?







Important statistics terms in data science

greatlearning
Learning for Life

1. What is Statistics?
2. What is population?
3. What is parameter?
4. What is sample?
5. What is mean?
6. Types of analysis in statistics
7. What is Outlier?
8. What is Interquartile Range IQR?
9. What is upper and lower limits in interquartile range
10. What is null hypothesis?
11. What is p value?

greatlearning
Learning for Life

What is Statistics?

Statistics is a part of integrated applied mathematics which deals with data

1 It helps to collect data and analyze them properly

2 With the help of statistics we can read the data and organize them in order to get the hidden information from them

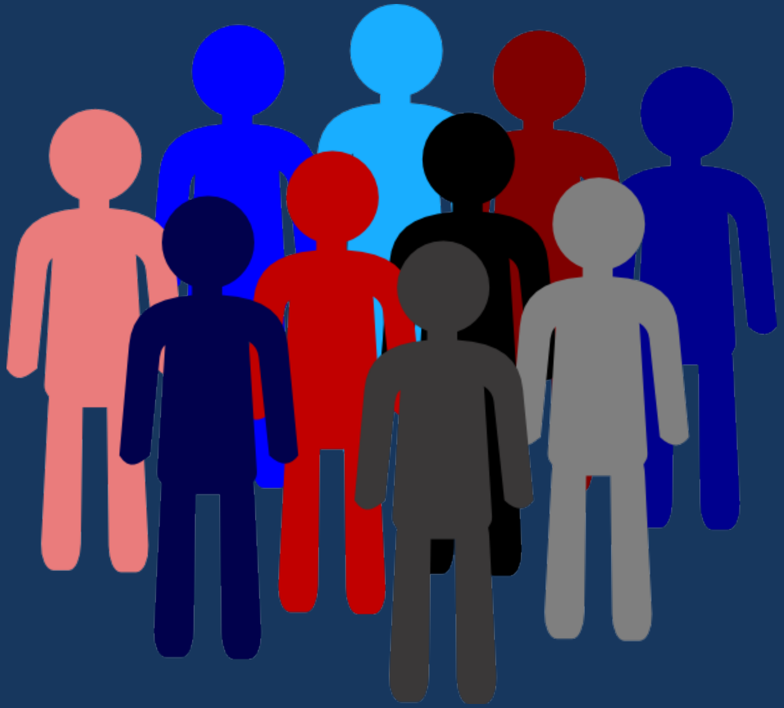
3 In data science domain statistics concepts are used to process the complex data to get the insights from them using mathematical computations

What is Population?

Population terms in statistics use to refer the total set of observations

Example:

Suppose,
If we want to study a diabetes dataset to understand the symptoms and the other factors then the whole dataset is referred as population



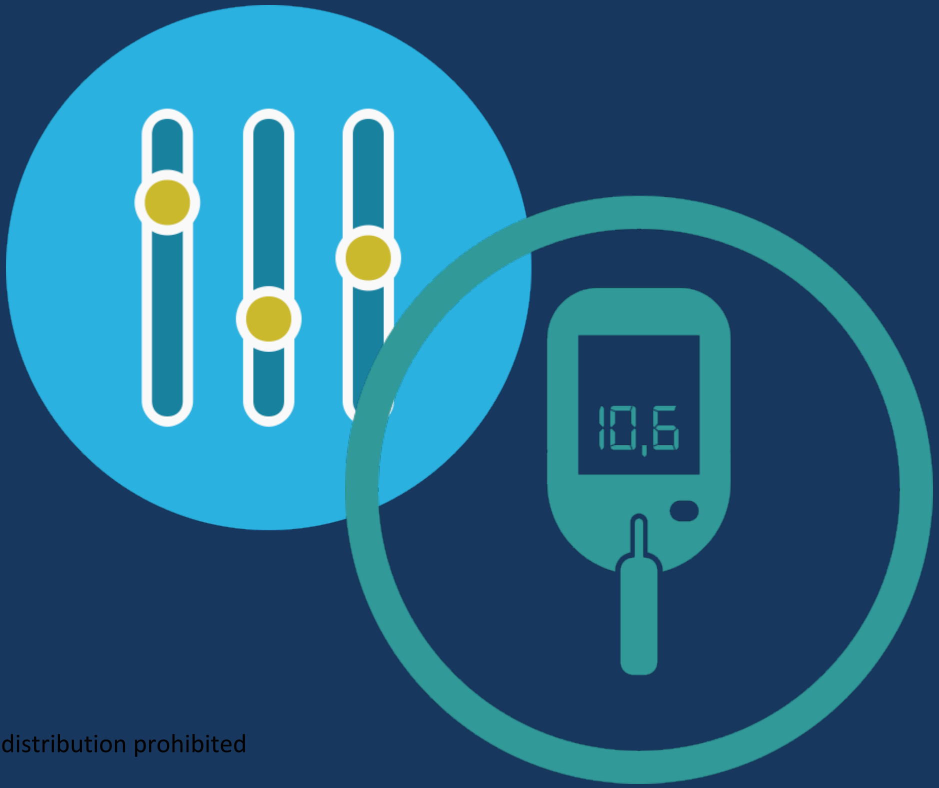
What is Parameter?

Parameters are referred to characteristics which describes the population

1 Parameters are like average or percentage which helps to describe the entire population

2 Mean and the standard deviation are two common parameters of population

3 Example: Average age for being diabetic is the parameter for whole diabetes dataset population

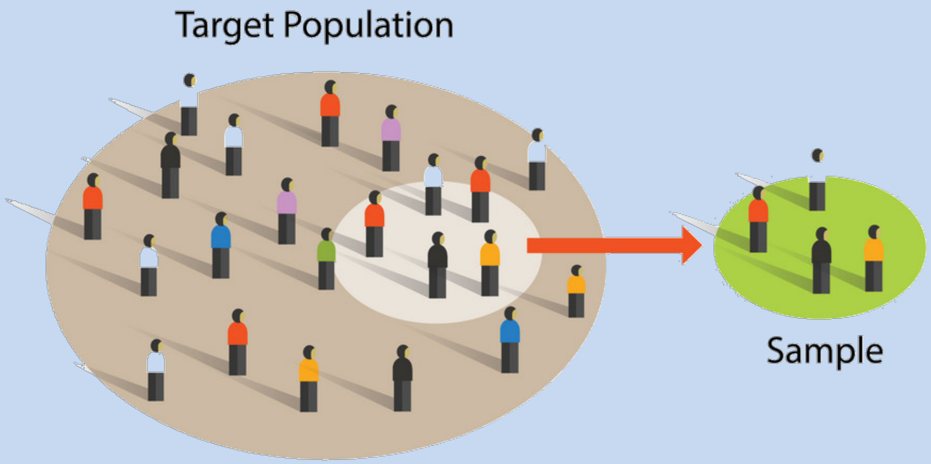


What is Sample?

Sample is basically a small part or portion of the large population

Example:

Suppose,
From the whole diabetes dataset you picked 100 rows of information to do the analysis, that 100 rows of information will be referred as **Sample**



Important statistics terms in data science

greatlearning
Learning for Life

What is Mean?

Mean term referred as average value of the whole population

What is Median?

Median is the middle value of the data when your data is sorted in manner

What is Mode?

Mode stands for the most occurring element in the dataset

Types Of Analysis In Statistics

Descriptive statistics

Inferential Statistics

It helps to describe the data in mathematical or graphical way

Inferential statistics split the data into samples and applies probability to arrive to theconclusion

What is Outlier?

Outliers in the dataset are referred as unusual value which can distort and violate statistical analysis

1. Outliers are basically experimental errors in the data
2. Some outliers are good for the dataset to detect anomaly like: detecting fraud transaction
3. It effects the mean and the standard deviation of the data and most of the machine learning technique does not perform good with outliers

What is Interquartile Range IQR?

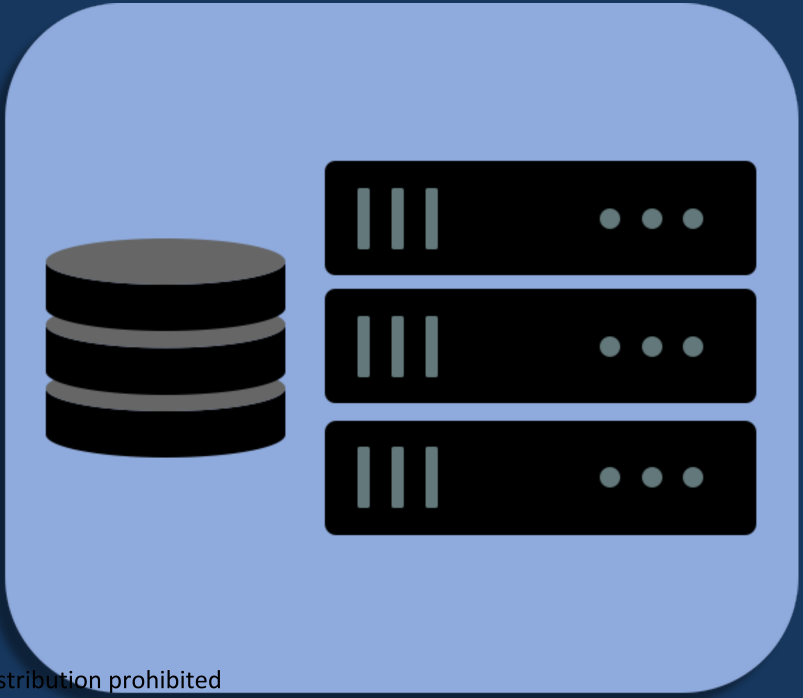
Interquartile range divides the dataset into quartiles to measure the variability and the spread of the dataset

