# Cassandra

## Workshop 1

# Plan of Action

**Activities:-**

- 5 Workshops
- Assignment releasing after every workshop (most probably next day)
- 2 Real time quizzes in between Workshops
- Workshop will be on theoretical cum discussion and innovation basis of previous assignment.
- 2nd workshop can be an exception.
- A Dummy competition instead of last assignment.

**Weightage(Total = 100%):-**

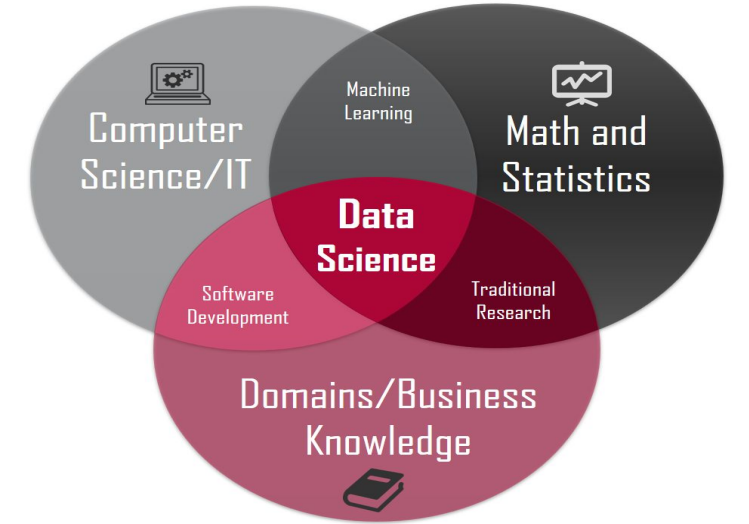
- First 3 Assignments (5% each)(5x3=15%)
- Last Assignment (10%)(10x1=10%)
- 2 Real Time Quiz (5% each)(5x2=10%)
- Dummy Competition(15%)
- Main Event(50%)

# DATA SCIENCE

**According to Wikipedia:-**

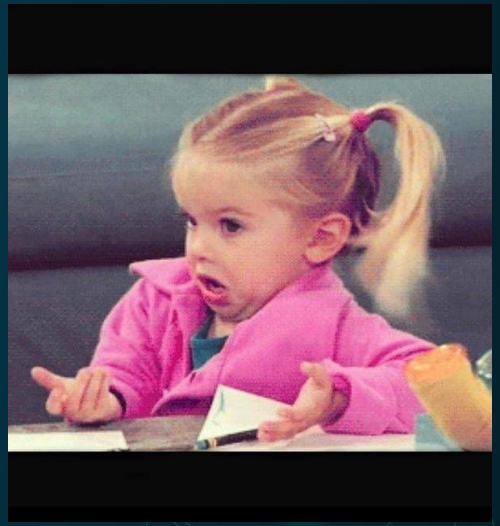
**Data science** is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data,[1][2] and apply knowledge and actionable insights from data across a broad range of application domains. Data science is related to data mining, machine learning and big data.

Diagrammatically:-



# Why Data Science?

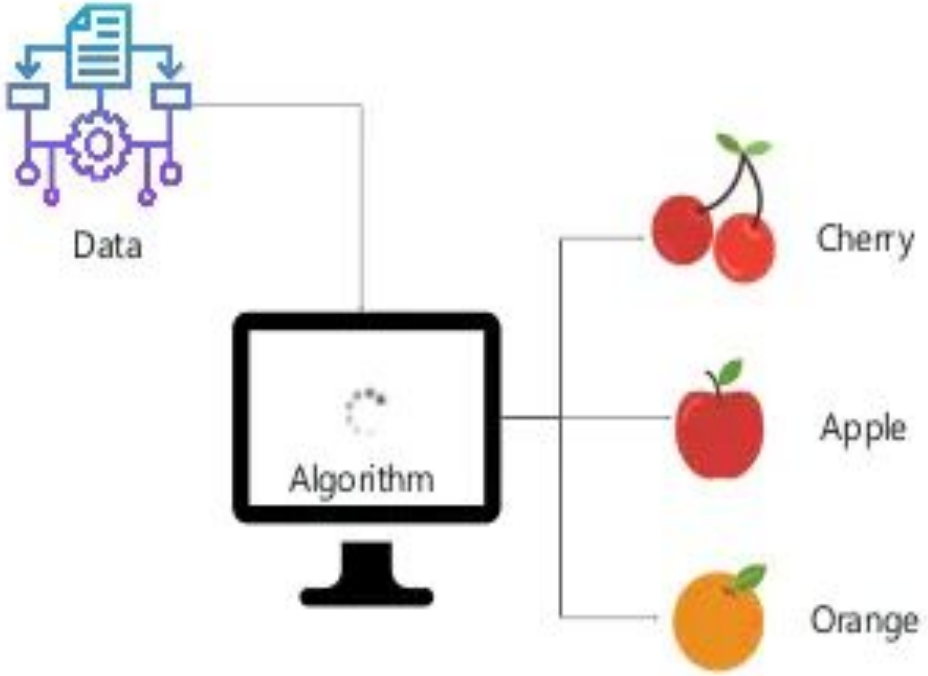
- Tons and Tons of Data Generation
- Need to analyse data behaviour => Business Growth
- Tech Giants investing in it, Ex -

