

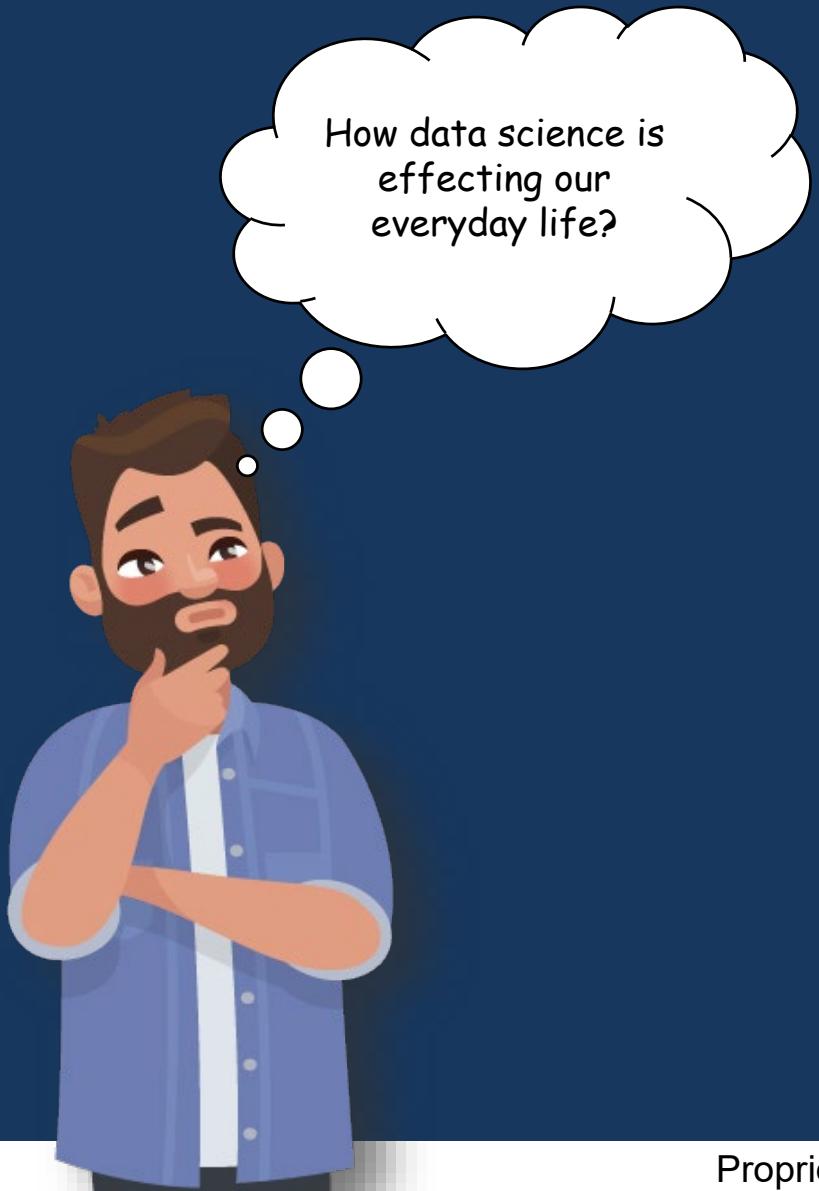
Uber Data Analysis with Python

Sampriti Chatterjee (Great Learning)

Agenda

- 1 Why do we need data science?
- 2 What is Data science?
- 3 Life cycle of Data science
- 4 Why Python is so popular?
- 5 Install python
- 6 Statistical visualization on Python user
- 7 Basic Syntax for Python
- 8 What is machine Learning?
- 9 Supervised Learning: Linear Regression
- 10 Uber data analysis to predict price

Why do we need Data Science?



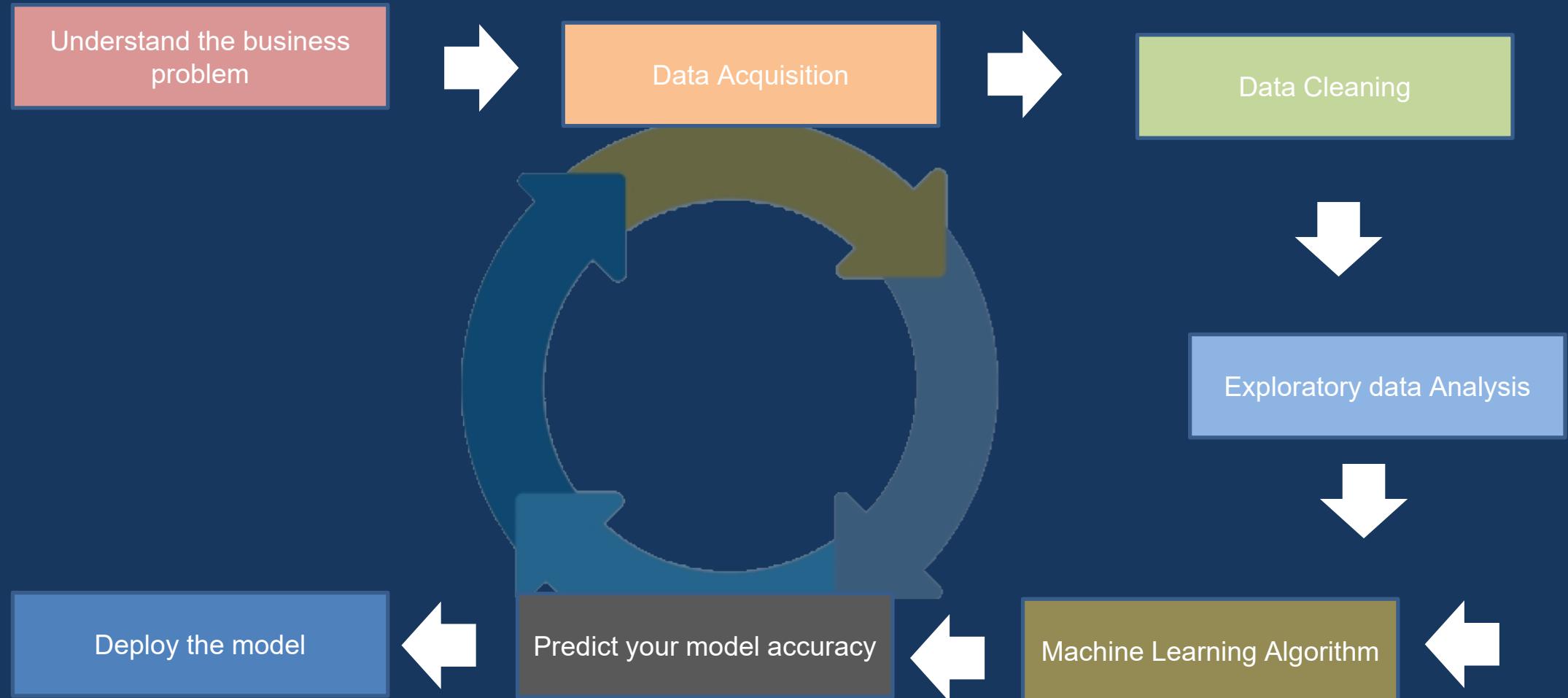
- In the past, we used to have data in a structured format but now as the volume of the data is increasing, so the number of structured data becomes very less, so to handle the massive amount of data we need data science techniques
- Those data can be used to get the proper business insights and the 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?