1 Splits the data into 4 equal part in sorted manner

2 Q1, Q2, Q3 are called first, second and third quartiles:

- Q1 → 25th percentile of the dataset
- Q2→ 50th percentile of the dataset
- Q3→ 75th percentile of the dataset

Formula: IQR→ Q3 – Q1



What is upper and lower limits in interquartile range

Lower and upper limit in the interquartile basically the range where data points lie

1 **Formula to find the lower limit:**

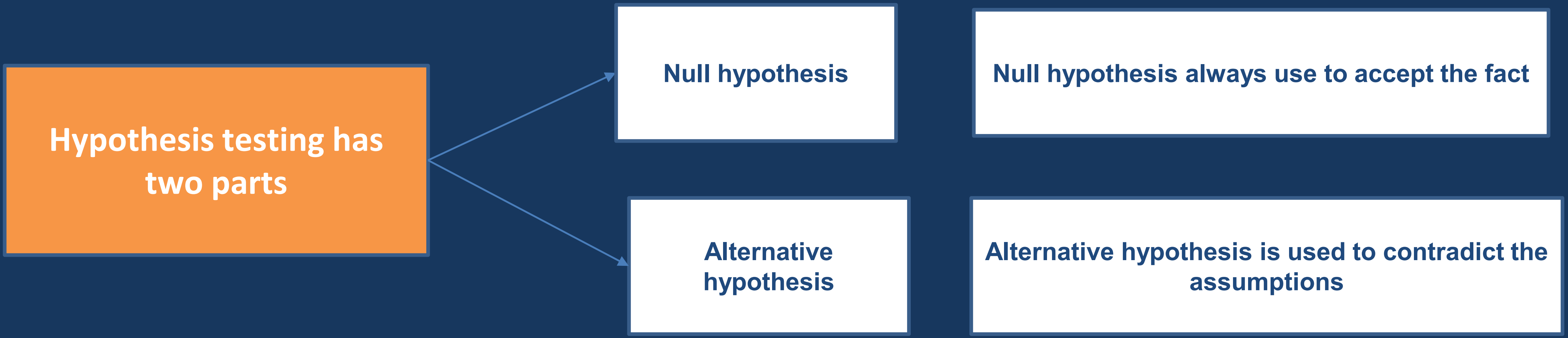
Lower_limit = Q1 - 1.5 IQR

2 **Formula to find the upper limit:**

Upper_limit = Q3 + 1.5 * IQR

What is Hypothesis testing?

Hypothesis testing is basically used to test the assumption which is taken based on observations and experiments



Hypothesis testing has two parts

Null hypothesis

Null hypothesis always use to accept the fact

Alternative hypothesis

Alternative hypothesis is used to contradict the assumptions

What is p value?

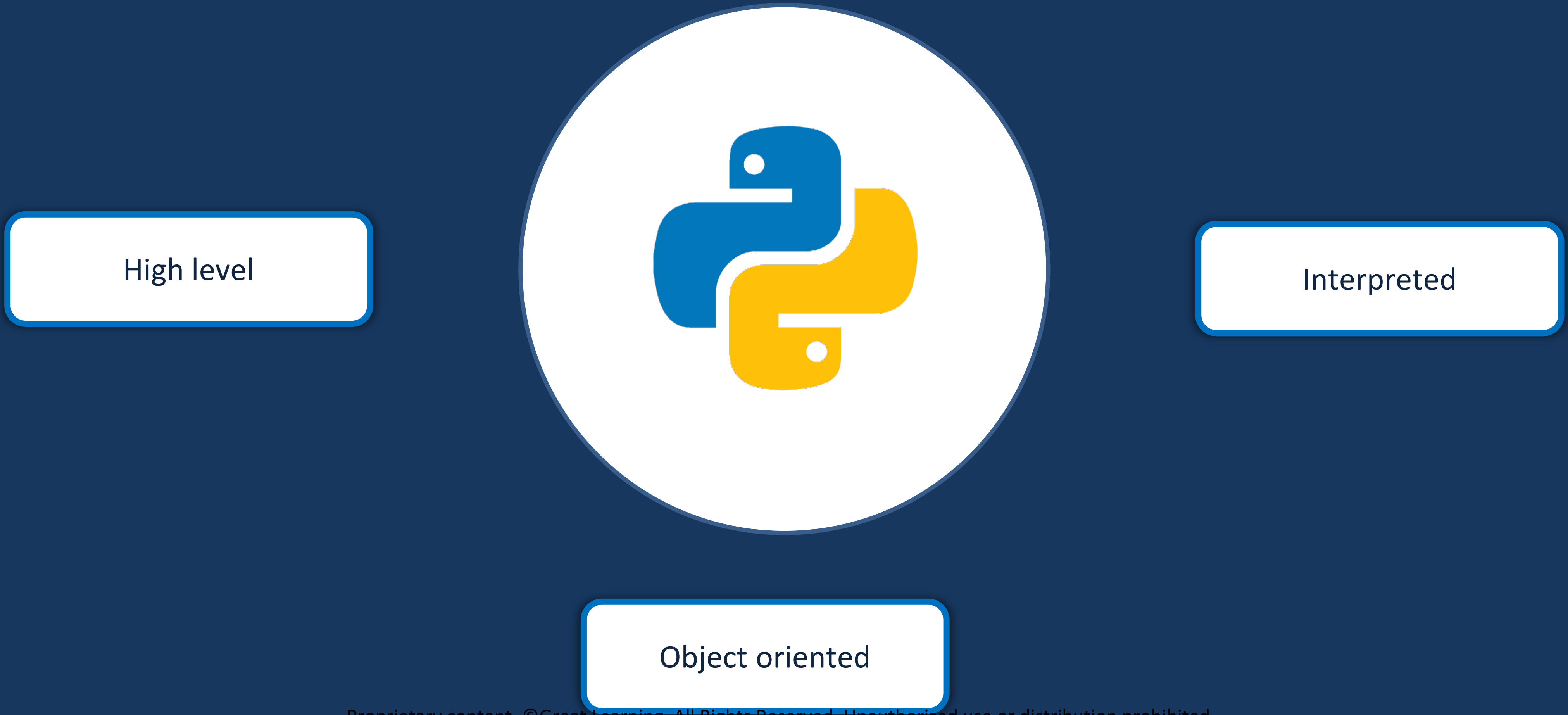
p value is used to support or reject the null hypothesis or the assumption

1 P value is basically the strong evidence to reject the null hypothesis

2 If p value is less than 0.05 then we accept the null hypothesis

Introduction to Python

Python is a popular high level, object oriented and interpreted language

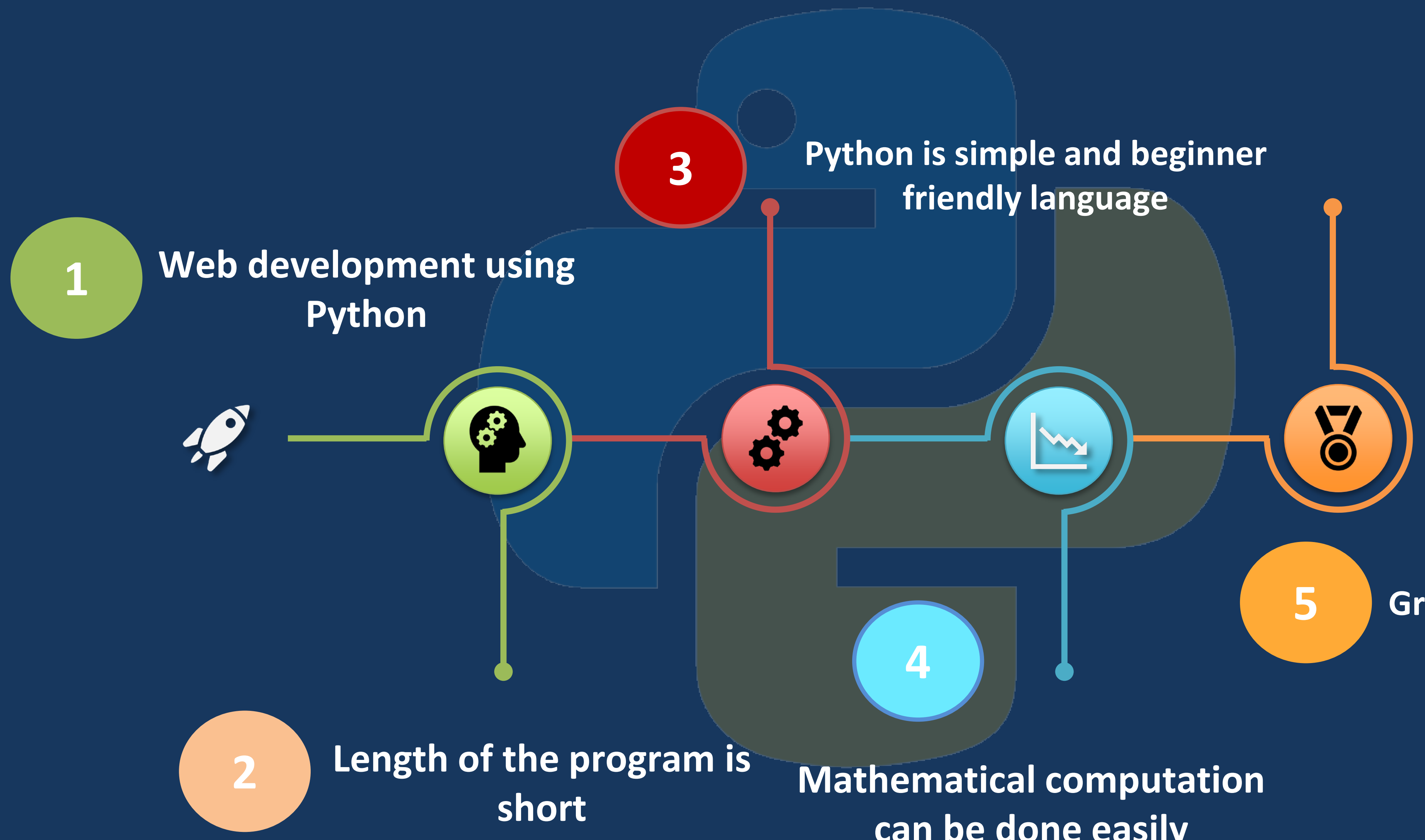


High level

Interpreted

Object oriented

Why should you learn Python?