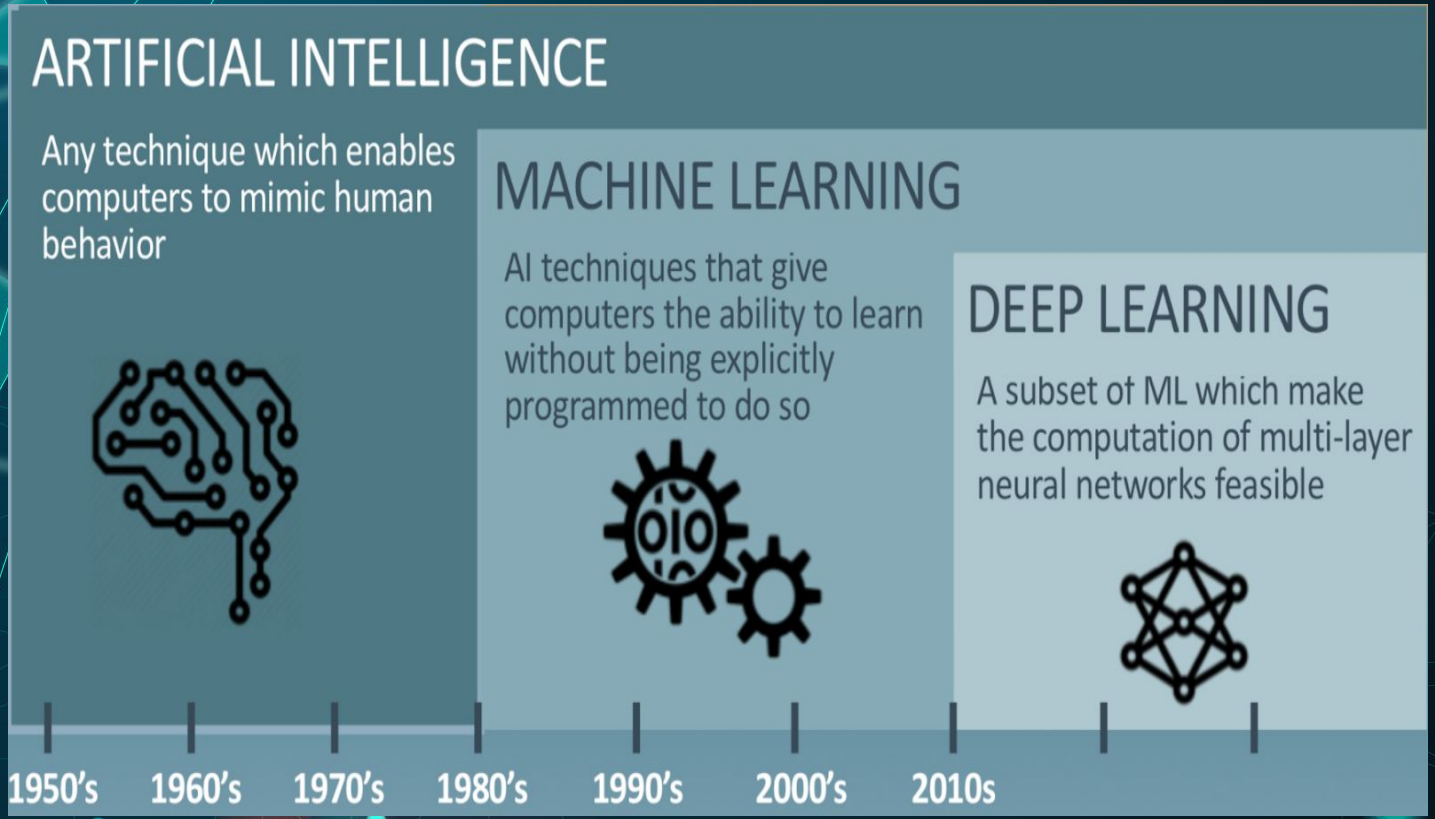
# LET’S START WITH TERMINOLOGIES

# What Is Machine Learning?

*Machine learning is a subset of artificial intelligence (AI) which provides machines the ability to learn automatically & improve from experience without being explicitly programmed.*