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

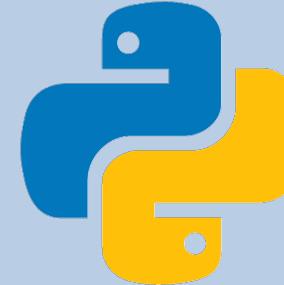
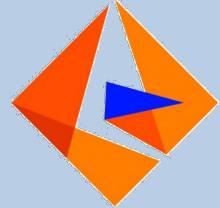
Life cycle of Data Science?



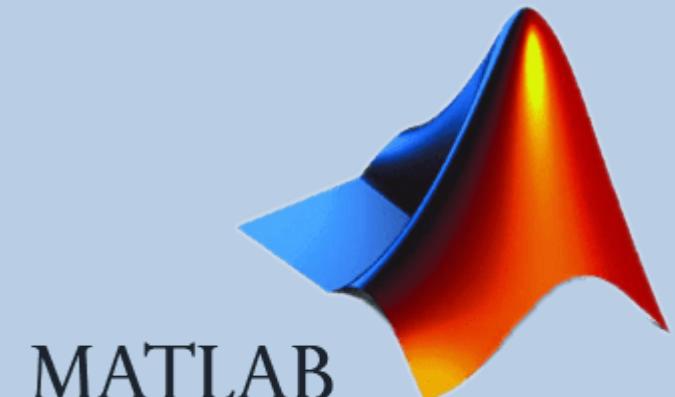
Most Popular Programming Languages For Data Science



Informatica™

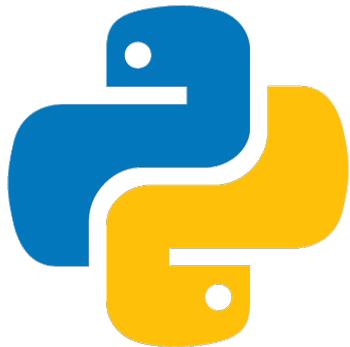


sas®

The SAS logo, featuring a stylized blue "S" followed by the word "sas" in a bold black sans-serif font with a registered trademark symbol.

MATLAB

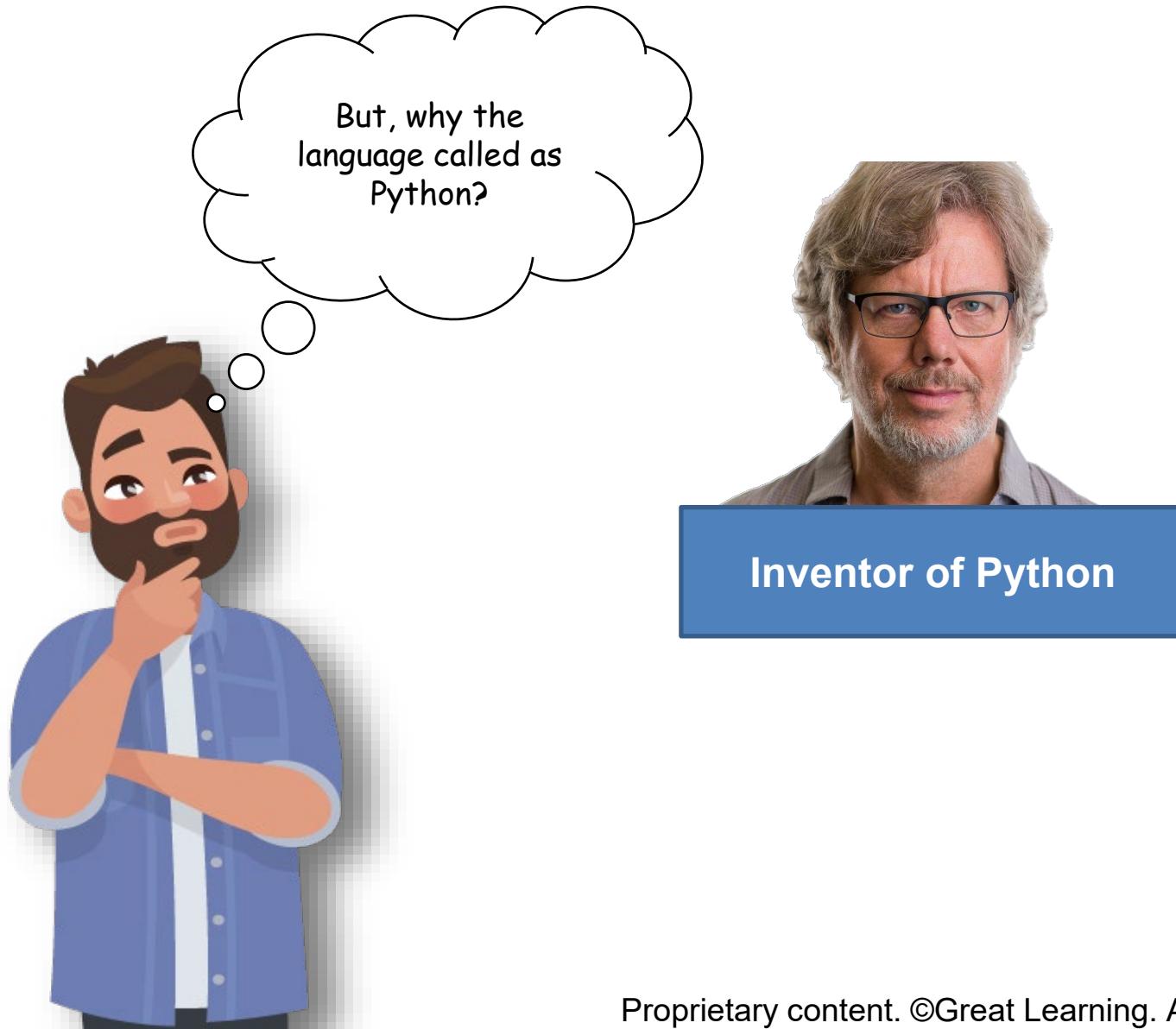
Python is a popular high level, object oriented and interpreted language



