Jawaharlal Nehru Technological University Hyderabad

Kukatpally, Hyderabad - 500 085, Telangana, India

DATA SCIENCE

Session 1

6 PM , 19 Sep 2022

Dr N V Ganapathi Raju

Professor, HOD IT,

G.R.I.E.T. , Hyderabad

Learning Outcomes

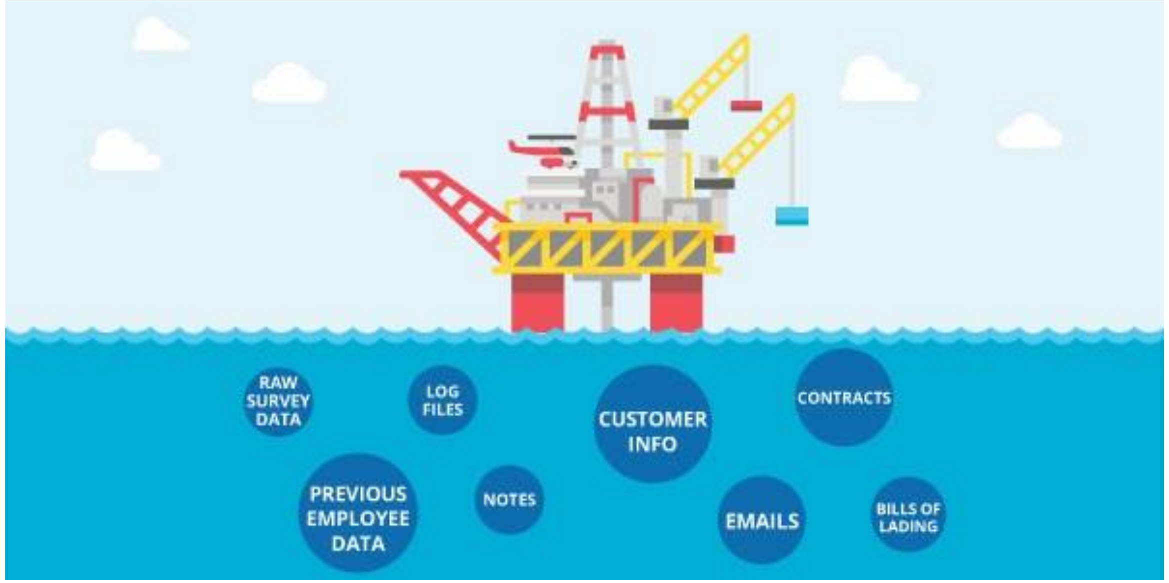
- Participants are going to learn what is data
- Understand kinds of data
- Case studies based on data
- Understanding of Data Analytics , Data Science , Machine Learning , Deep Learning
- Applications of Data Science
- Steps in Data Science Life Cycle
- Job Roles for Data Science

Data

- Quantities 1,2,3,…100,…1000…
- Characters A, B, C…Z, a, b, c,…z
- Symbols ! @ # $ % & * ()…
- Quantities , Characters, Symbols are stored in digital format.
- Data → plural Datum→ singular
- Data is Every Where
- Machines, Robots, Sensors, Our self are products of data.
- All roads lead to **DATA.**

Numerous kinds of data

- **Text data** (.doc, .txt, .pdf…)
- **Excel data** (.csv, .tsv)
- **HTML data**
- **XML data**
- **JSON data**
- **Relational Database** (Oracle, MySql, Sqlite…)
- **Log files / Transactional data**
- **Sensors/Web servers**
- **Social Media data** (FB, Twitter, WhatsApp, YouTube…)
- **Image / Audio / Video / Signal….**



Data

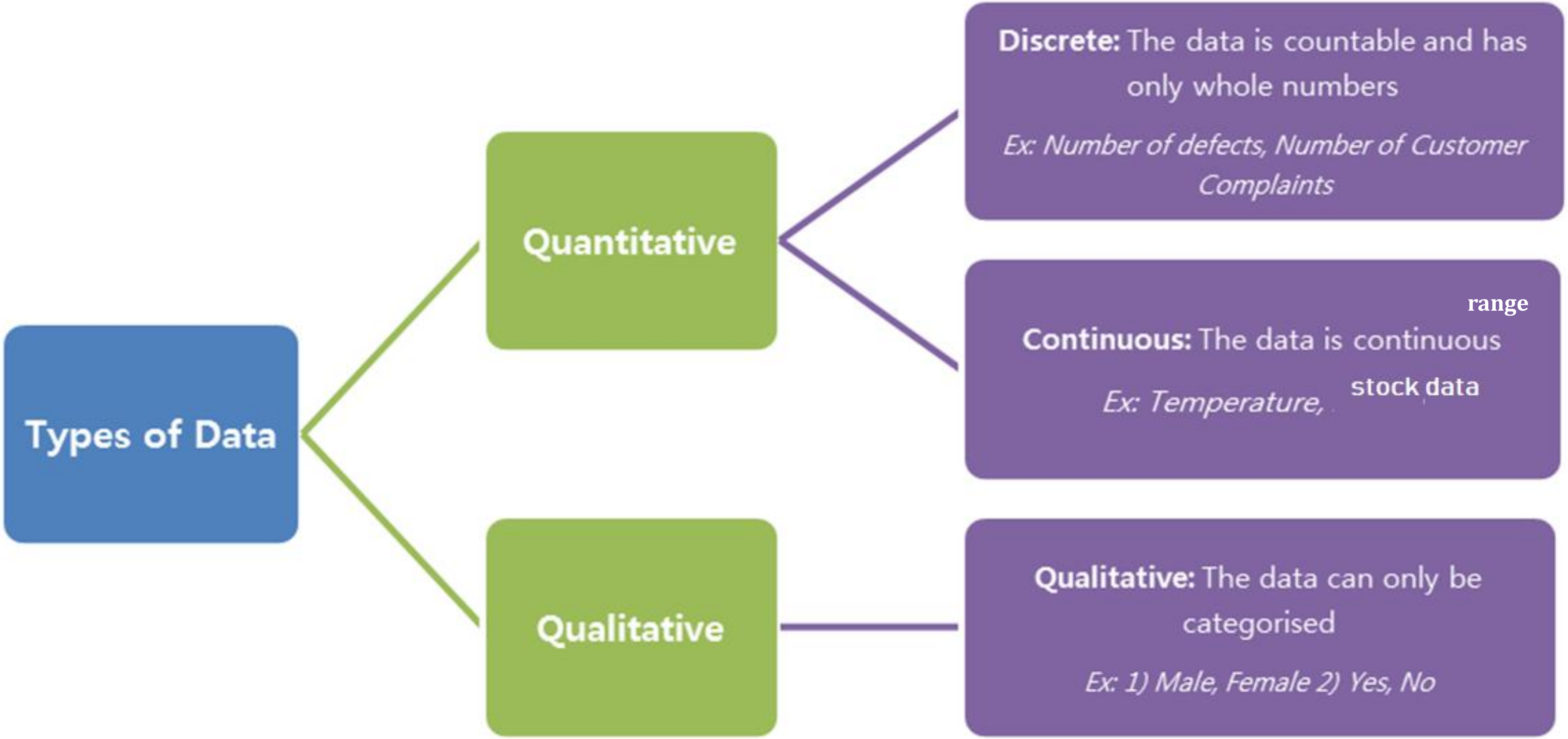
- **Data is useful in a refined form.**
  - **Data to Information**
  - **Information to Knowledge**

    (hidden knowledge/pattern/relationships/insights)
  - **Knowledge to Decision making**
  - **Decision making to Result cycle**.

Data

DATA IS GROWING FASTER THAN EVER BEFORE.
Each person-
1.7 megabytes
created
1sec

- Data is growing at an incredible rate.
- Gartner state that data is doubling every 18 months
- Current estimate is that there is over 4 zeta bytes of data in the world
- **Facebook** users send on average 31.25 million messages and view 2.77 million videos every minute.
- A massive growth in video and photo data, where every minute up to 300 hours of video are uploaded to **YouTube** alone.



Types of Data

Quantitative

Qualitative

Discrete: The data is countable and has only whole numbers

Ex: Number of defects, Number of Customer Complaints

range

Continuous: The data is continuous

Ex: Temperature, stock data

Qualitative: The data can only be categorised

Ex: 1) Male, Female 2) Yes, No

Types of Data

- Structured data

- Unstructured data

- Semi Structured data

Types of Data

Structured Data	Unstructured Data
Structured data is quantitative and is often displayed as numbers, dates, values, and strings.	Unstructured data is qualitative data and includes text, video, audio, images, and more.
Structured data is stored in rows and columns.	Unstructured data is stored as audio, text, and video files, or NoSQL databases.
Estimated 20% of business data.	Estimated 80% of business data
Stored in data warehouses.	Stored in applications, NoSQL databases, data lakes, and data warehouses.
Reveals patterns and trends that show you what’s happening.	Reveals patterns and trends that explain why something is happening.
Requires less storage space.	Needs more storage space.
Easy to analyze with tools like Excel.	Hard to analyze without AI tools.

Students for DS Course

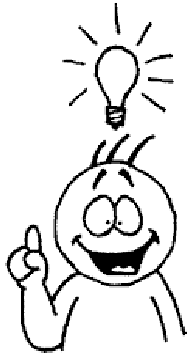
USA	China	USA	Sweden	China
Canada	China	Japan	Mexico	USA
China	Germany	India	India	Japan
USA	USA	USA	China	China
India	Japan	England	India	Japan
Englad	India	China	Mexico	USA
Mexico	USA	Canada	Pakistan	India
Japan	China	USA	Japan	Germany
China	India	India	China	China
Germany	Japan	China	USA	Japan

Think?