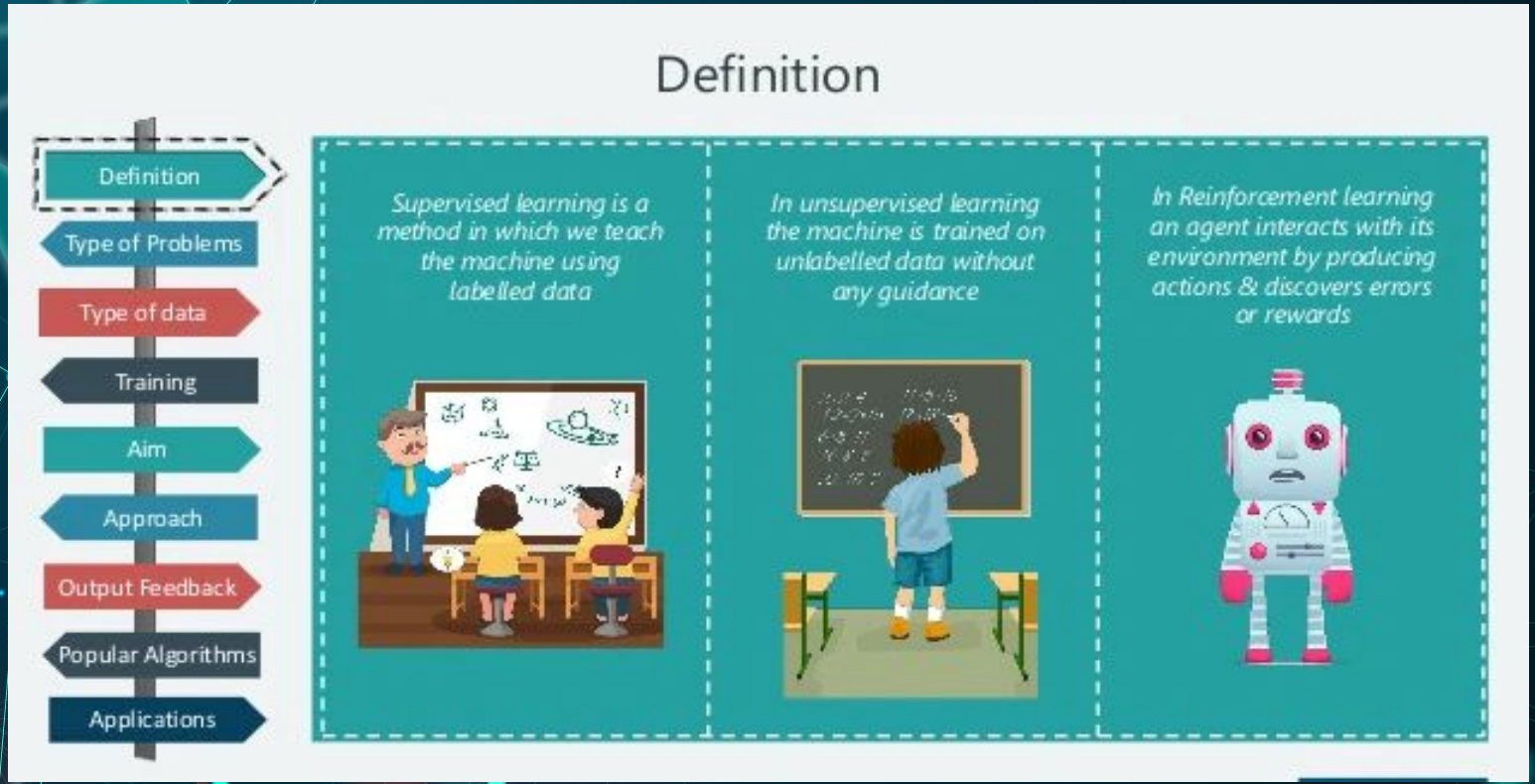
# AI/ML/DL


ARTIFICIAL INTELLIGENCE
Any technique which enables computers to mimic human behavior
MACHINE LEARNING
AI techniques that give computers the ability to learn without being explicitly programmed to do so
DEEP LEARNING
A subset of ML which make the computation of multi-layer neural networks feasible
1950's
1960's
1970's
1980's
1990's
2000's
2010s