High level

Interpreted

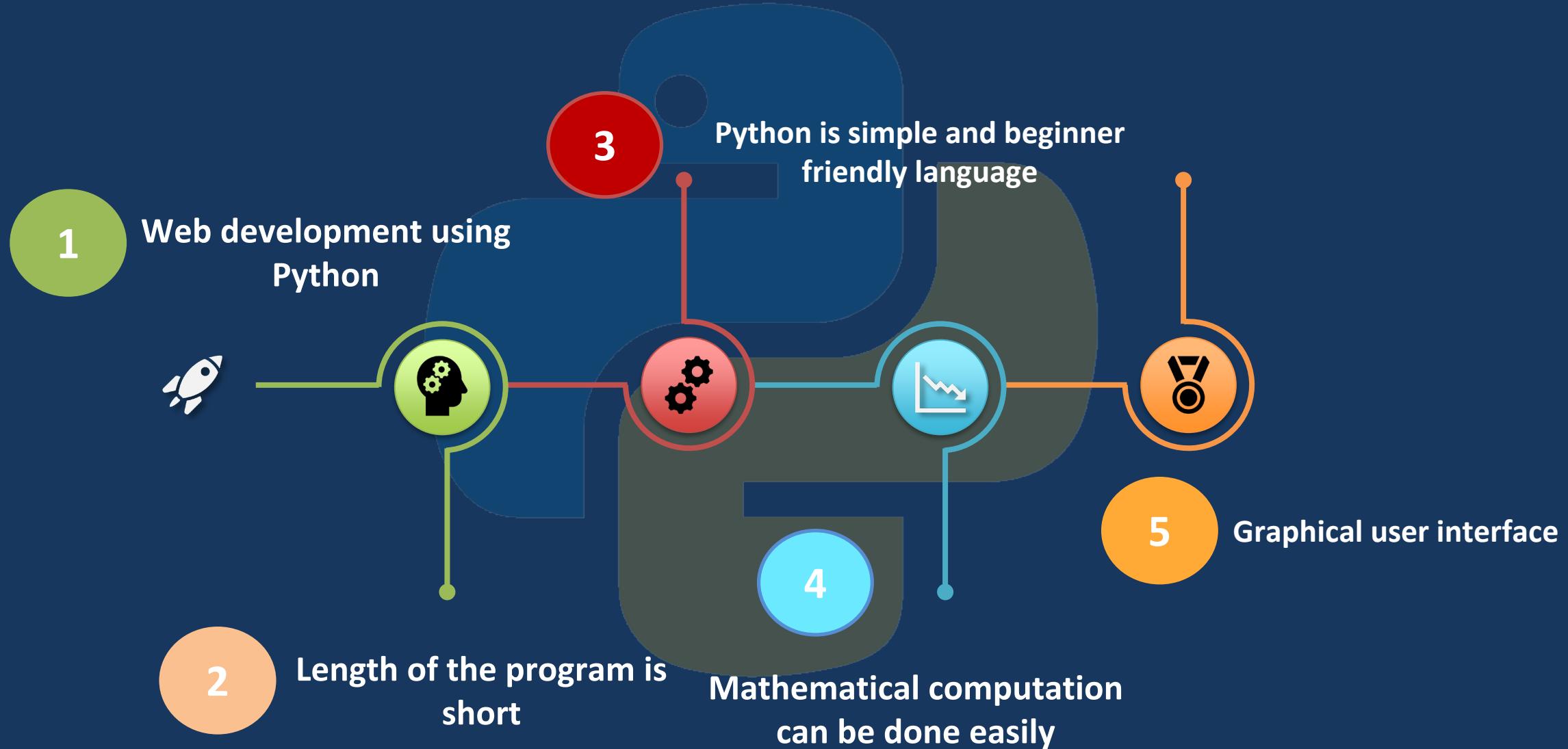
Object oriented



Important Facts

- Python is invented by Guido van Rossum in 1989
- Rossum used to love watching comedy movies from late seventies
- He needed a short, unique, and slightly mysterious name for his language
- In that time he was watching Monty Python's Flying Circus and from that series he decided to keep his language name python.
- This how Python invented

Why should you learn Python?