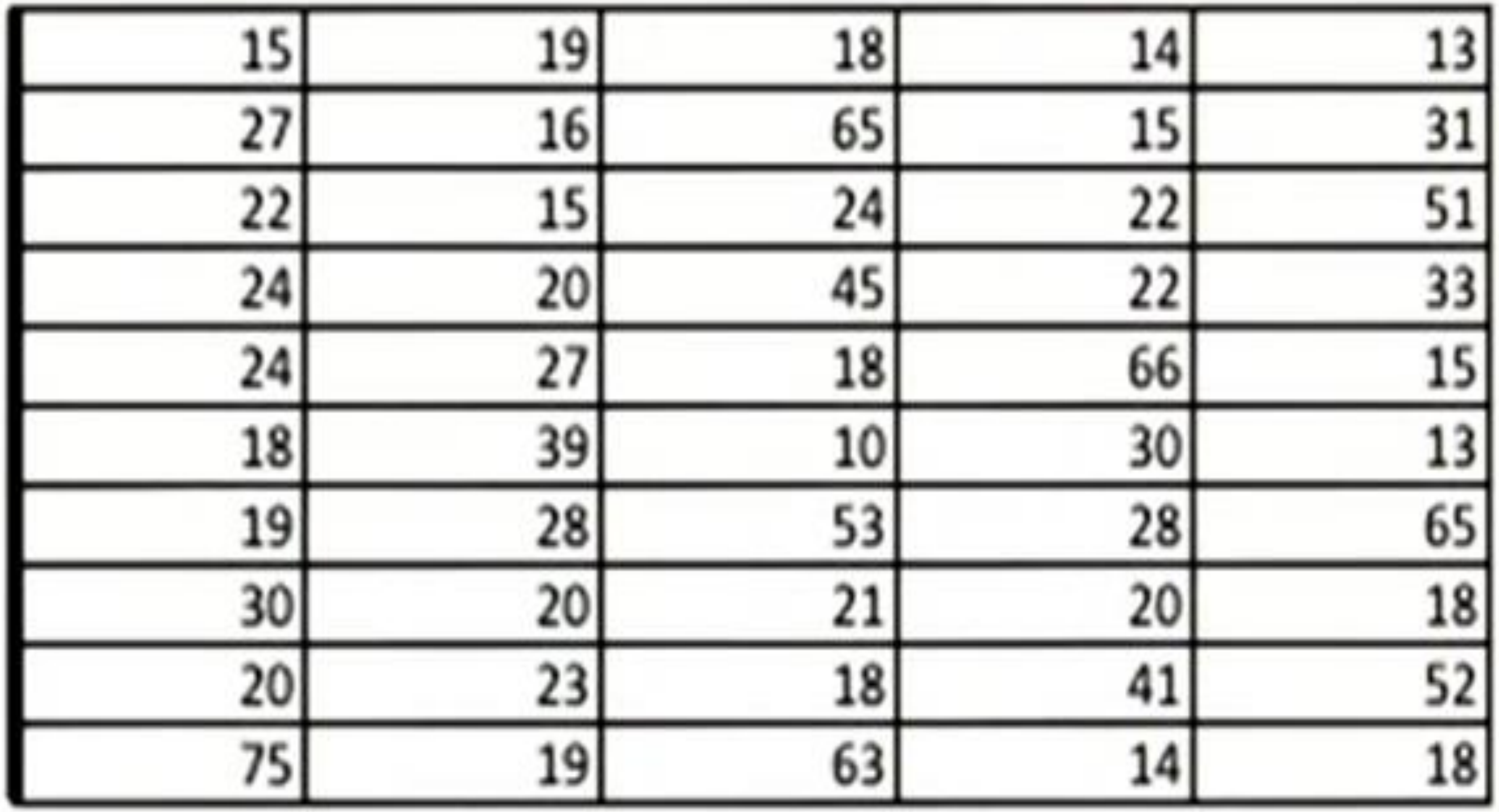
Students for a DS course

Country	Frequency	Proportion	Percent
Canada	2	0.04	4
China	12	0.24	24
England	2	0.04	4
Germany	3	0.06	6
India	8	0.16	16
Japan	8	0.16	16
Mexico	3	0.06	6
Pakistan	1	0.02	2
Sweden	1	0.02	2
USA	10	0.2	20
Total	**50**	**1**	**100**

Analyze?

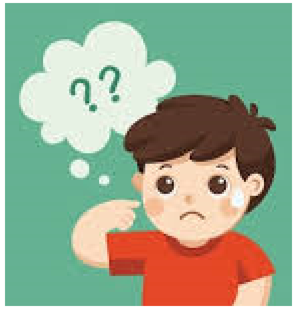


Sample Student Ages



15	19	18	14	13
27	16	65	15	31
22	15	24	22	51
24	20	45	22	33
24	27	18	66	15
18	39	10	30	13
19	28	53	28	65
30	20	21	20	18
20	23	18	41	52
75	19	63	14	18

Think?

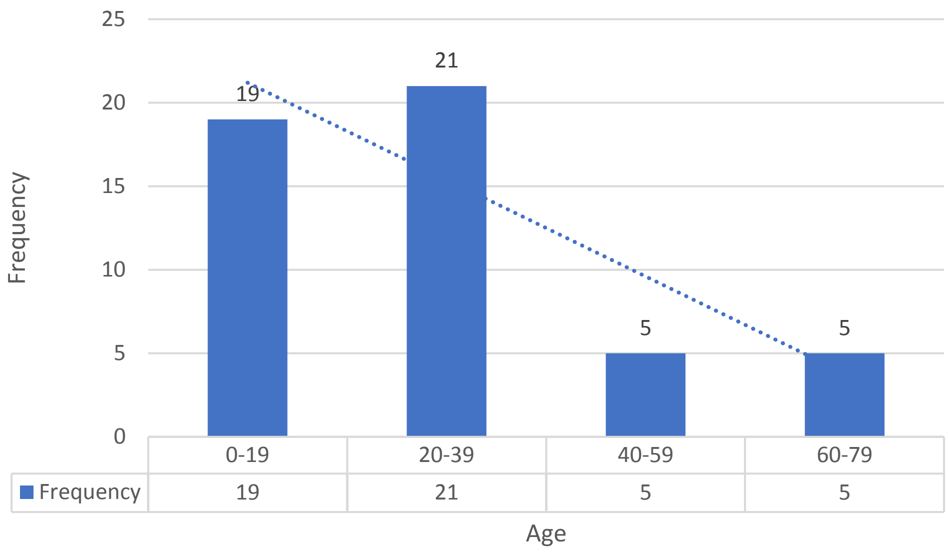


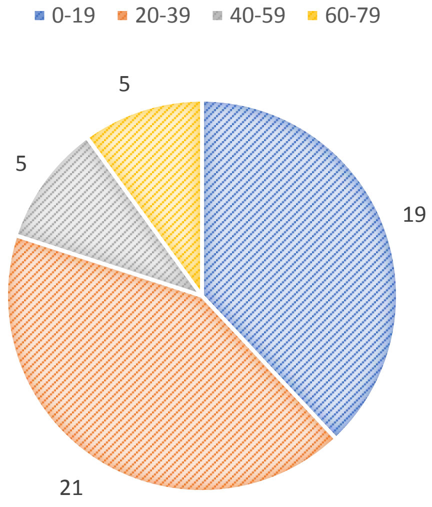
Sample Student Ages

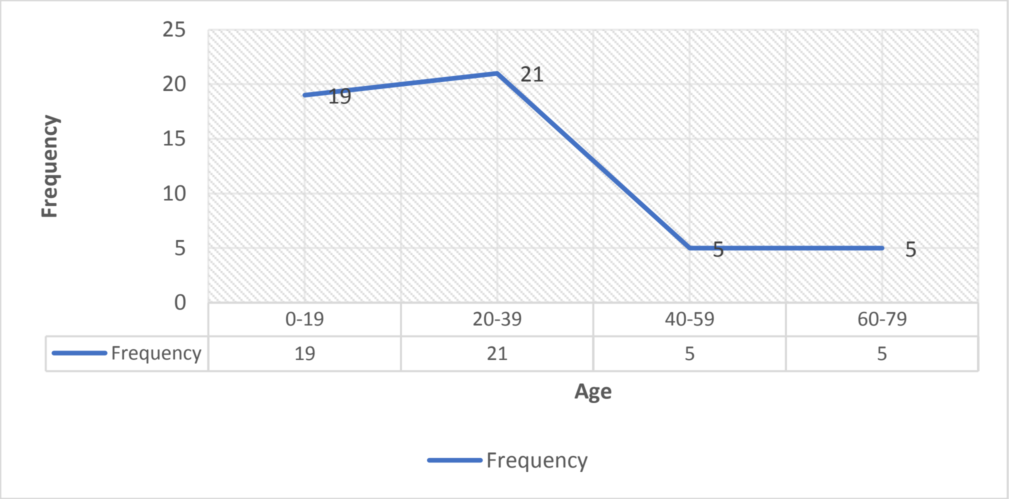
Age	Frequency
0-19	19
20-39	21
40-59	5
60-79	5

Analyze?

Sample Student Ages







Interpret ?

Data Analytics

- Analytics : Exploring and analyzing large datasets to find hidden pattern/ unseen trends / discover correlations / derive valuable insights to make business predictions, it improves speed and efficiency of business.

- Analytics is the transformation of data into insights

- Analytics involves
  - Understanding the **past and current performance to predict future performance**
  - Understanding the **relations**, identifying **patterns** and **translating them to meaningful, useful and relevant business insights and intelligent strategies**
  - Laying foundation for a data driven **decision making process** in an enterprise.