1 Web development using Python

2 Length of the program is short

3 Python is simple and beginner friendly language

4 Mathematical computation can be done easily

5 Graphical user interface

Why Python is so popular?

1. Largest community for Learners and Collaborators

2. **Open source**

3. **Easy to learn and usable flexibility**

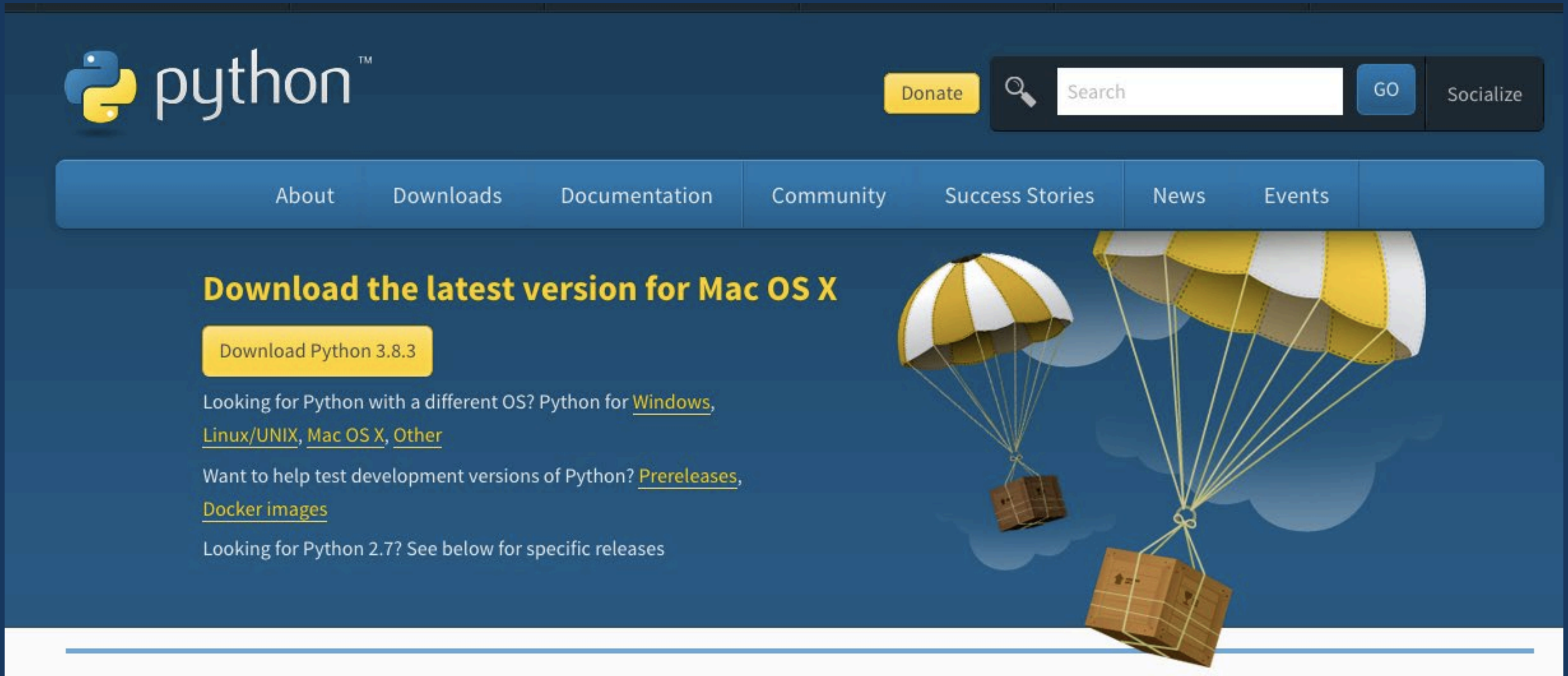
4. **Huge numbers of Python libraries and Frame work**

5. **Supports Big Data, Machine Learning and Cloud computing**

6. **Supports Automation**

Installing Python

This is the site to install Python -> https://www.python.org/downloads/



Download the latest version for Mac OS X

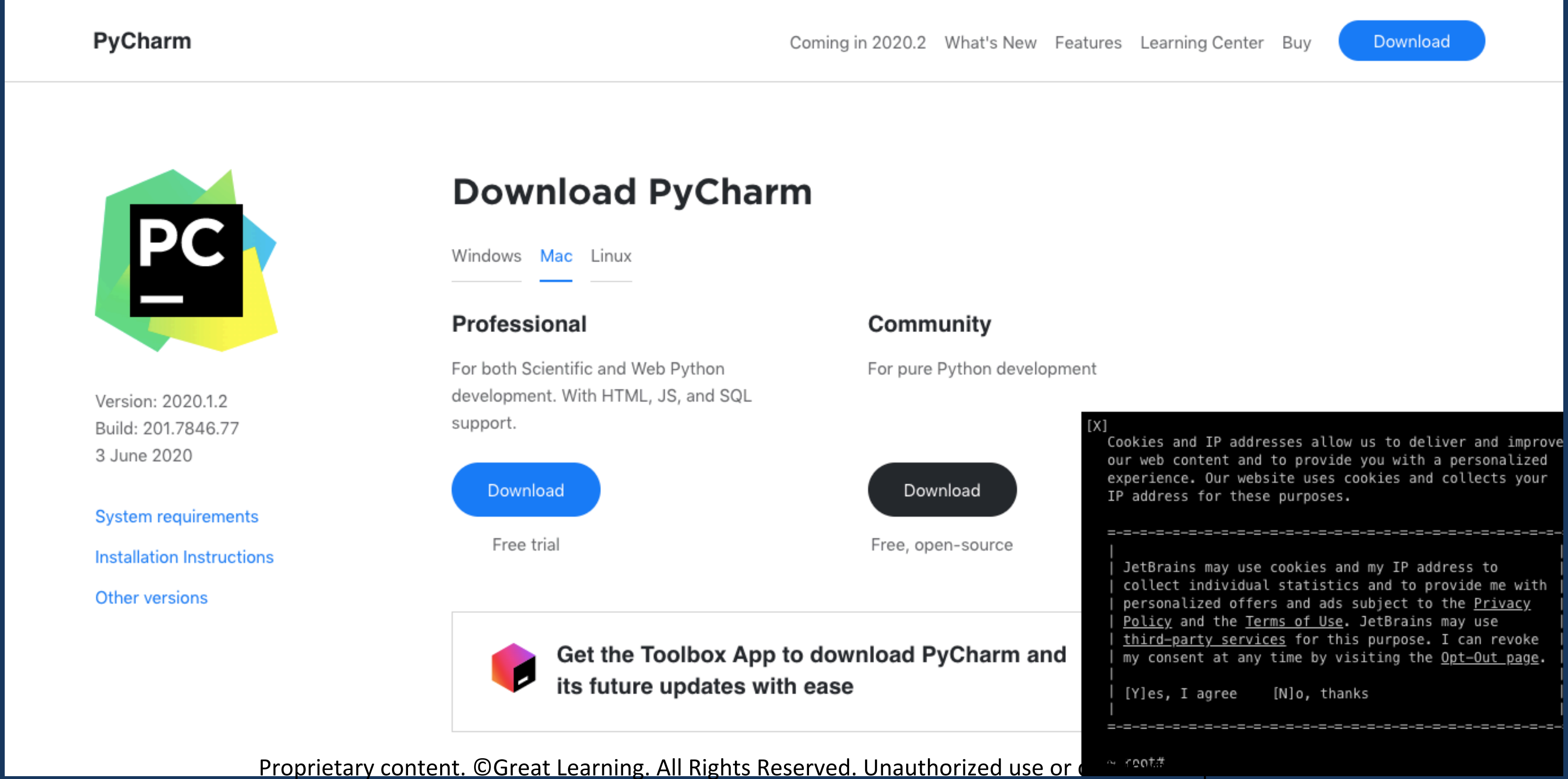
Looking for Python with a different OS? Python for Windows, Linux/UNIX, Mac OS X, Other

Want to help test development versions of Python? Prereleases, Docker images

Looking for Python 2.7? See below for specific releases

Popular IDE for Python: Pycharm

Site to install Python ->
https://www.jetbrains.com/pycharm/download/#section=mac



Version: 2020.1.2
Build: 201.7846.77
3 June 2020

System requirements

Installation Instructions

Other versions

Download PyCharm

Professional

For both Scientific and Web Python development. With HTML, JS, and SQL support.