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

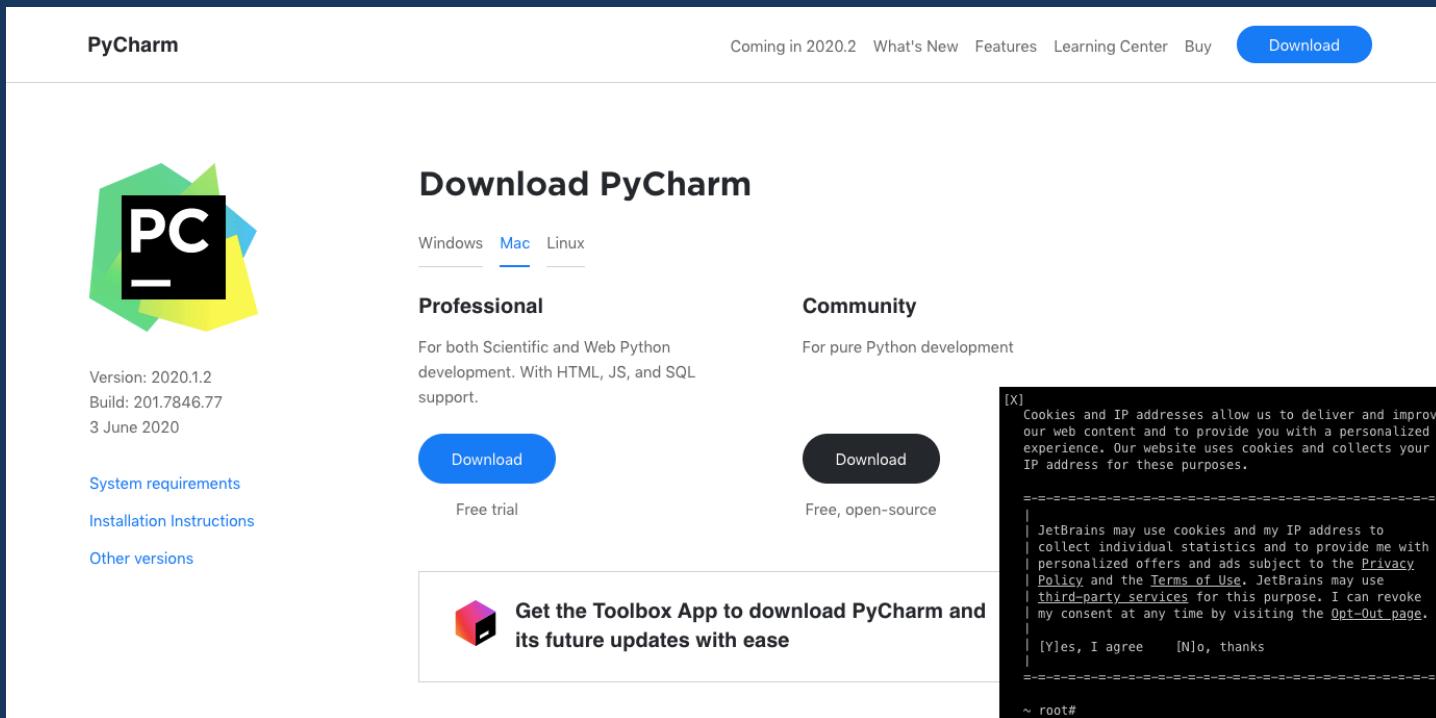
This is the site to install Python -> <https://www.python.org/downloads/>



Popular IDE for Python: Pycharm

Site to install Python ->

<https://www.jetbrains.com/pycharm/download/#section=mac>



The screenshot shows the official PyCharm download page. At the top, there's a navigation bar with links to 'Coming in 2020.2', 'What's New', 'Features', 'Learning Center', 'Buy', and a prominent blue 'Download' button. Below the navigation, there's a large image of the PyCharm logo (a stylized 'PC' inside a hexagon) and some version information: 'Version: 2020.1.2', 'Build: 201.7846.77', and '3 June 2020'. To the right, there's a section titled 'Download PyCharm' with tabs for 'Windows', 'Mac' (which is selected), and 'Linux'. Under the 'Mac' tab, there are two main options: 'Professional' and 'Community'. The 'Professional' section is described as being for 'both Scientific and Web Python development. With HTML, JS, and SQL support.' It includes a 'Download' button and a 'Free trial' link. The 'Community' section is described as being for 'pure Python development' and is labeled as 'Free, open-source'. It also has a 'Download' button. At the bottom of the page, there's a call-to-action: 'Get the Toolbox App to download PyCharm and its future updates with ease' with an icon of the Toolbox app. On the right side of the page, there's a dark sidebar containing a cookie consent message from JetBrains.

PyCharm

Coming in 2020.2 What's New Features Learning Center Buy Download

Version: 2020.1.2 Build: 201.7846.77 3 June 2020

System requirements Installation Instructions Other versions

Download PyCharm

Windows Mac Linux

Professional

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

Download Free trial

Community

For pure Python development

Download Free, open-source

[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.