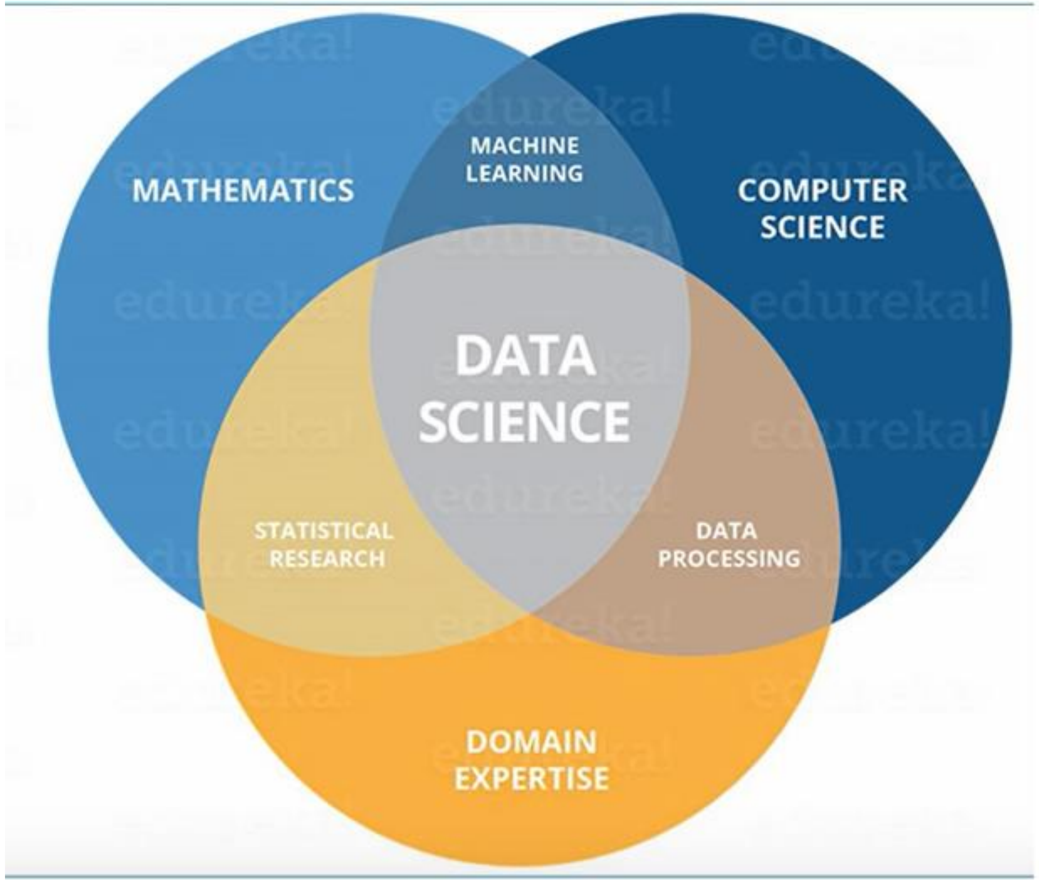
Data Science

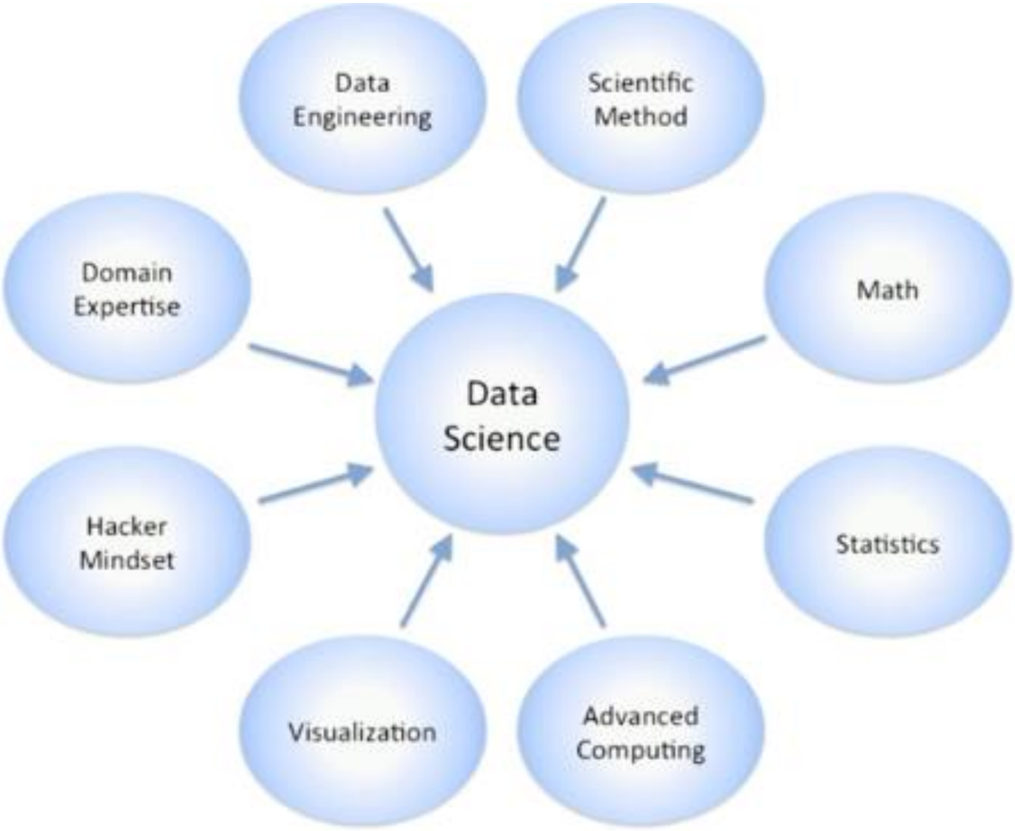
- **“The ability to extract knowledge and insights from large and complex data sets.”**

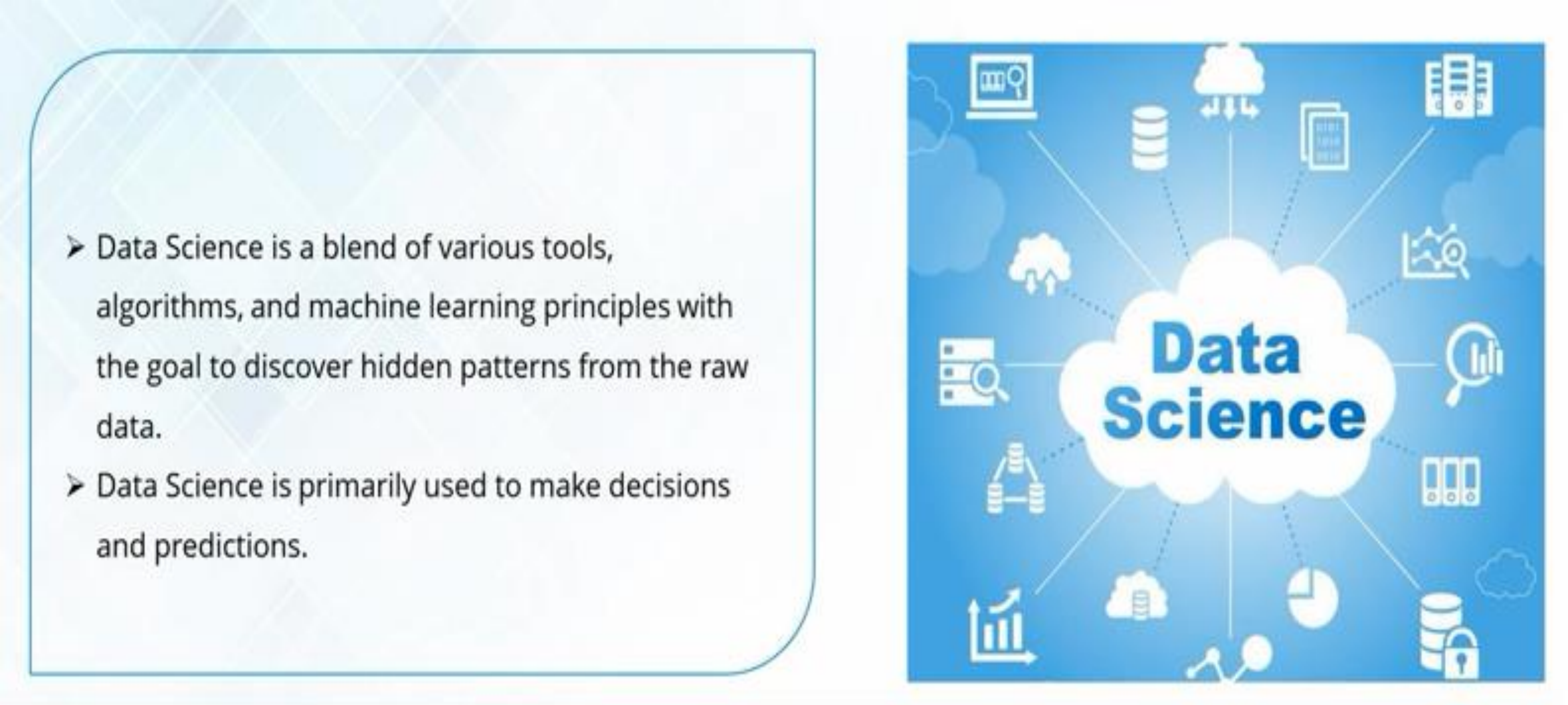
  [D J Patil]

- Data scientists combine statistics, mathematics, programming, problem-solving, capturing data in ingenious ways, the ability to look at things differently to **find patterns**, along with the activities of **cleansing, preparing, and aligning** the data.