Free trial

Community

For pure Python development

Free, open-source

Get the Toolbox App to download PyCharm and its future updates with ease

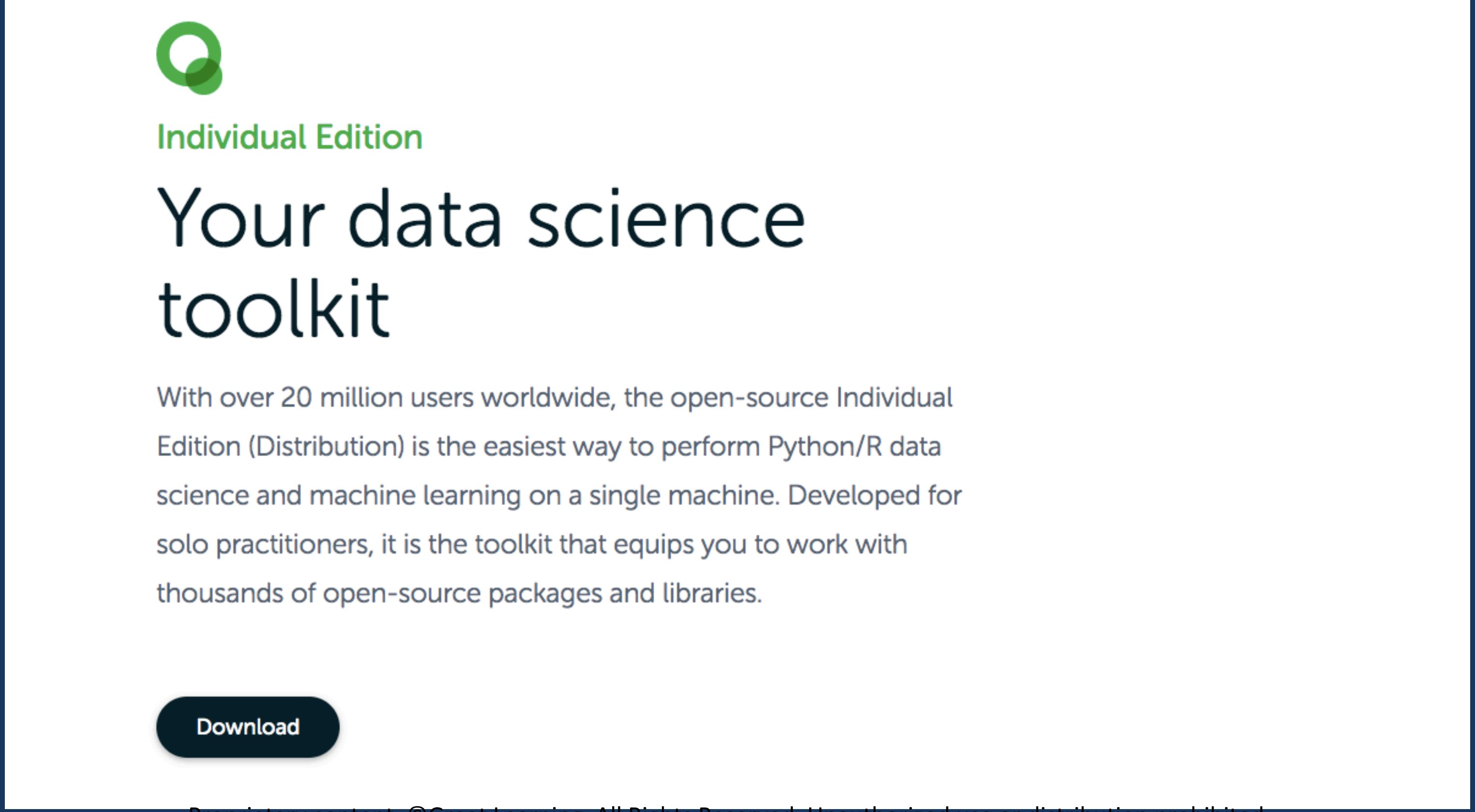
```
[X]
  Cookies and IP addresses allow us to deliver and improve
  our web content and to provide you with a personalized
  experience. Our website uses cookies and collects your
  IP address for these purposes.

  =-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-
  |                                                        |
  | JetBrains may use cookies and my IP address to         |
  | collect individual statistics and to provide me with   |
  | personalized offers and ads subject to the Privacy     |
  | Policy and the Terms of Use. JetBrains may use         |
  | third-party services for this purpose. I can revoke    |
  | my consent at any time by visiting the Opt-Out page.   |
  |                                                        |
  | [Y]es, I agree    [N]o, thanks                         |
  |                                                        |
  =-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-

  ~ root#
```

Popular IDE for Python: Anaconda

Anaconda installation site->
https://www.anaconda.com/products/individual



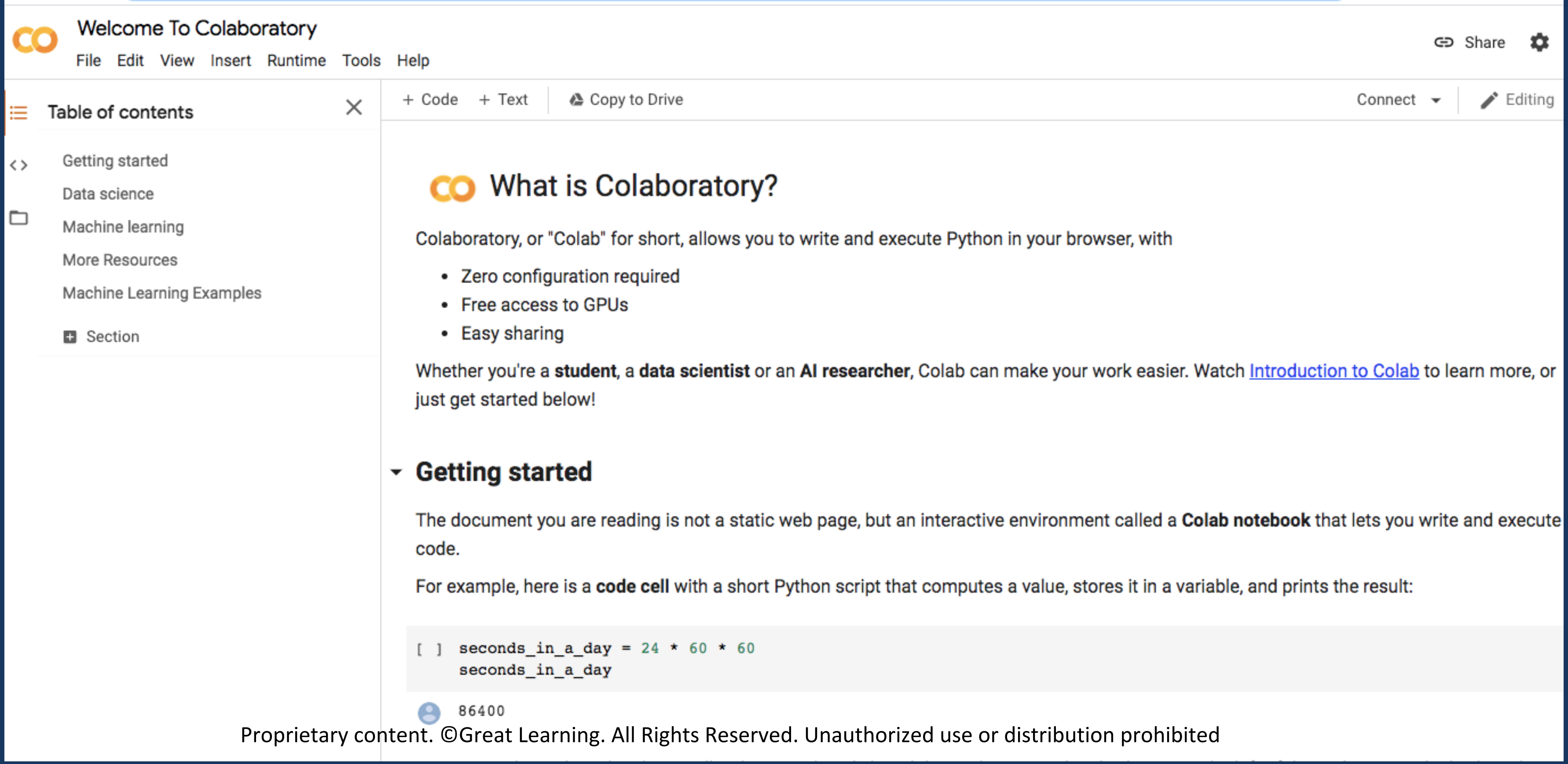
Individual Edition

Your data science toolkit

With over 20 million users worldwide, the open-source Individual Edition (Distribution) is the easiest way to perform Python/R data science and machine learning on a single machine. Developed for solo practitioners, it is the toolkit that equips you to work with thousands of open-source packages and libraries.

Popular IDE for Python: Google colab

Google collaboratory link-> https://colab.research.google.com/notebooks/intro.ipynb



What is Colaboratory?

Colaboratory, or "Colab" for short, allows you to write and execute Python in your browser, with

- Zero configuration required
- Free access to GPUs
- Easy sharing

Whether you're a **student**, a **data scientist** or an **AI researcher**, Colab can make your work easier. Watch Introduction to Colab to learn more, or just get started below!

Getting started

The document you are reading is not a static web page, but an interactive environment called a **Colab notebook** that lets you write and execute code.