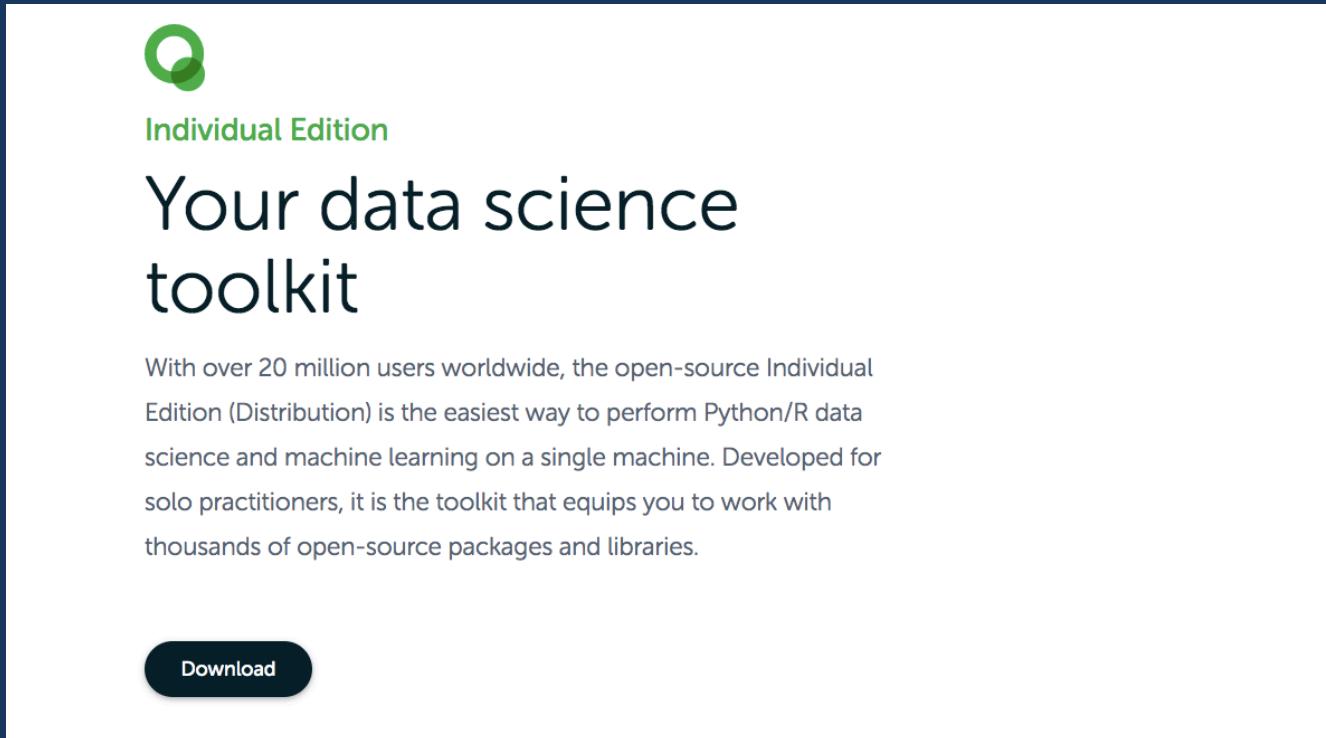
JetBrains may use cookies and my IP address to collect individual statistics and to provide me with personalized offers and ads subject to the [Privacy Policy](#) and the [Terms of Use](#). JetBrains may use [third-party services](#) for this purpose. I can revoke my consent at any time by visiting the [Opt-Out page](#).

[Y]es, I agree [N]o, thanks

~ root#

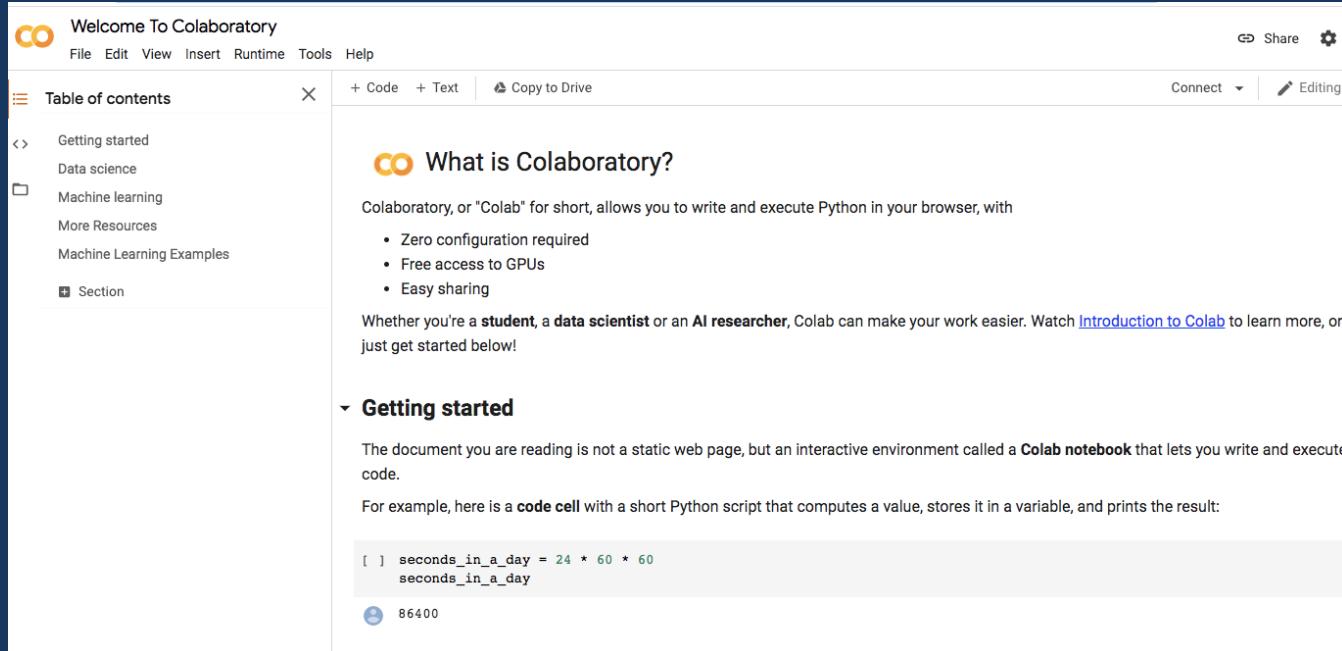
Anaconda installation site->

<https://www.anaconda.com/products/individual>



Popular IDE for Python: Google colab

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



Welcome To Colaboratory

File Edit View Insert Runtime Tools Help

Share

Table of contents

- Getting started
- Data science
- Machine learning
- More Resources
- Machine Learning Examples
- Section