Areas of Data Science

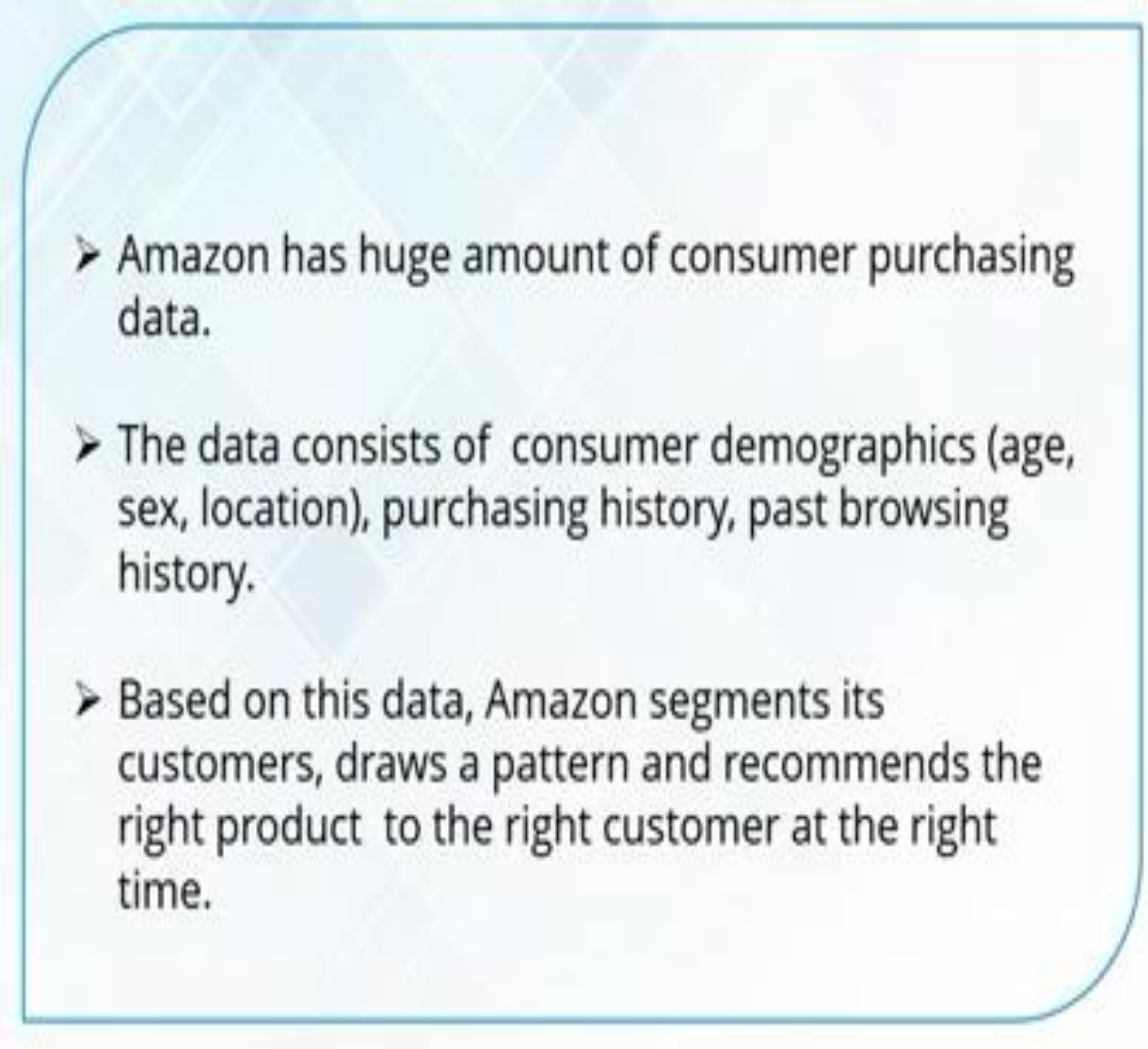






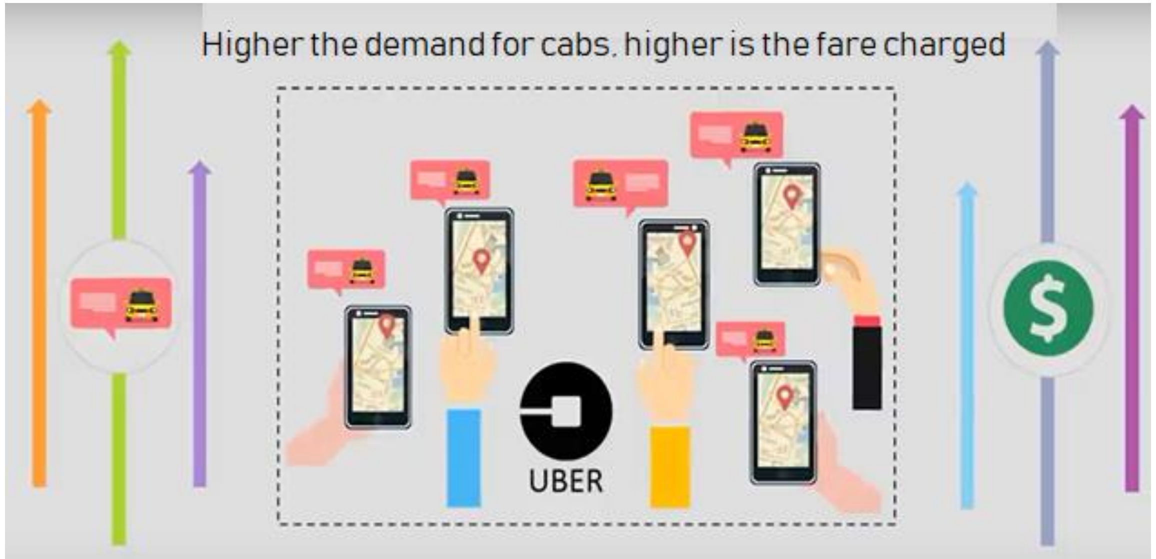
- Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data.
- Data Science is primarily used to make decisions and predictions.

Data
Science

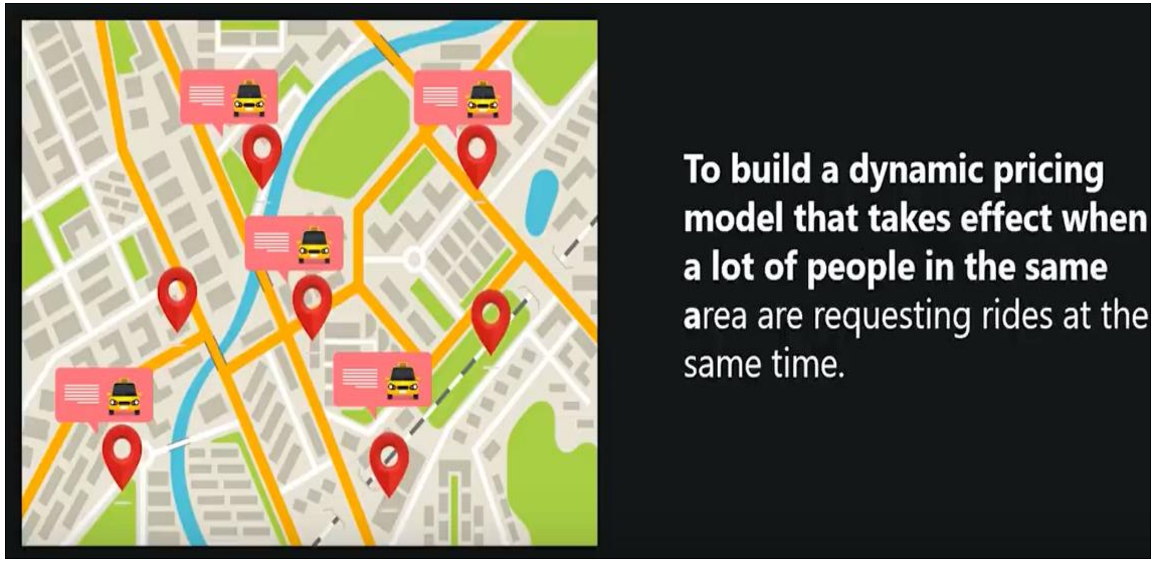


- Amazon has huge amount of consumer purchasing data.
- The data consists of consumer demographics (age, sex, location), purchasing history, past browsing history.
- Based on this data, Amazon segments its customers, draws a pattern and recommends the right product to the right customer at the right time.