For example, here is a **code cell** with a short Python script that computes a value, stores it in a variable, and prints the result:

```
[ ] seconds_in_a_day = 24 * 60 * 60
    seconds_in_a_day
```

```
86400
```

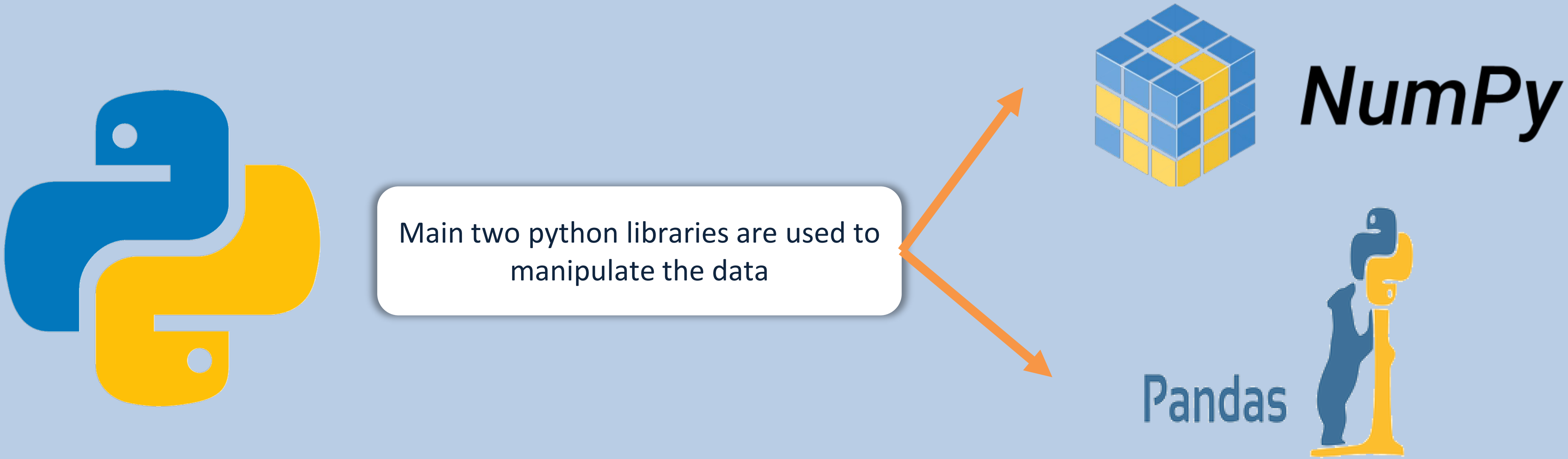
Getting started with Python

greatlearning
Learning for Life

Python Libraries

Data manipulation using Python

Data manipulation is a technique which allows to transform, extract, and filter your data efficiently with less time.



Main two python libraries are used to manipulate the data

Data manipulation using Numpy

Numpy stands for Numerical Python and it is used to perform mathematical and logical operations on arrays

1 Numpy is a python library

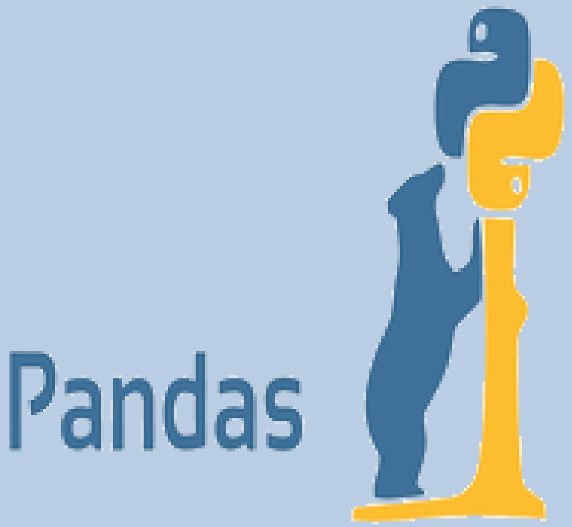
2 Install Numpy: !pip install numpy

3 Import the Library: import numpy as np

Data manipulation using Pandas

Pandas is a popular data manipulation and analysis library in python which is based on Numpy

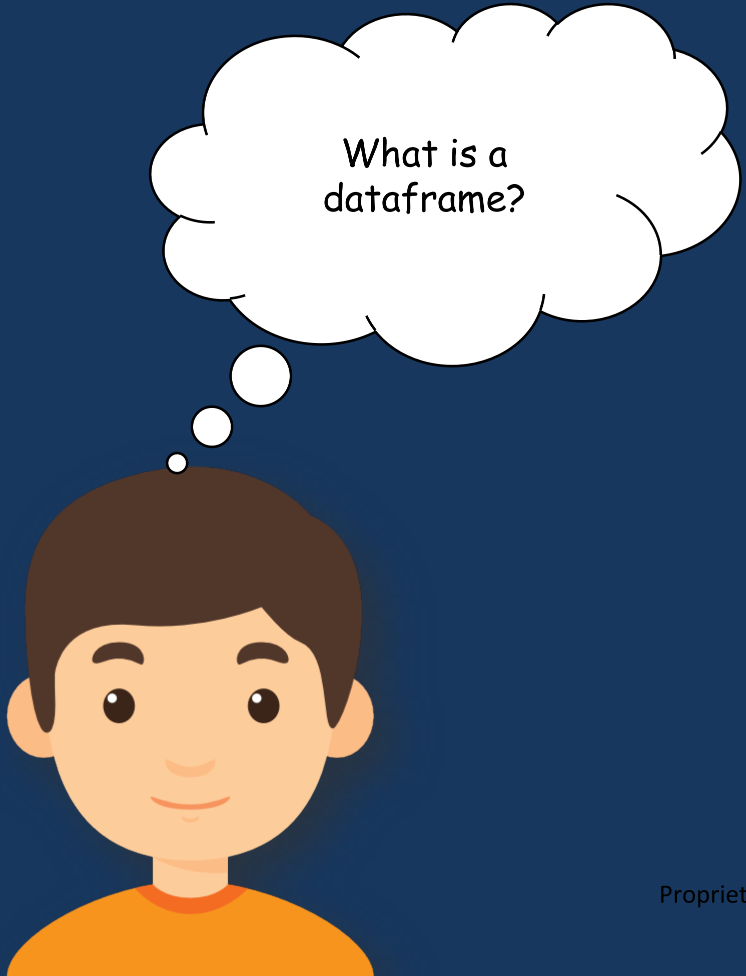
1. Python is a python library built on top of Numpy
2. Install Numpy: !pip install pandas
3. Import the Library: import pandas as pd

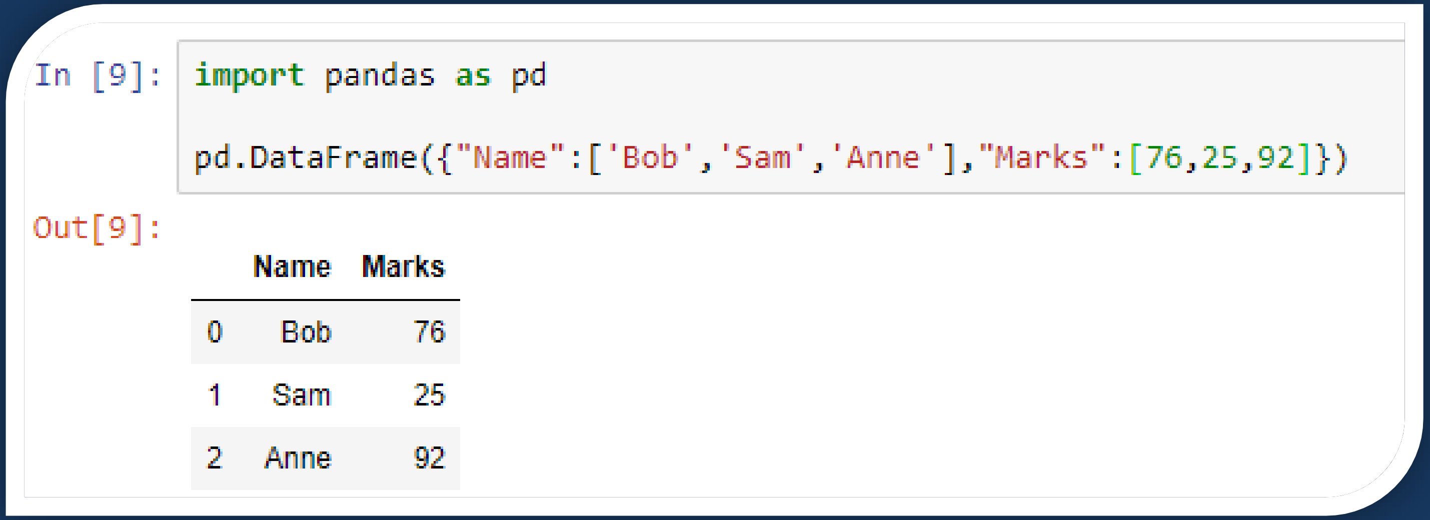


Demo on Numpy and pandas

Creating a Dataframe

Panda's data frame is a two-dimensional data structure which is aligned in a tabular fashion with rows and columns