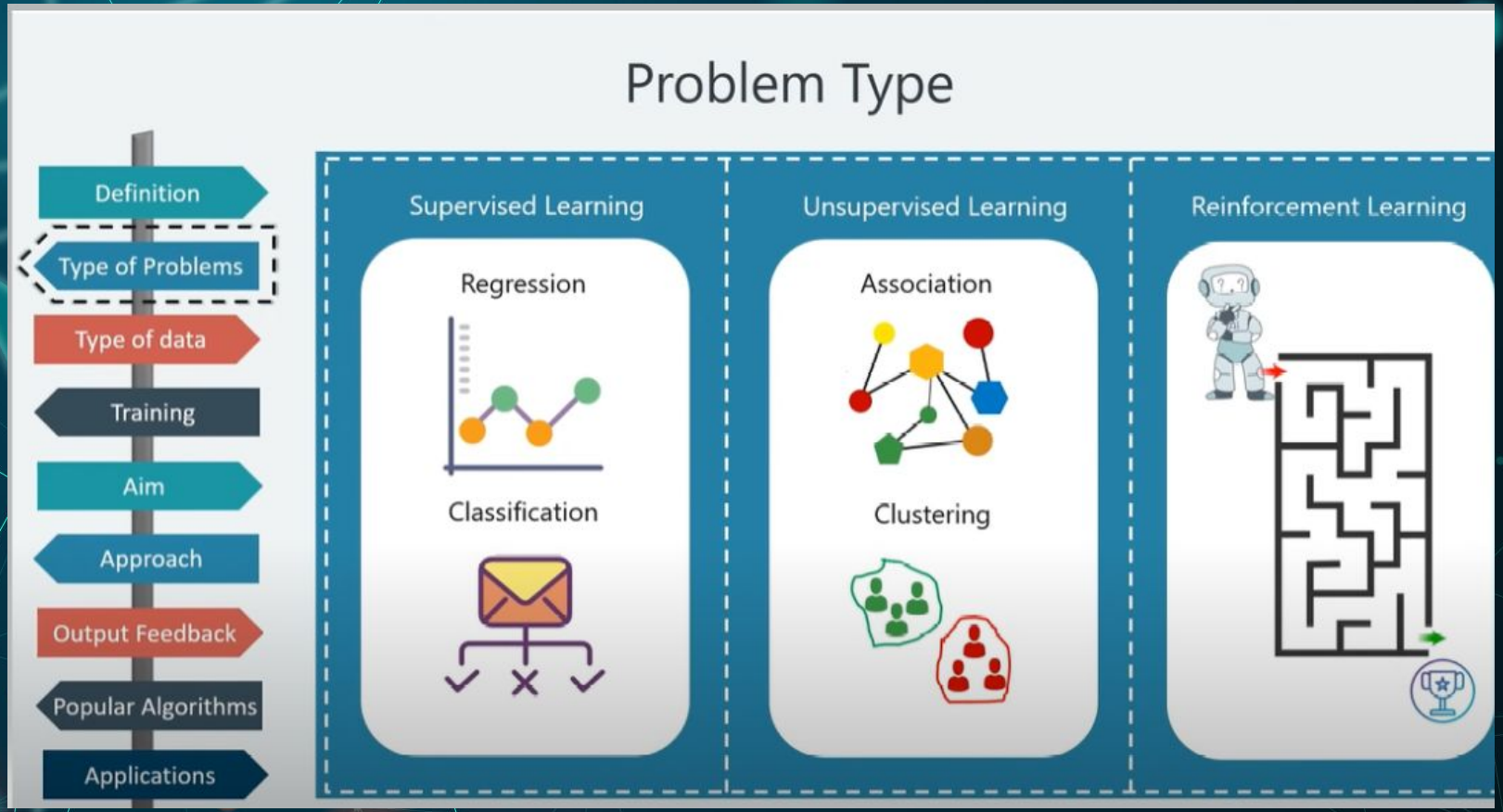
# Mind-Blowing Applications Of AI

- Tonight Showbotics: Jimmy Meets Sophia the Human-Like Robot
- 4 Mind-Blowing Ways Facebook Uses Artificial Intelligence
- The Machine Learning Behind Alexa's AI Systems
- Introduction to Deep Learning and Self-Driving Cars