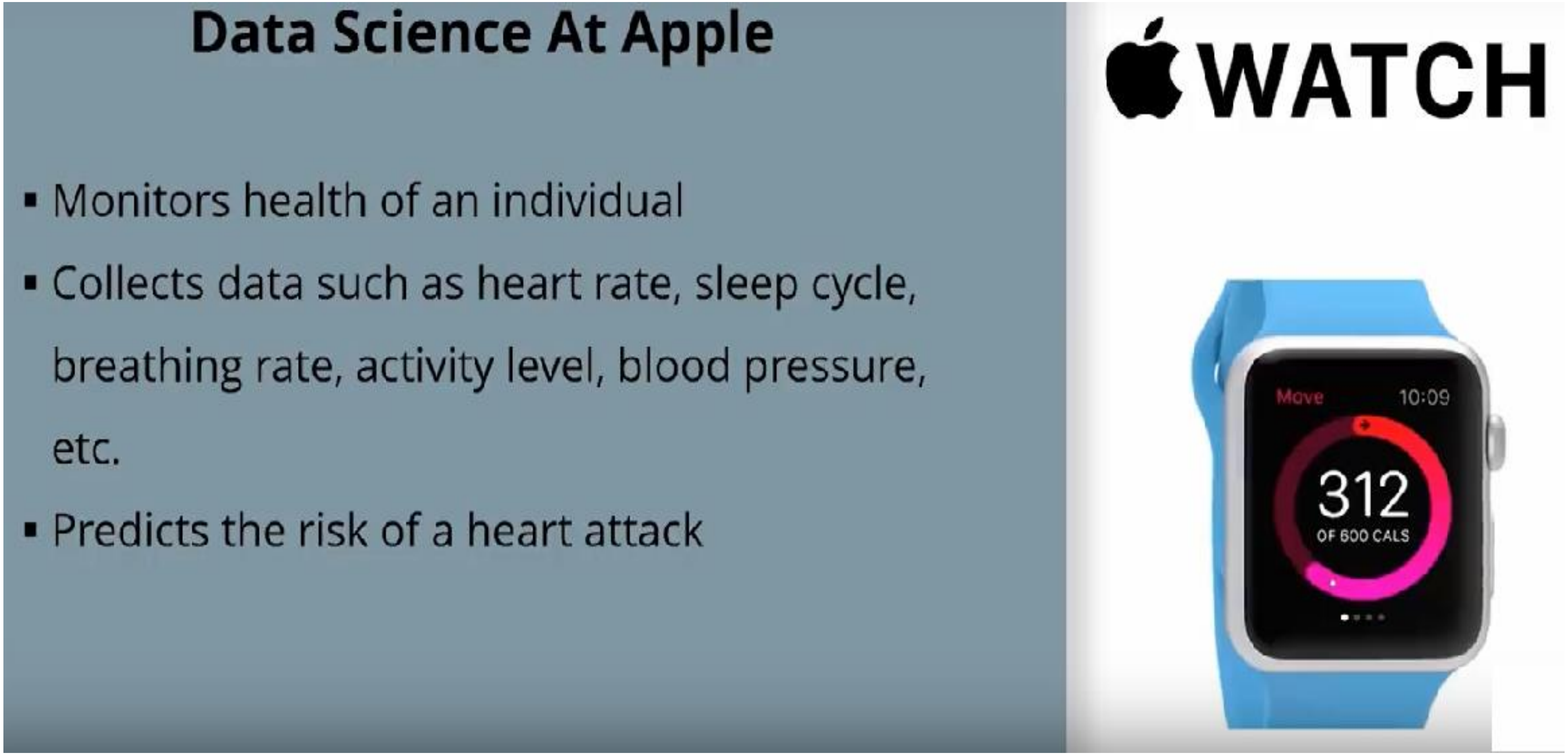
UBER



Higher the demand for cabs, higher is the fare charged

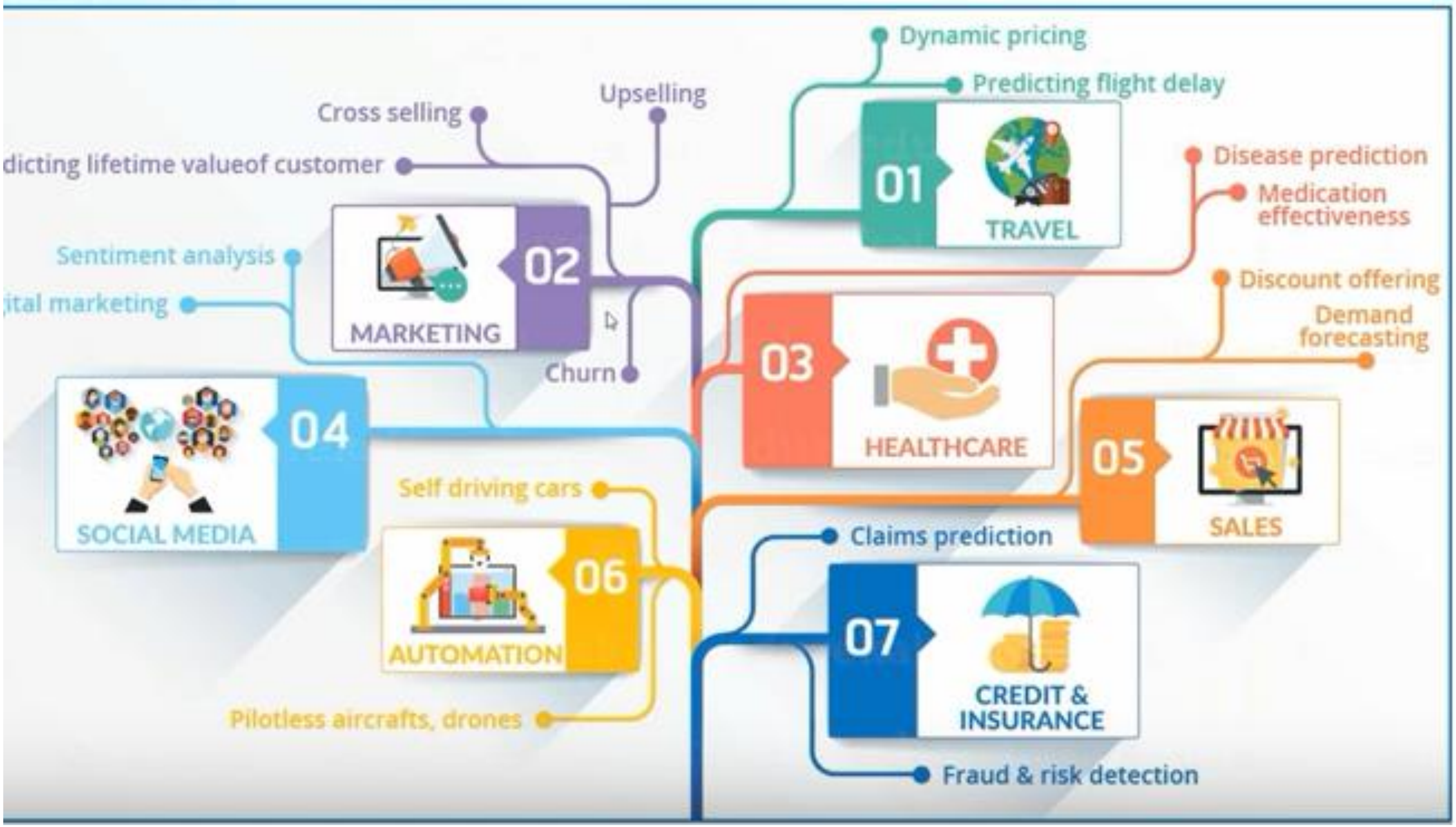


To build a dynamic pricing model that takes effect when a lot of people in the same area are requesting rides at the same time.

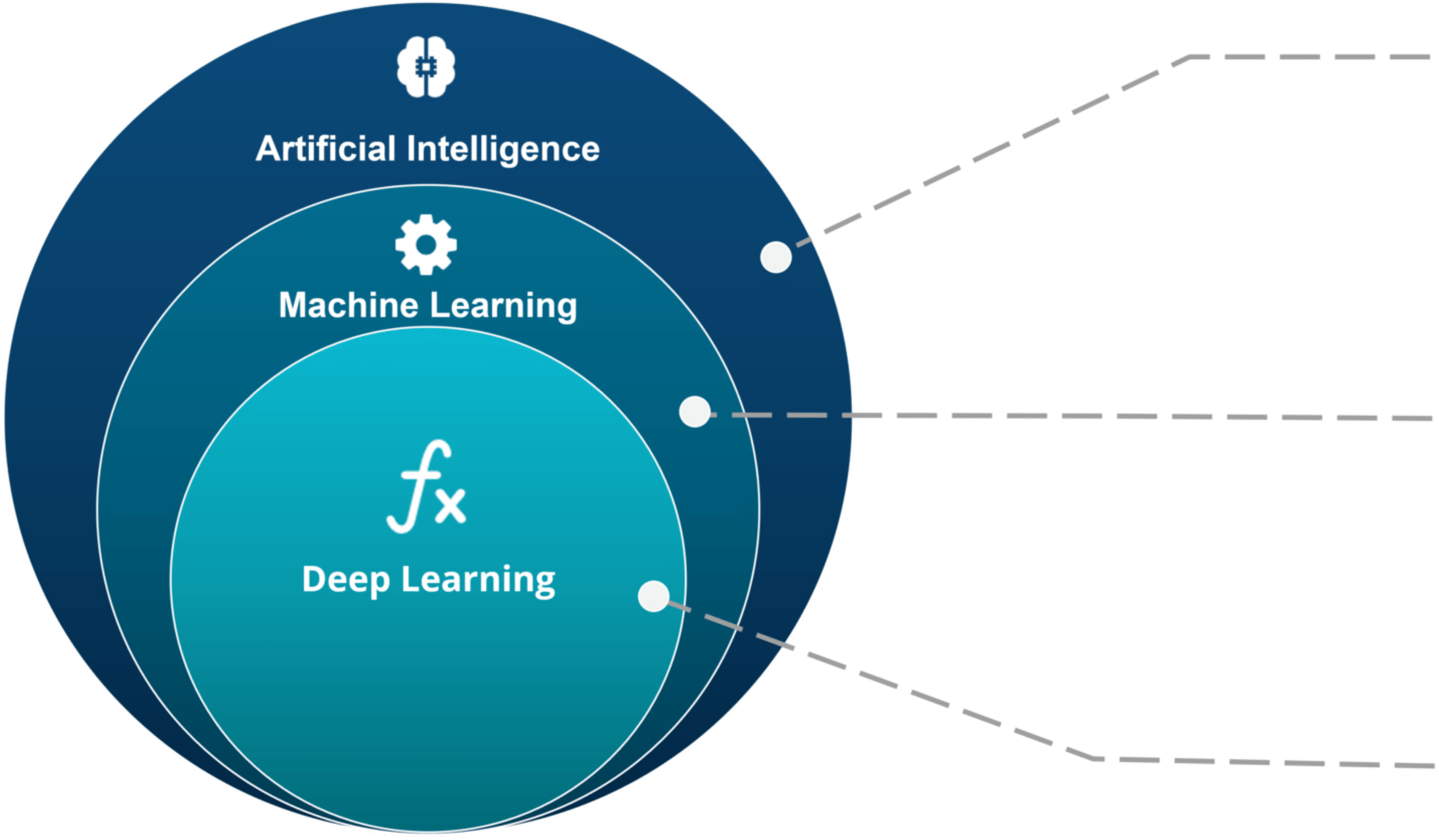


Data Science At Apple

- Monitors health of an individual
- Collects data such as heart rate, sleep cycle, breathing rate, activity level, blood pressure, etc.
- Predicts the risk of a heart attack