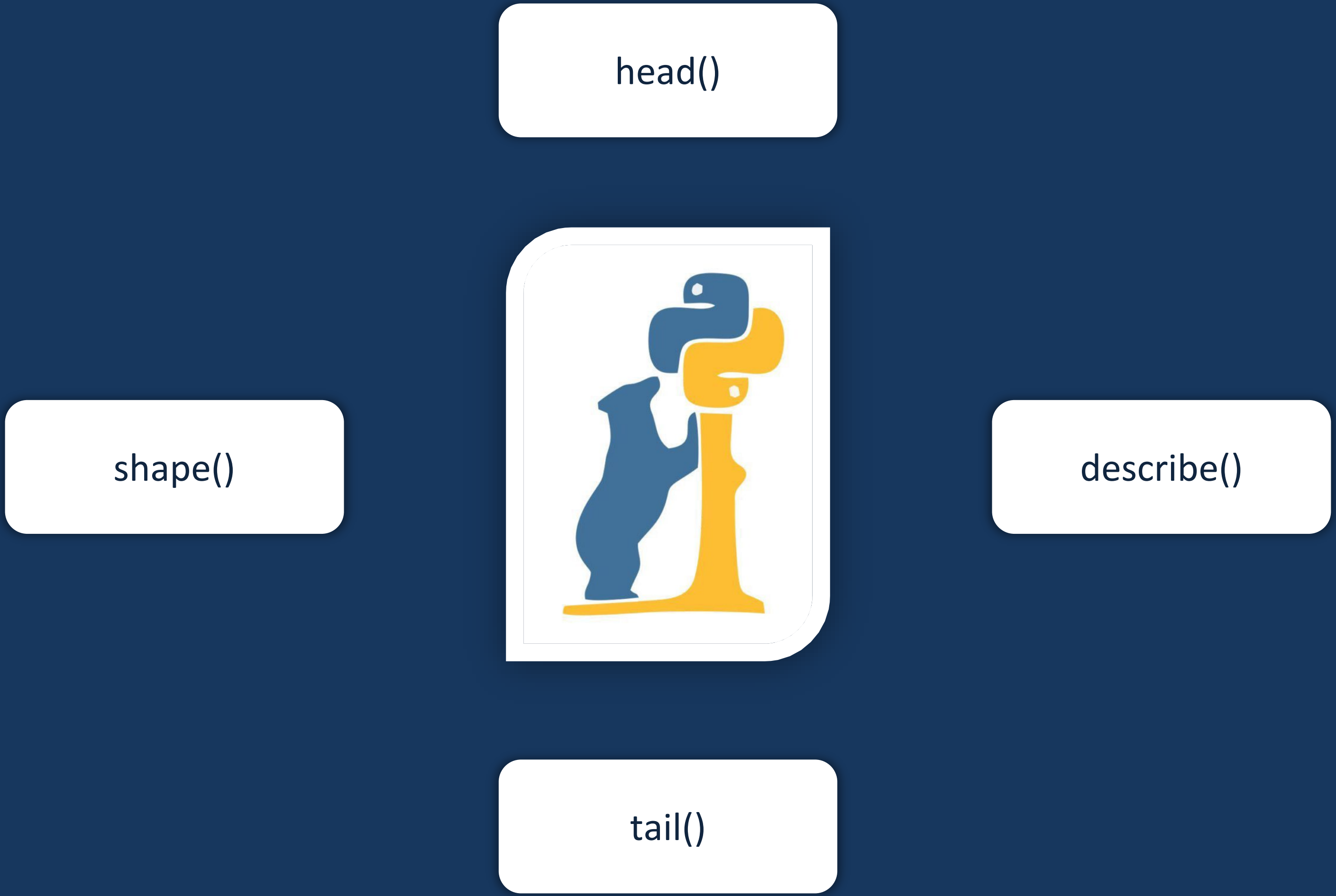


```
In [9]: import pandas as pd

        pd.DataFrame({"Name":['Bob','Sam','Anne'],"Marks":[76,25,92]})

Out[9]:
           Name  Marks
        0   Bob     76
        1   Sam     25
        2  Anne     92
```

Dataframe In-Built Functions



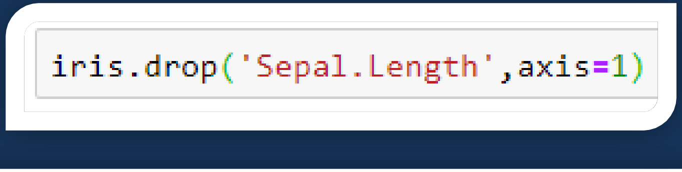
head()

shape()

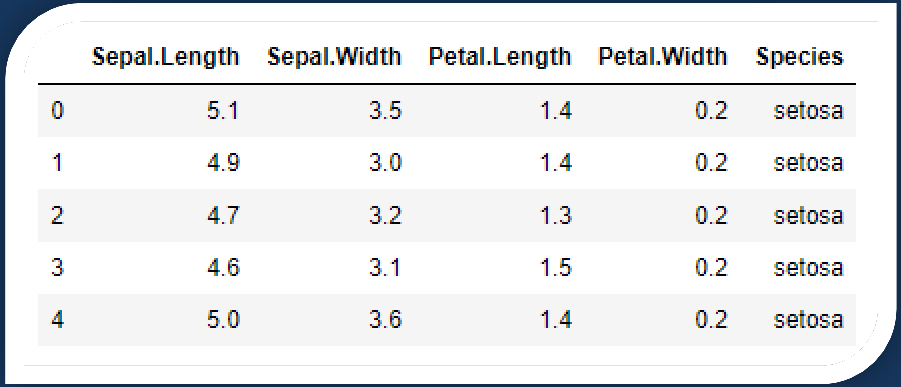
describe()

tail()

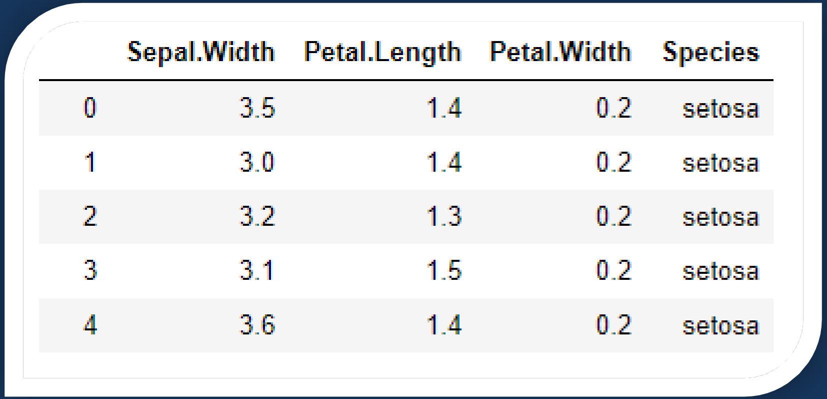
Dropping Columns



```
iris.drop('Sepal.Length',axis=1)
```

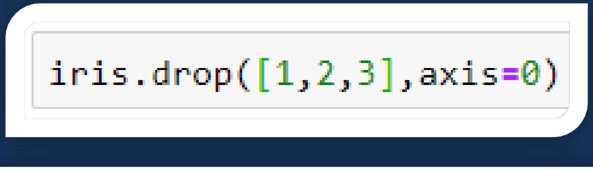


	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

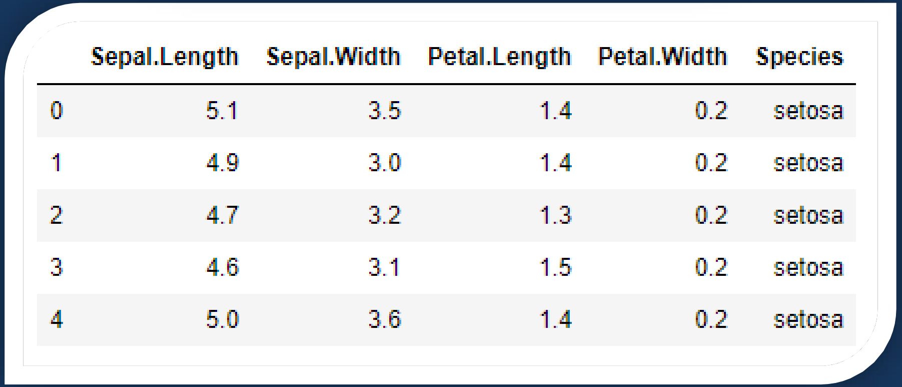


	Sepal.Width	Petal.Length	Petal.Width	Species
0	3.5	1.4	0.2	setosa
1	3.0	1.4	0.2	setosa
2	3.2	1.3	0.2	setosa
3	3.1	1.5	0.2	setosa
4	3.6	1.4	0.2	setosa

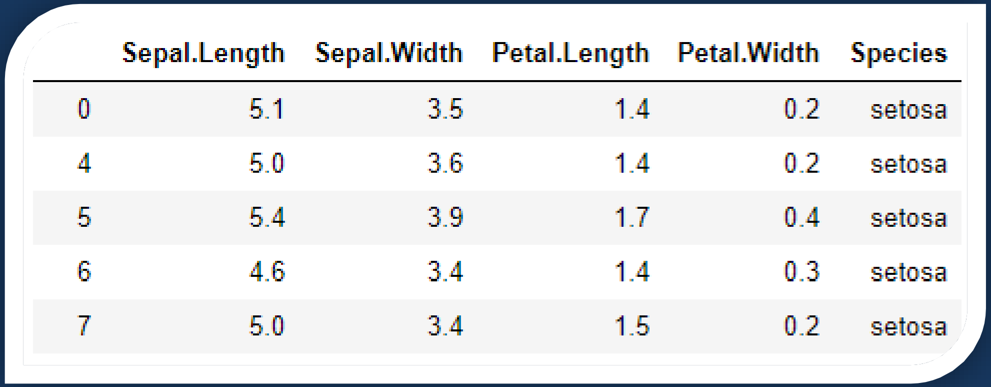
Dropping Rows



```
iris.drop([1,2,3],axis=0)
```



	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa



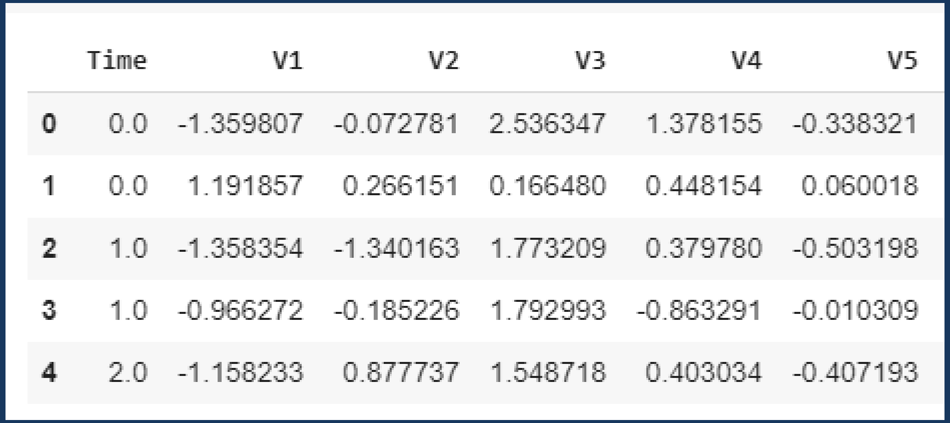
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
0	5.1	3.5	1.4	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
5	5.4	3.9	1.7	0.4	setosa
6	4.6	3.4	1.4	0.3	setosa
7	5.0	3.4	1.5	0.2	setosa

greatlearning
Learning for Life

Machine Learning to build the Model

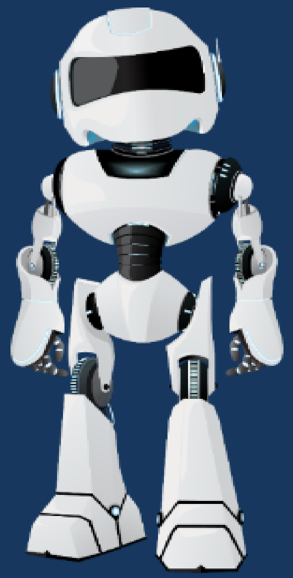
What is Machine Learning?