+ Code + Text Copy to Drive Connect Editing

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

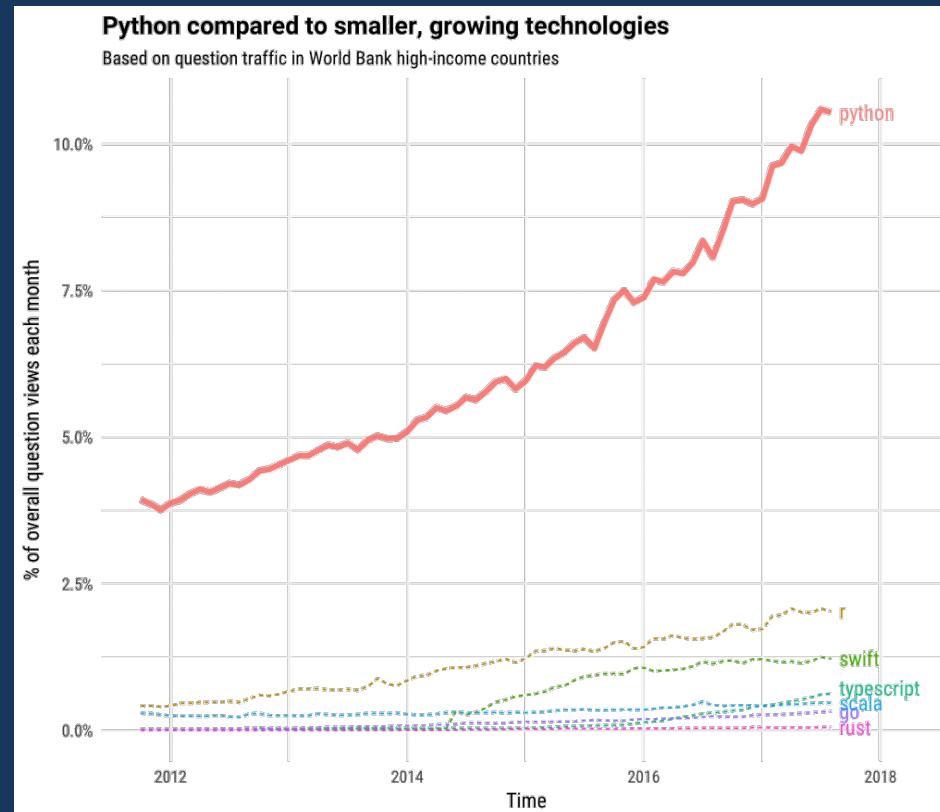
The document you are reading is not a static web page, but an interactive environment called a **Colab notebook** that lets you write and execute code.

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

Statistical measurement on Python user



In recent time it is prominent that Python is one of the most popular language because of it's simplicity

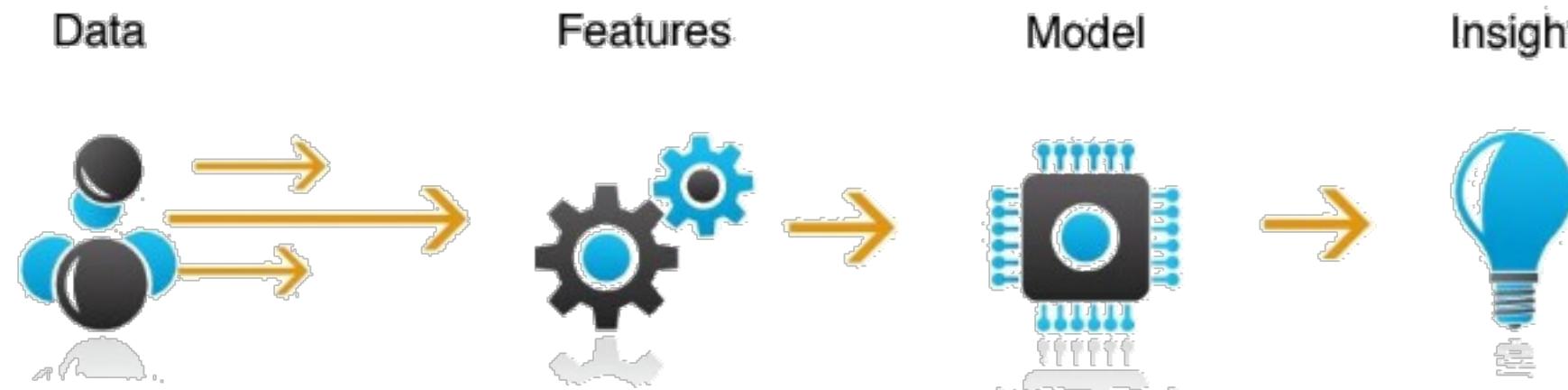


Uber Data Analysis

Exploratory data analysis is basically a technique to understand your data using statistics concept



Feature engineering process is basically used to create new feature from the existing data which helps to understand the data more deeply