Dynamic pricing
Predicting flight delay
01
TRAVEL
Upselling
Cross selling
dicting lifetime valueof customer
02
MARKETING
Churn
Sentiment analysis
ital marketing
Disease prediction
Medication effectiveness
03
HEALTHCARE
Discount offering
Demand forecasting
04
SOCIAL MEDIA
05
SALES
Self driving cars
06
AUTOMATION
Pilotless aircrafts, drones
Claims prediction
07
CREDIT & INSURANCE
Fraud & risk detection

AI / ML / DL



ARTIFICIAL INTELLIGENCE

A technique which enables machines to mimic human behaviour

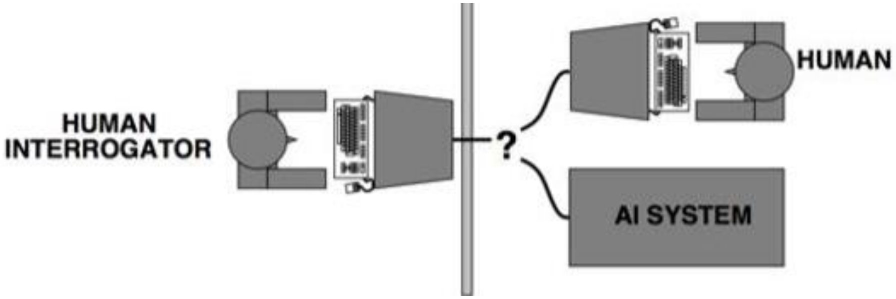
MACHINE LEARNING

Subset of AI technique which use statistical methods to enable machines to improve with experience

DEEP LEARNING

Subset of ML which make the computation of multi-layer neural network feasible

Turing Test approach



- A computer passes the test of intelligence, if it can fool a human int

- The computer passes the test if a human interrogator, after posing some written questions, cannot tell whether the written responses come from a person or not.

- The computer would need to possess the following capabilities:

  - ✓ **natural language processing** to enable it to communicate successfully in English,
  - ✓ **knowledge representation** to store what it knows or hears;
  - ✓ **automated reasoning** to use the stored information to answer questions and to draw new conclusions;
  - ✓ **machine learning** to adapt to new circumstances and to detect and extrapolate patterns
  - ✓ **computer vision** to perceive objects, and
  - ✓ **robotics** to manipulate objects and move about.

Result of Turing Test

- If the interrogator can not reliably distinguish the human from the computer

- Then the computer does posses artificial intelligence

Vocabulary

- **Target:** Predicted category or value of the data (discrete / continuous)

  Column to be predicted

  Response, Output, Dependent Variable, Labels

- **Features:** Properties of the data used for prediction

  Non-Target columns

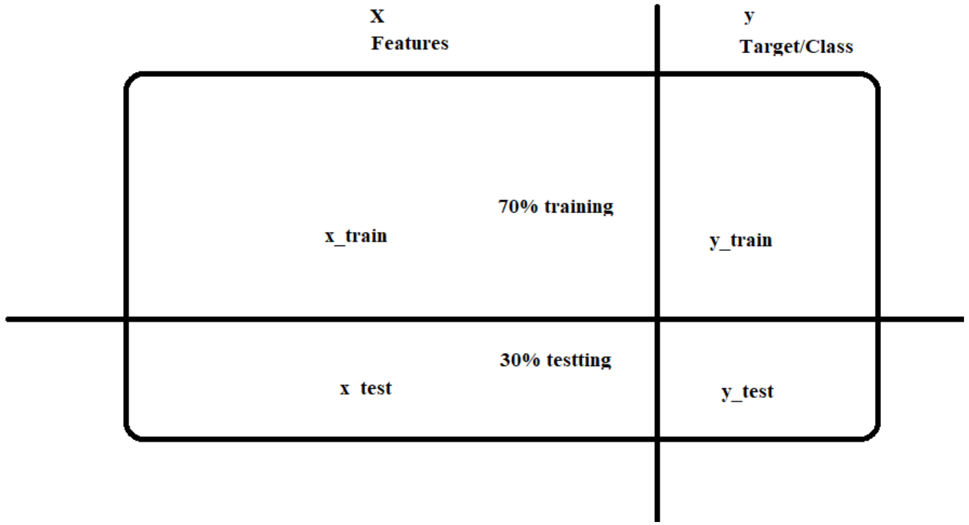
  Predictors, Input, Independent var, attributes

- **Example:** a single data point within the data (one row)

  Observations, Record, Instance, row, data points

- **Label:** The target value for a single data point

  answer, Category, Y axis



X
Features

y
Target/Class

70% training

x_train

y_train

30% testting

x test

y_test

Machine Learning

- Allows computers to learn and
- infer from data

Types of Machine Learning

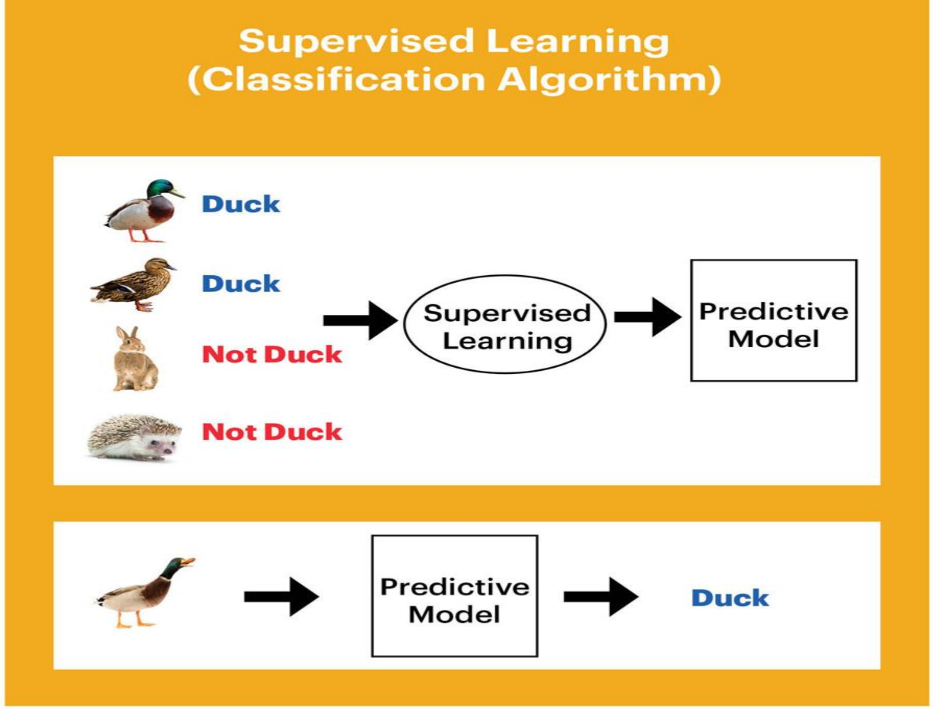
- Supervised
- Unsupervised
- Reinforcement Learning