Machine learning is a sub-set of artificial intelligence (AI) that allows the system to automatically learn and improve from experience without being explicitly programmed

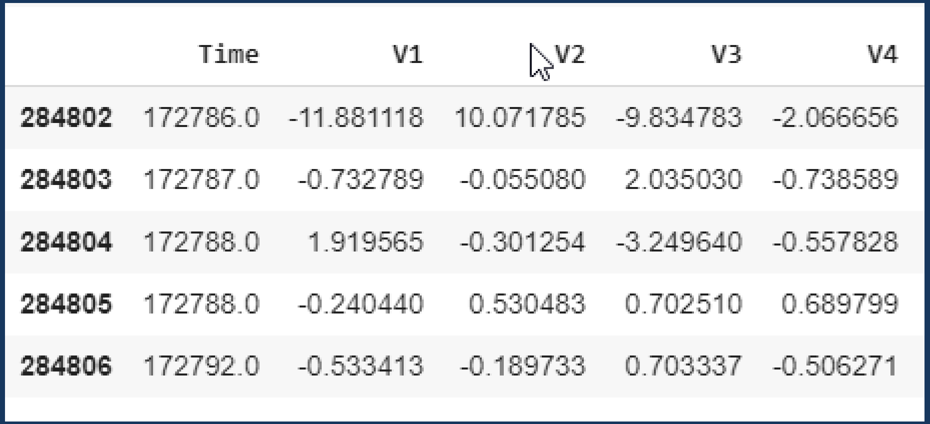


	Time	V1	V2	V3	V4	V5
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193

Training Data



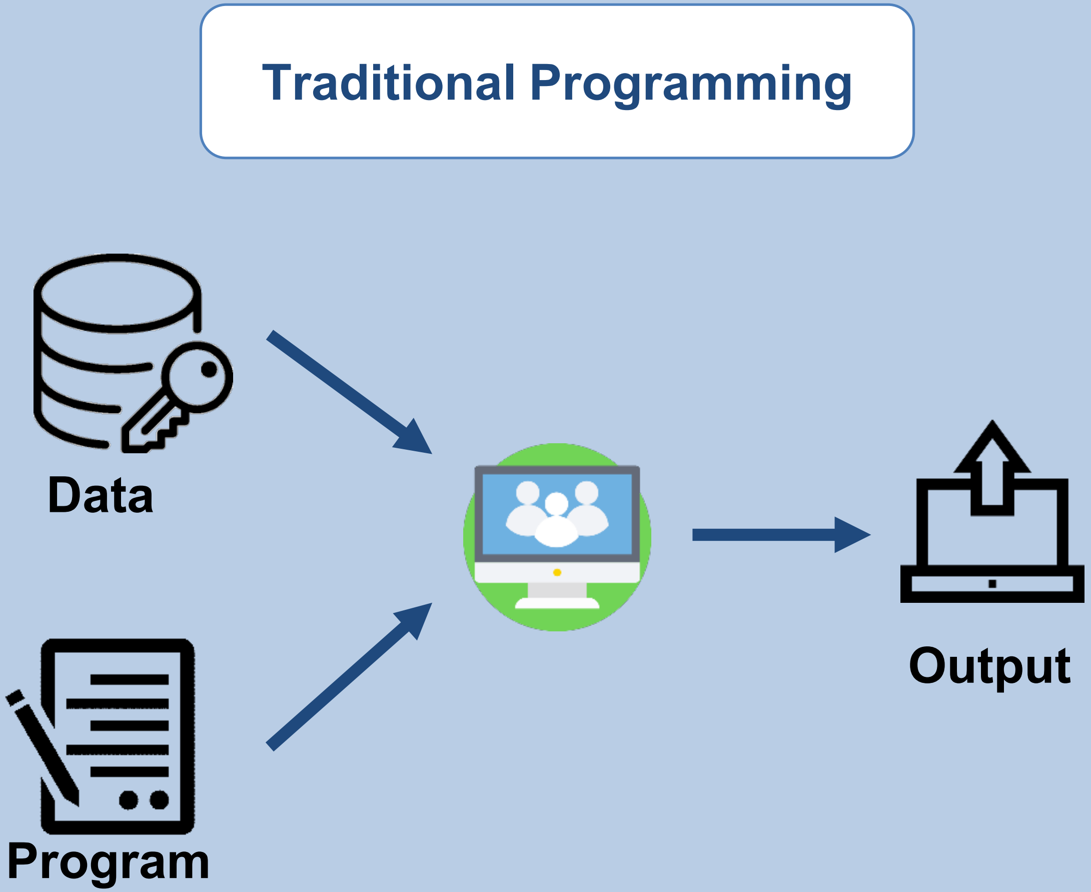
Model Building

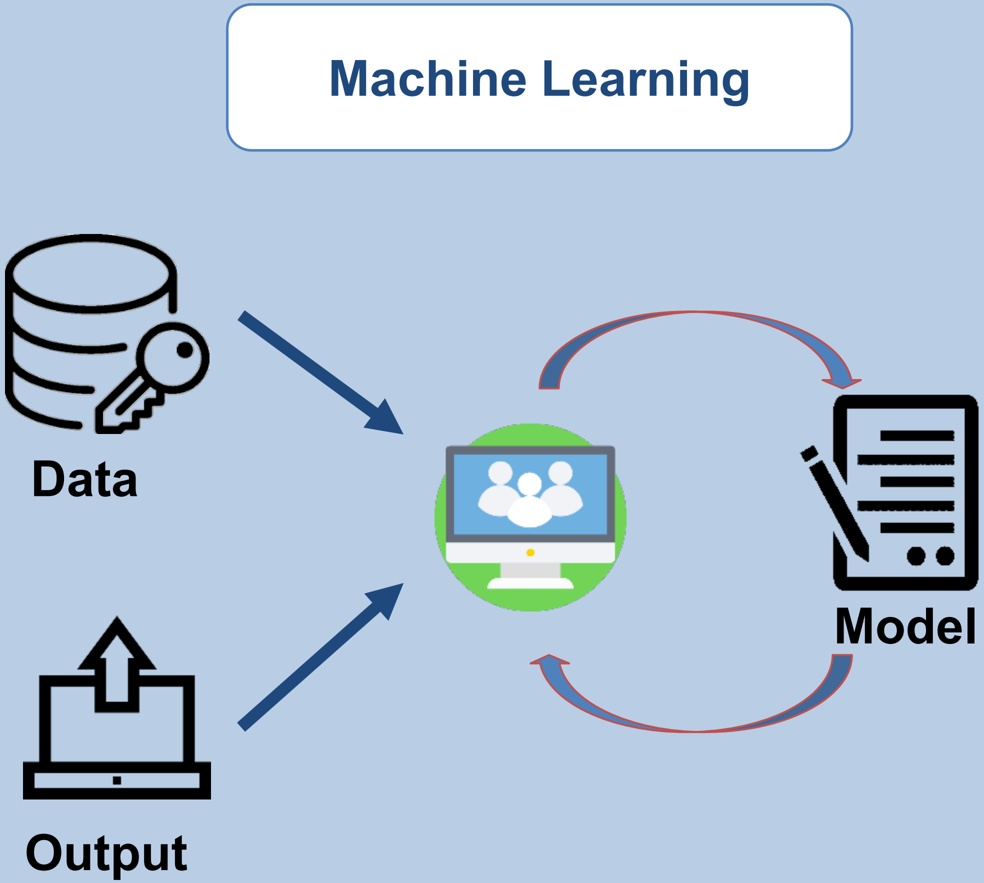


	Time	V1	V2	V3	V4
284802	172786.0	-11.881118	10.071785	-9.834783	-2.066656
284803	172787.0	-0.732789	-0.055080	2.035030	-0.738589
284804	172788.0	1.919565	-0.301254	-3.249640	-0.557828
284805	172788.0	-0.240440	0.530483	0.702510	0.689799
284806	172792.0	-0.533413	-0.189733	0.703337	-0.506271