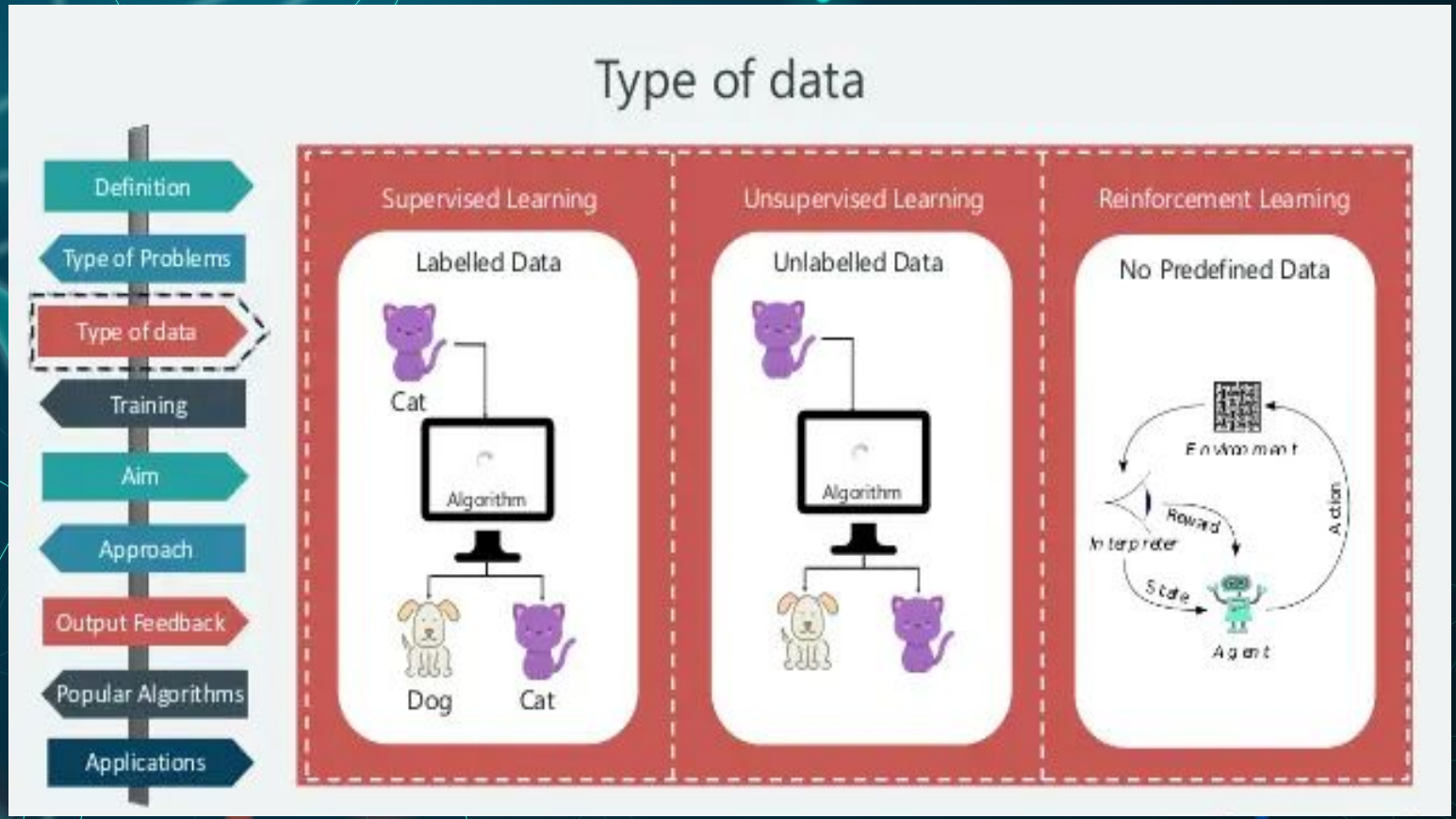
# SUPERVISED/UNSUPERVISED/REINFORCEMENT




Source: - Edureka

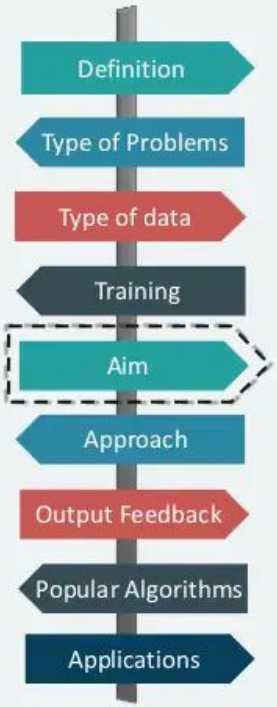


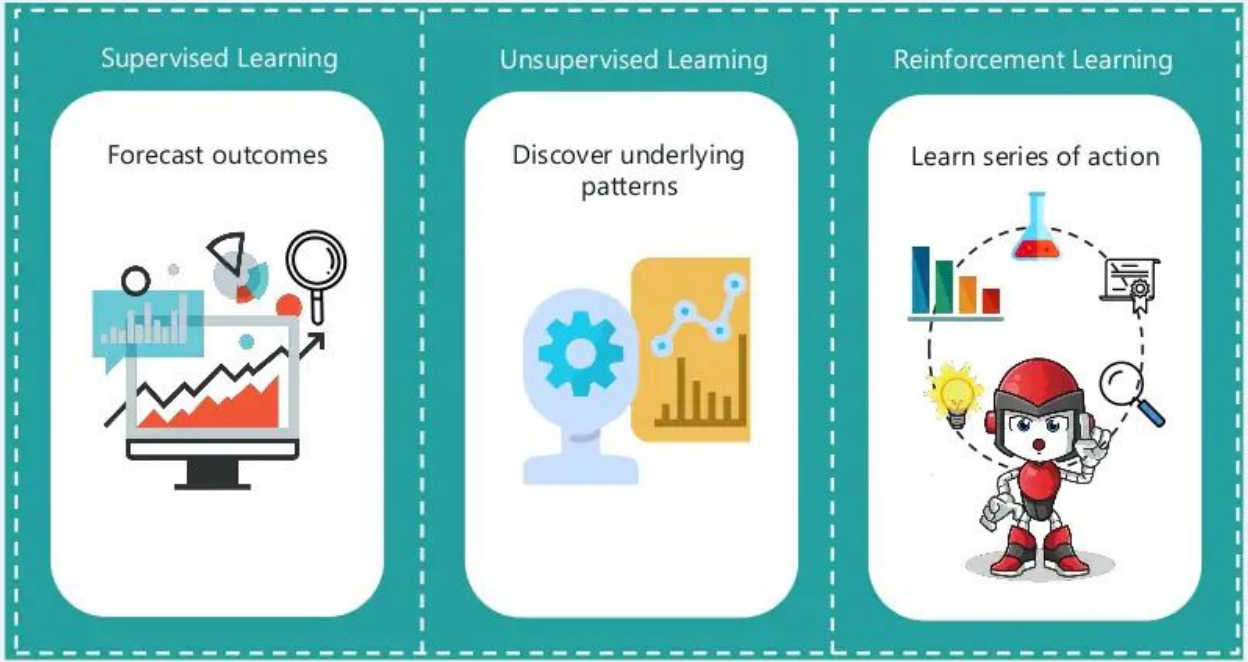

**Source: - Edureka**




Source: - Edureka

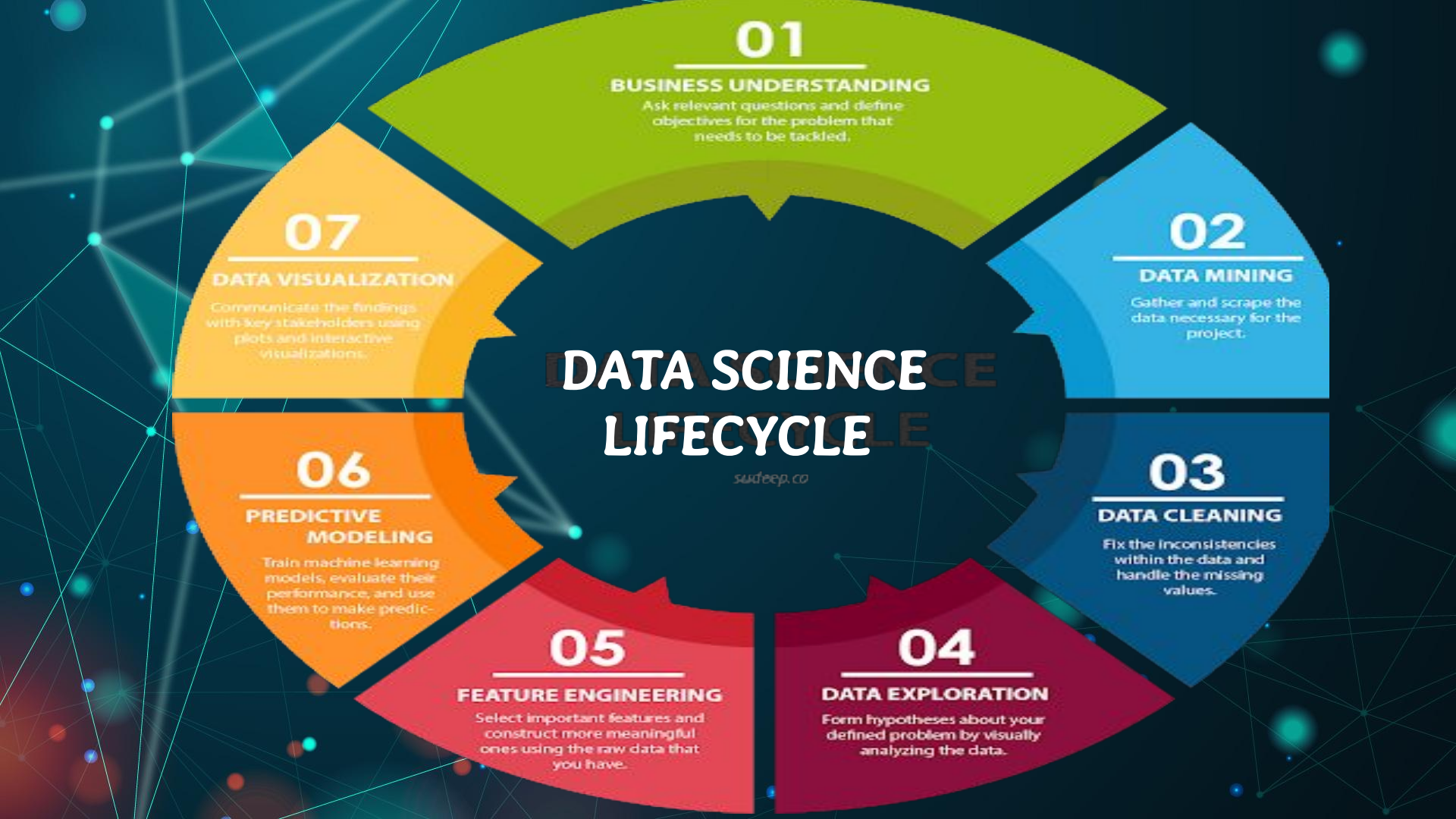
# Aim

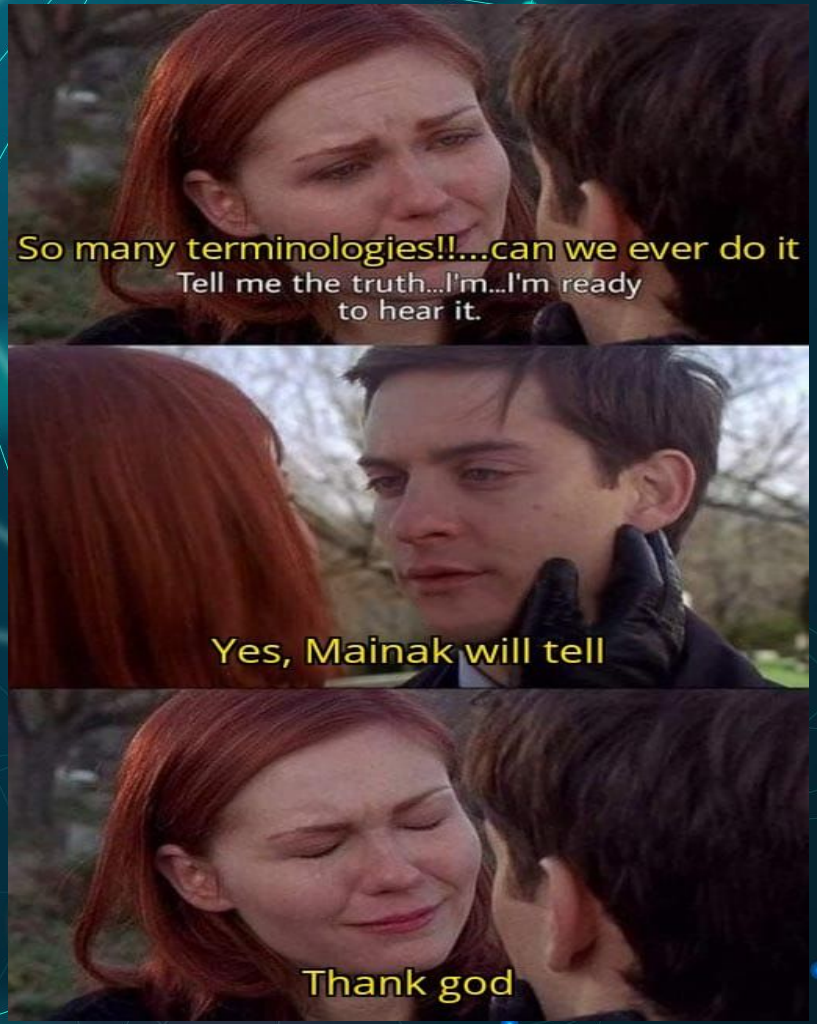



edureka!

Source: - Edureka


01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.
DATA SCIENCE LIFECYCLE
sudeep.co


So many terminologies!!...can we ever do it
Tell me the truth...I'm...I'm ready
to hear it.
Yes, Mainak will tell
Thank god

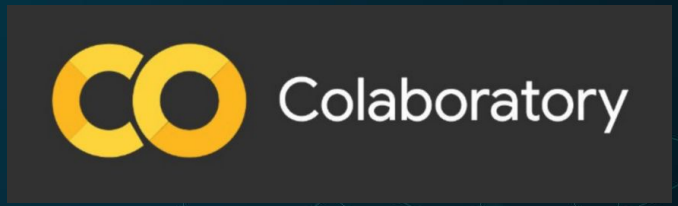
# Data Science Process

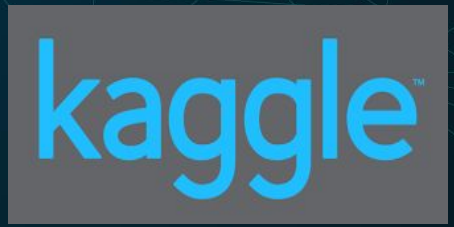

| O | S | E | M | N |
|---|---|---|---|---|