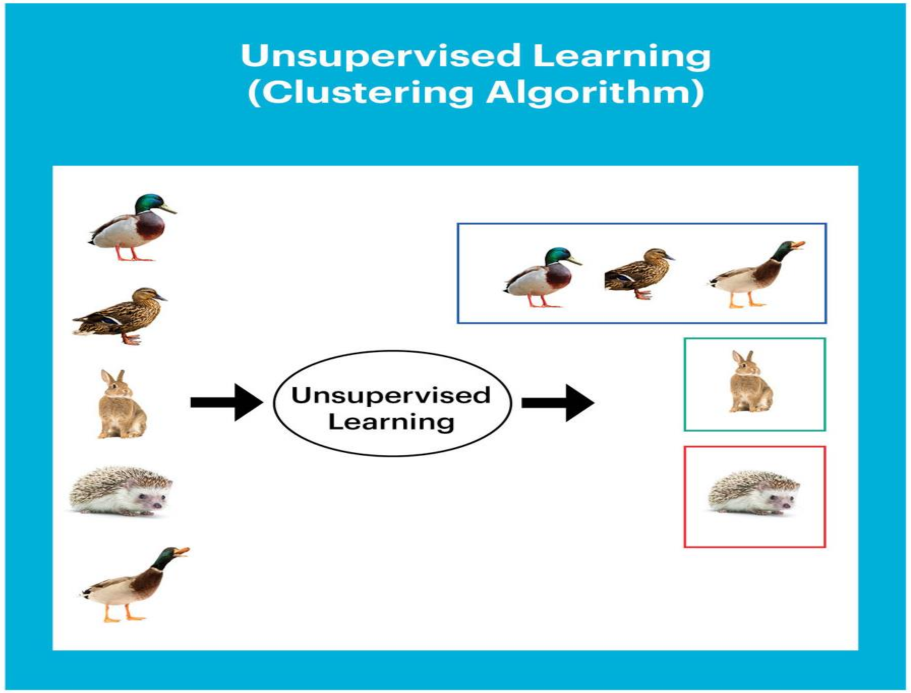
Supervised Learning

- Data points have a known outcome

Unsupervised Learning

- Data points have unknown outcome





Types of Supervised Learning

Regression

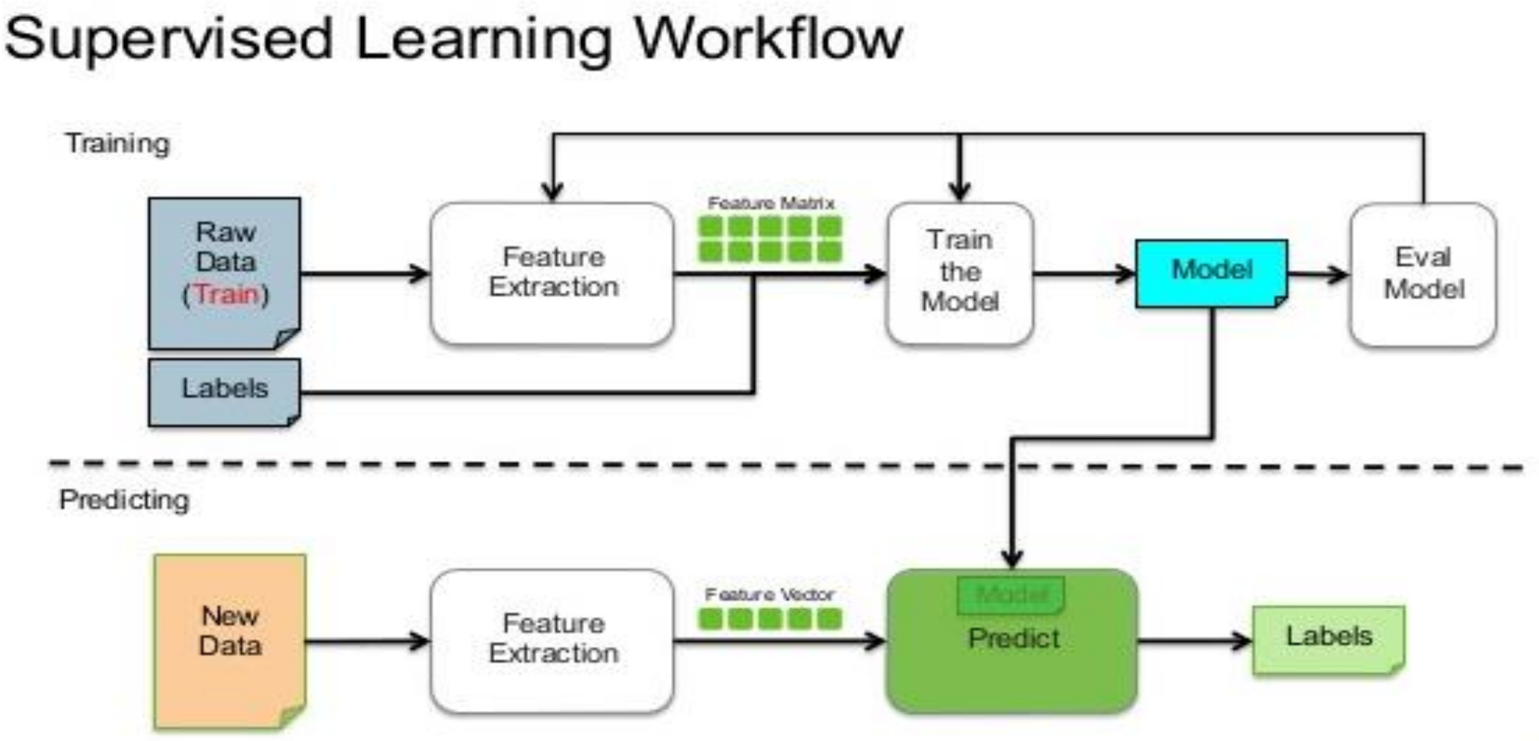
Outcome is continuous (numerical)

Ex:- home prices, happiness index

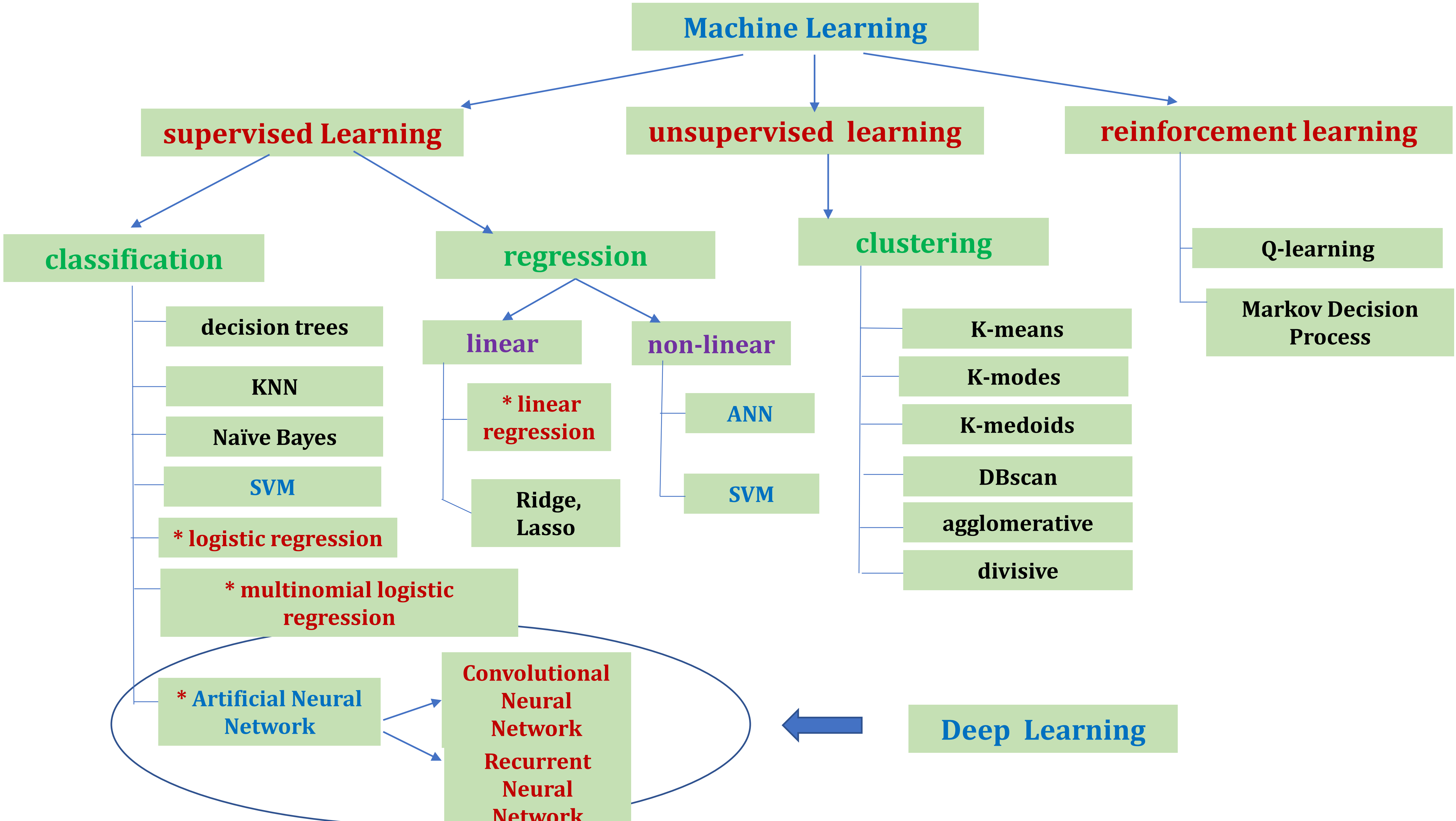
Classification

Outcome is a Category

Ex:- Object classes in Images



Supervised Learning Workflow

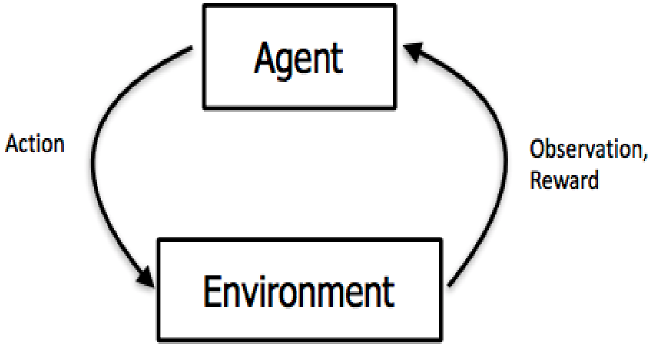


- Machine Learning
  - supervised Learning
    - classification
      - decision trees
      - KNN
      - Naïve Bayes
      - SVM
      - * logistic regression
      - * multinomial logistic regression
      - * Artificial Neural Network
        - Convolutional Neural Network
        - Recurrent Neural Network
    - regression
      - linear
        - * linear regression
        - Ridge, Lasso
      - non-linear
        - ANN
        - SVM
  - unsupervised learning
    - clustering
      - K-means
      - K-modes
      - K-medoids
      - DBscan
      - agglomerative
      - divisive
  - reinforcement learning
    - Q-learning
    - Markov Decision Process

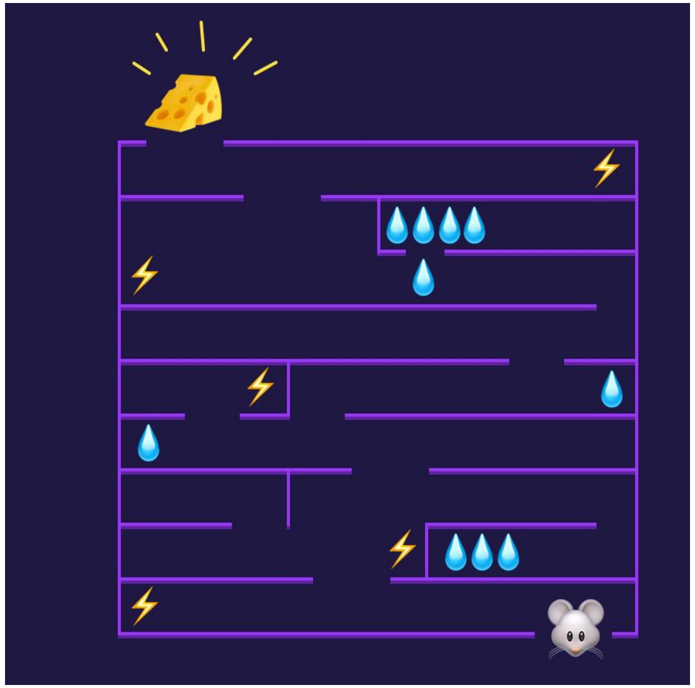
Deep Learning

Reinforcement Learning

- In supervised learning, training data comes with an answer key from some godlike “supervisor
- In **reinforcement learning (RL)** there’s no answer key, but your reinforcement learning **agent** still must decide how to act to perform its task.
- In the absence of existing training data, the agent learns from experience.
- It collects the training examples (“this action was good, that action was bad”) through **trial-and-error** as it attempts its task, with the goal of maximizing long-term **reward**.