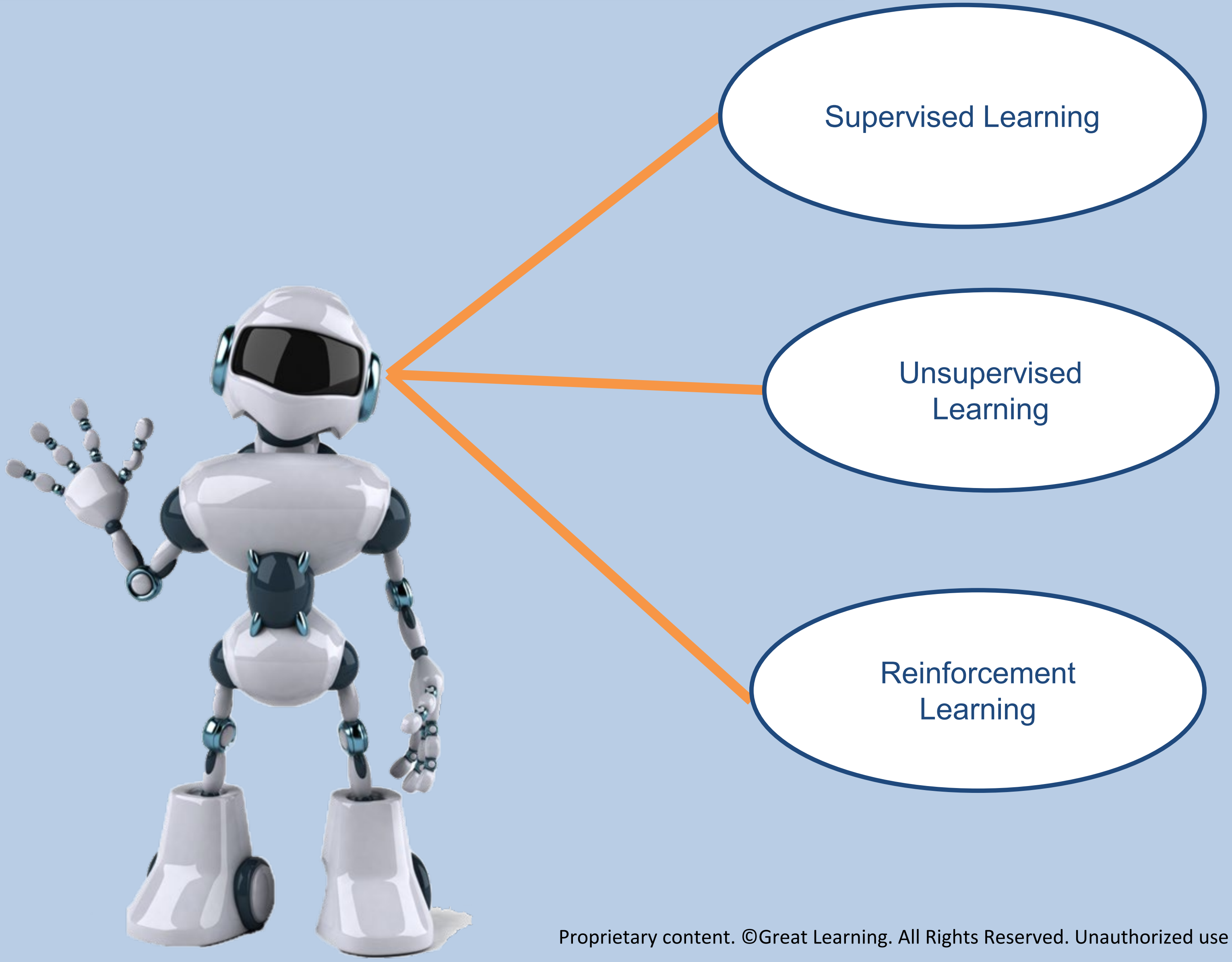
Testing Data

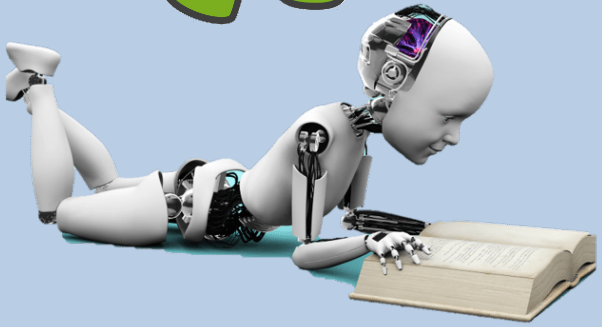
Traditional Vs Machine Learning

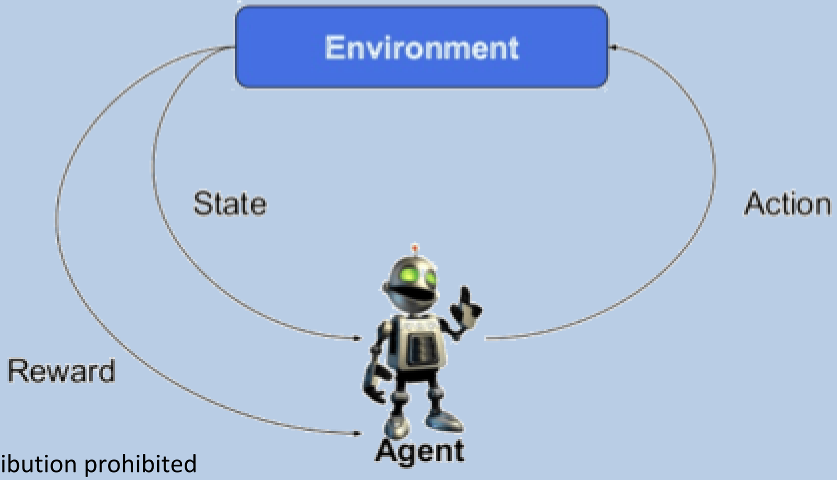




Types Of Machine Learning







What is Supervised Learning?

Supervised learning works as a supervisor or teacher. Basically, In supervised learning, we teach or train the machine with labeled data (that means data is already tagged with some predefined class). Then we test our model with some unknown new set of data and predict the level for them

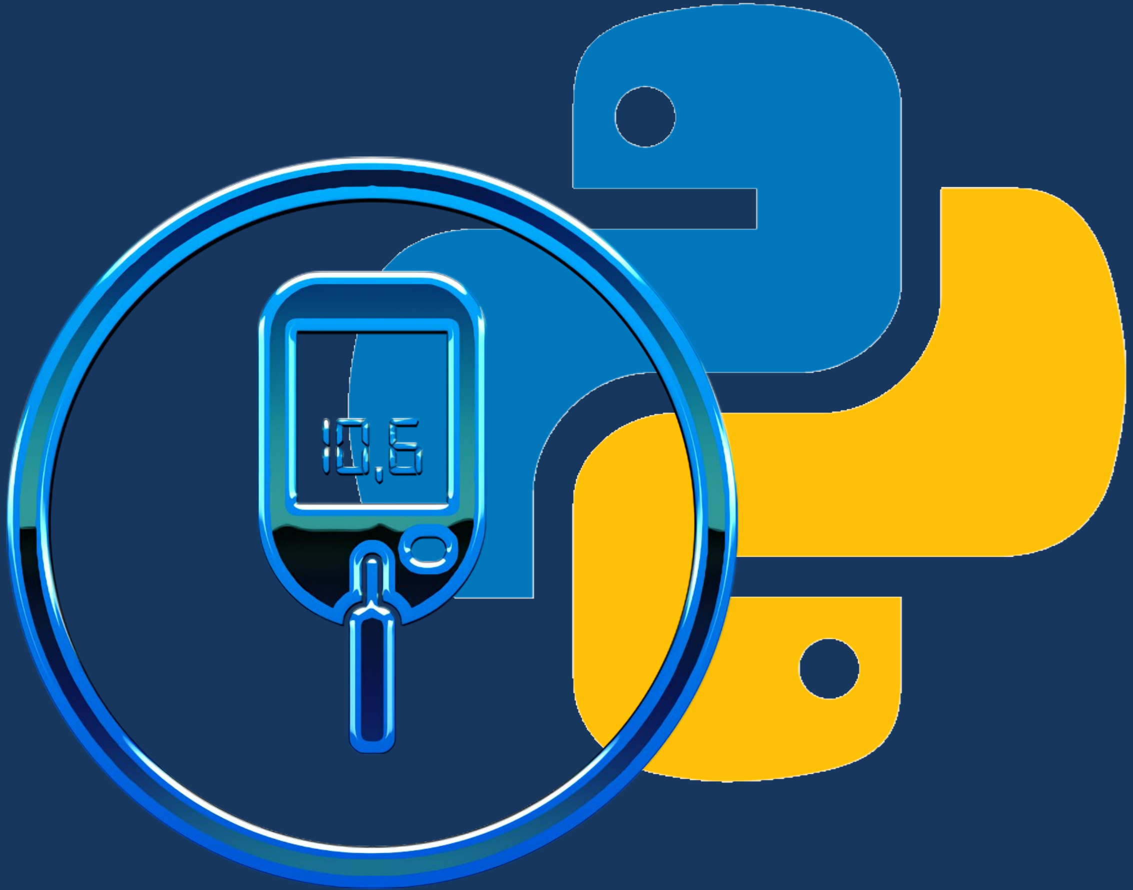
- Learning from the labelled data and applying the knowledge to predict the label of the new data(test data), is known as ***Supervised Learning***
- ***Types of Supervised Learning:***
  - **Linear Regression**
  - **Logistic regression**
  - **Decision Tree**
  - **Random Forest**
  - **Naïve Bayes Classifier**

What is Logistic Regression?

Logistic regression is also a part of supervised learning classification algorithm. It is used to predict the probability of a target variable and the nature of target or dependent variable is discrete, so for the output there will be only two class will be present

- The dependent variable is binary in nature so that can be either 1 (stands for success/yes) or 0 (stands for failure/no).
- Logistic regression is also known as sigmoid function
- *Sigmoid function = 1 / (1 + e^-value)*

greatlearning
Learning for Life



Diabetes Prediction using Python

greatlearning
Learning for Life

Thank You