| Gather data from relevant sources | Clean data to formats that machine understands | Find significant patterns and trends using statistical methods | Construct models to predict and forecast | Put the results into good use |

Originally by Hilary Mason and Chris Wiggins

# Platform setup

- First of all we have to set up our python environment.
- For our data science project we can use different platforms like…….
  - Kaggle
  - Google Colab
  - Anaconda with Jupyter Notebook in your PC

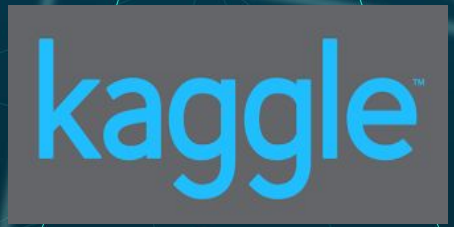



# Obtain

For Data science projects the first requirement is to have some proper data.
In this step we need to collect our own data or fetch some open source datasets from some web sites.

For example: Kaggle, UCI Machine Learning Repository, Data.gov etc.





# Scrub

- In this part we have to convert the data from one format to another and consolidate everything into one standardized format across all data.
- Also in this step we need to check for missing or garbage values and replace them accordingly.
- Libraries used: Python, Pandas, Numpy etc.



# Explore

- In this step we have to go through all the features of the data and inspect its properties to extract features and test significant variables and visualize them to identify significant patterns in our data.
- Libraries used: Numpy, Pandas, Matplotlib, Seaborn etc.




# Model

- In this step we pass the preprocessed dataset in to our machine learning algorithm to predict future states or classify them into different categories or group values using clustering.
- We can also optimize our algorithms to get better results.
- Libraries used: Sci-kit Learn, Keras, XGBoost, Optuna etc.




# Interpret

- Interpreting data refers to the presentation of your data to a non-technical layman.
- We need to visualize the results and findings obtained throughout the process.
- Libraries used: Matplotlib, Seaborn etc.


# Hands On Time

# USEFUL YOUTUBE PLAYLIST AND VIDEOS

- Krish Naik Machine Learning Playlist (Very Useful)
- StatQuest Machine Learning Playlist (Very Useful)
- Downloading Python and Pycharm Installation
- How to Read Dataset in Google Colab from Google Drive
- What is Kaggle?
- Install Anaconda with Jupyter Notebook
- Opening Jupyter Notebook on Windows

# USEFUL LINKS ,REFERENCES AND SITES

- Colaboratory
- Jupyter
- Kaggle
- Medium:Data Science
- Analytics Vidhya
- Coursera