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 |



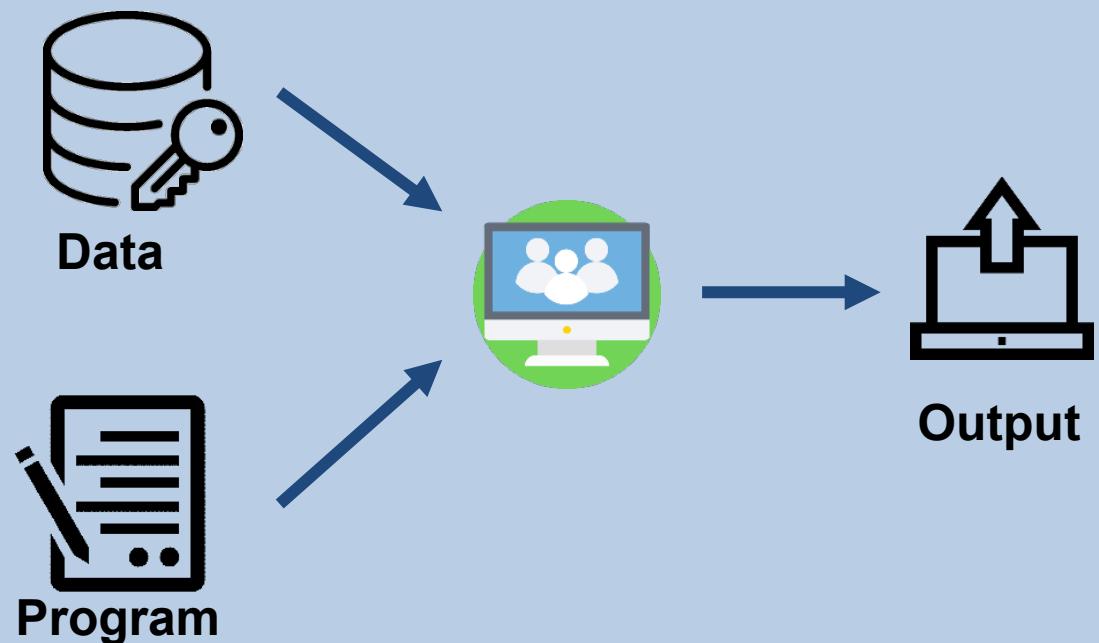
| | Time | V1 | V2 | V3 | V4 |
|--------|----------|------------|-----------|-----------|-----------|
| 284802 | 172786.0 | -11.881118 | 10.071785 | -9.834783 | -2.066656 |
| 284803 | 172787.0 | -0.732789 | -0.055080 | 2.035030 | -0.738589 |
| 284804 | 172788.0 | 1.919565 | -0.301254 | -3.249640 | -0.557828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

Training Data

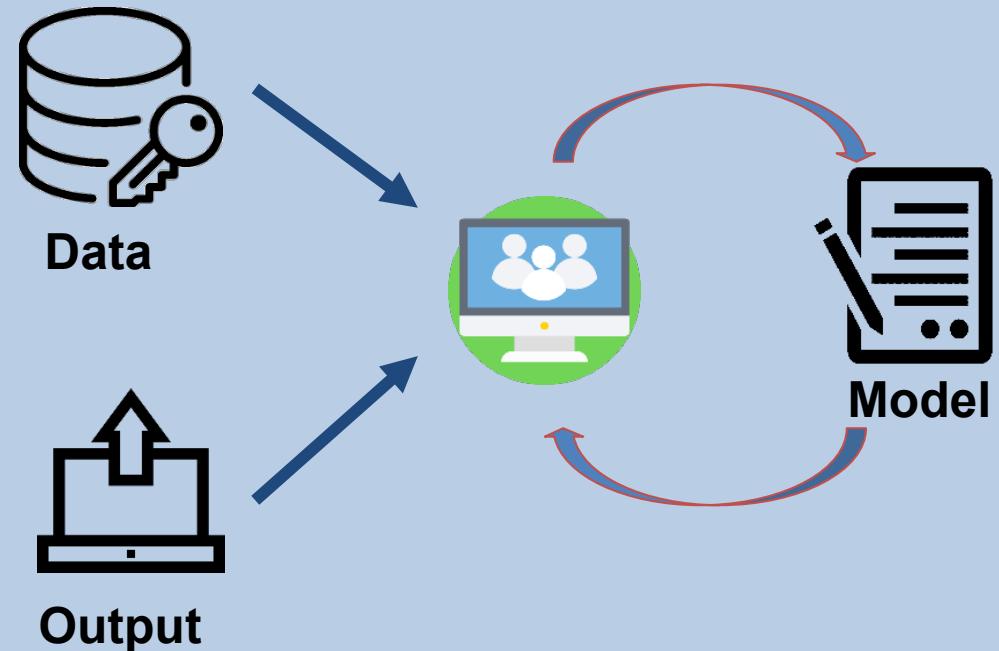
Model Building

Testing Data

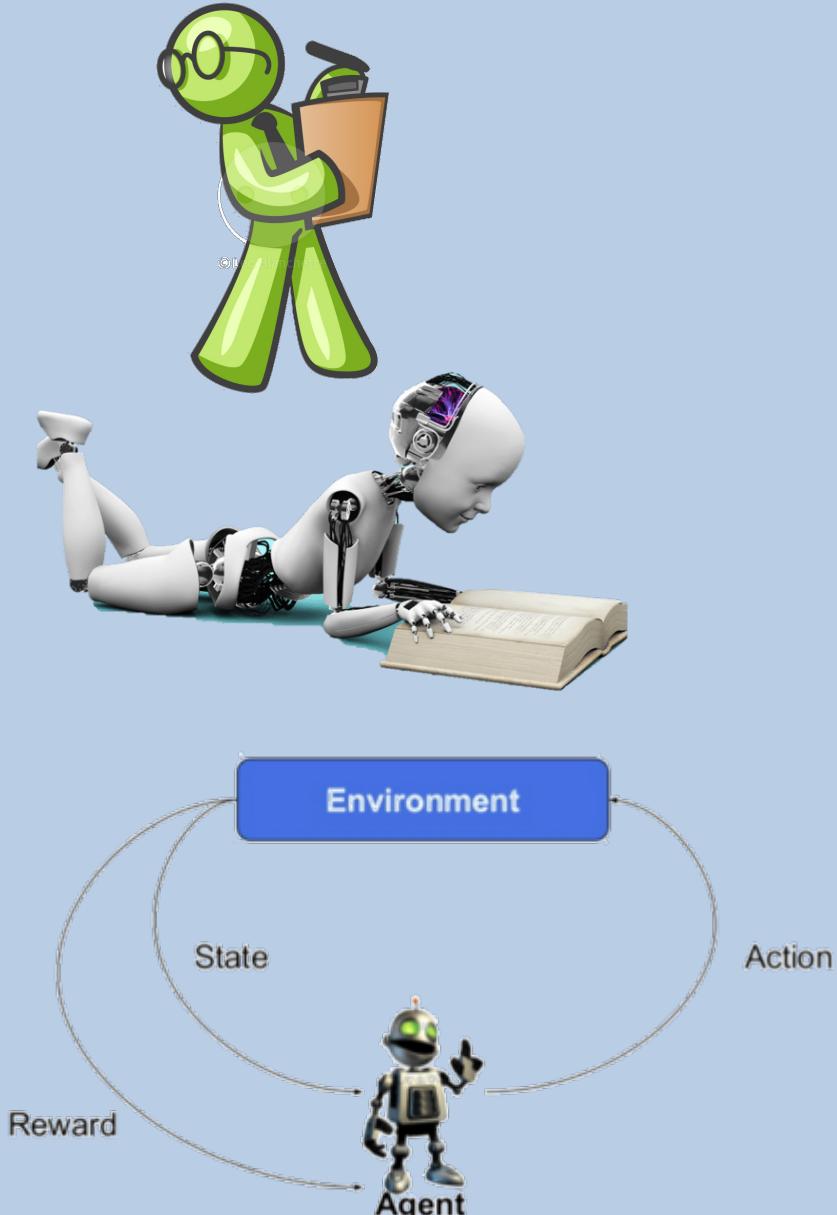
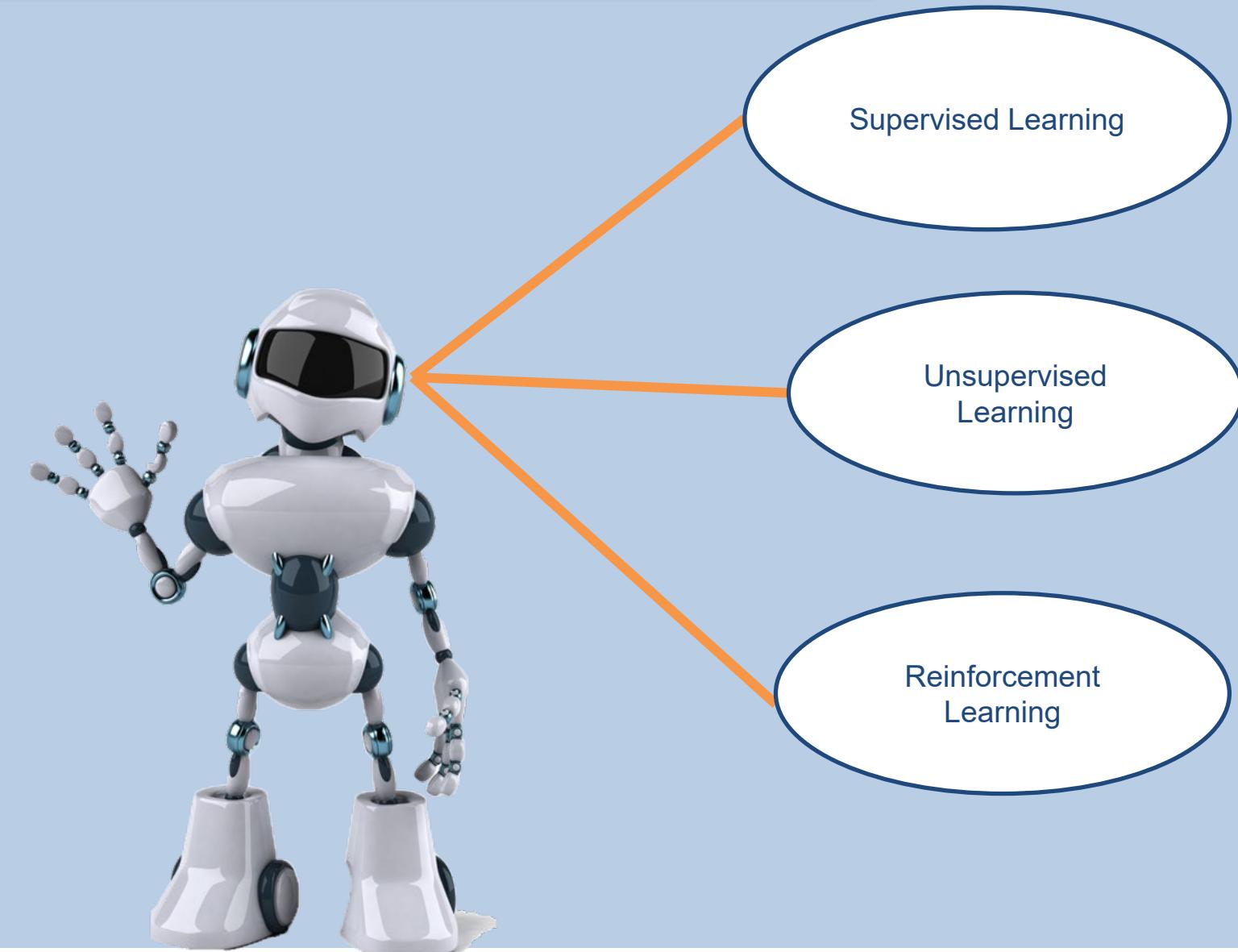
Traditional Programming



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**



© Leo Blanchette

What is Linear Regression?

Regression stands for to model a target value based on independent variables and Linear Regression is used to find the relationship between dependent(y) and independent variable(x)

- Linear regression is a supervised machine learning algorithm
- Always works with continuous value
- Formula: $y = mx+c$ m=slope of line and c= intercept
- Main target for linear regression to find the best value for X and Y



Thank You