The agent **observes** the environment, takes an **action** to interact with the environment, and receives positive or negative **reward.**

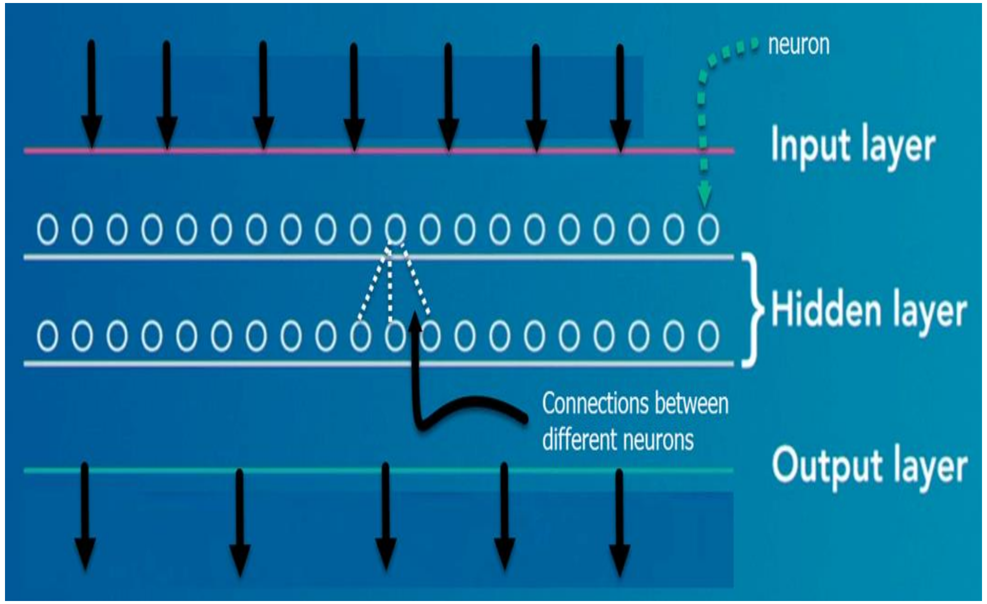


Deep Learning

- Deep Learning is about learning multiple levels of representation and abstraction that help to make sense of data such as images, sound, and text. it makes use of deep neural networks.

- Deep learning mimics the network of neurons in a brain.

- It is a subset of machine learning and is called deep learning because Deep learning algorithms are constructed with connected layers.



Algorithms

- ANN (Artificial neural networks)
- CNN (Convolutional neural networks)
- RNN (Recurrent neural networks)

Applications

- Object detection and Recognition
- Image Captioning
- Computer Vision

Natural language processing (NLP)

- NLP refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.

- Applications of NLP:
  - Chatbots/Voice Assistants
  - Machine Translation
  - Hiring and Recruitment
  - Grammar Checkers
  - Text Classification/Clustering/Summarization
  - Advertisement to Targeted Audience/Survey Analysis

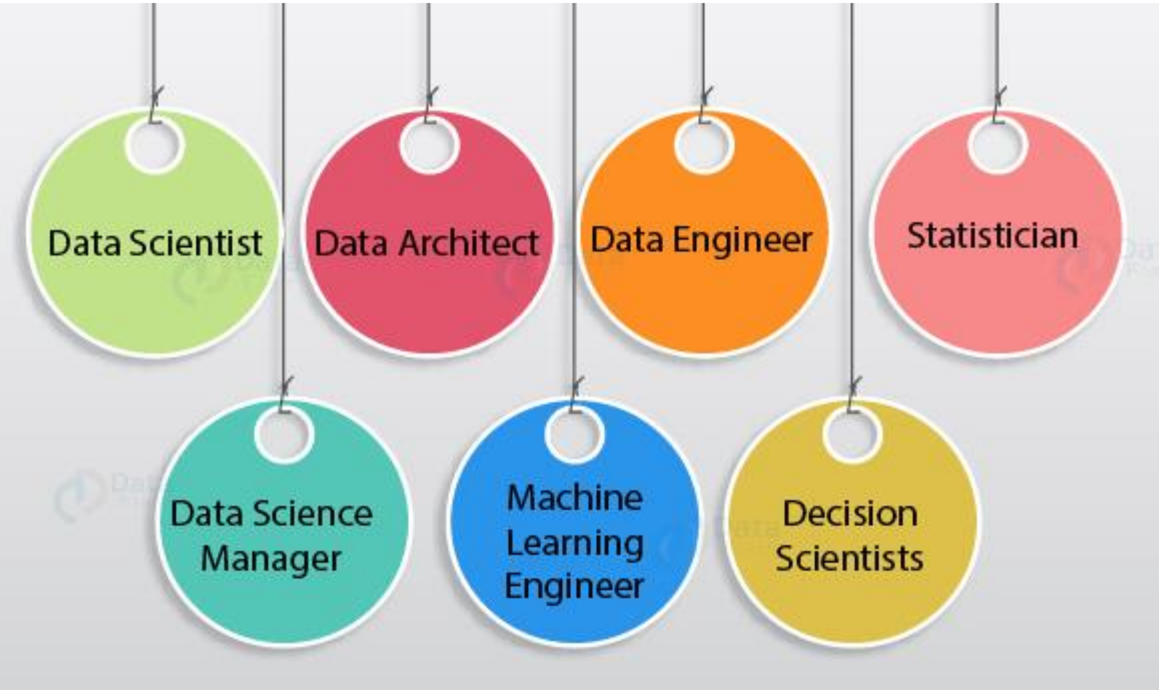
Steps in Data Science

- Problem Identification
- Data Collection/Generation
- Data Preprocessing
- Data Exploration (EDA)
- Feature Selection
- Model Building
- Model Evaluation
- Analyze Results

Data Wrangling / Munging

- Data Imputation
- Data Integration
- Data Encoding / Decoding
- Data Transformation / Normalization
- Dimensionality Reduction
- Feature Engineering

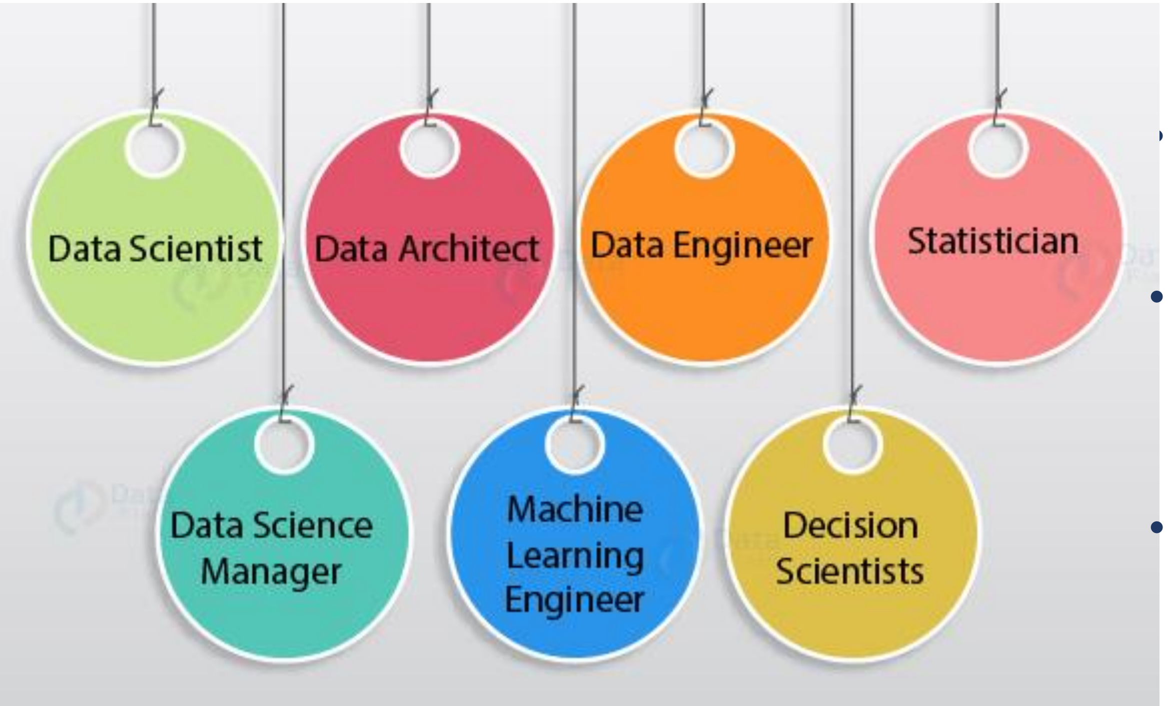
Job Roles Data Science



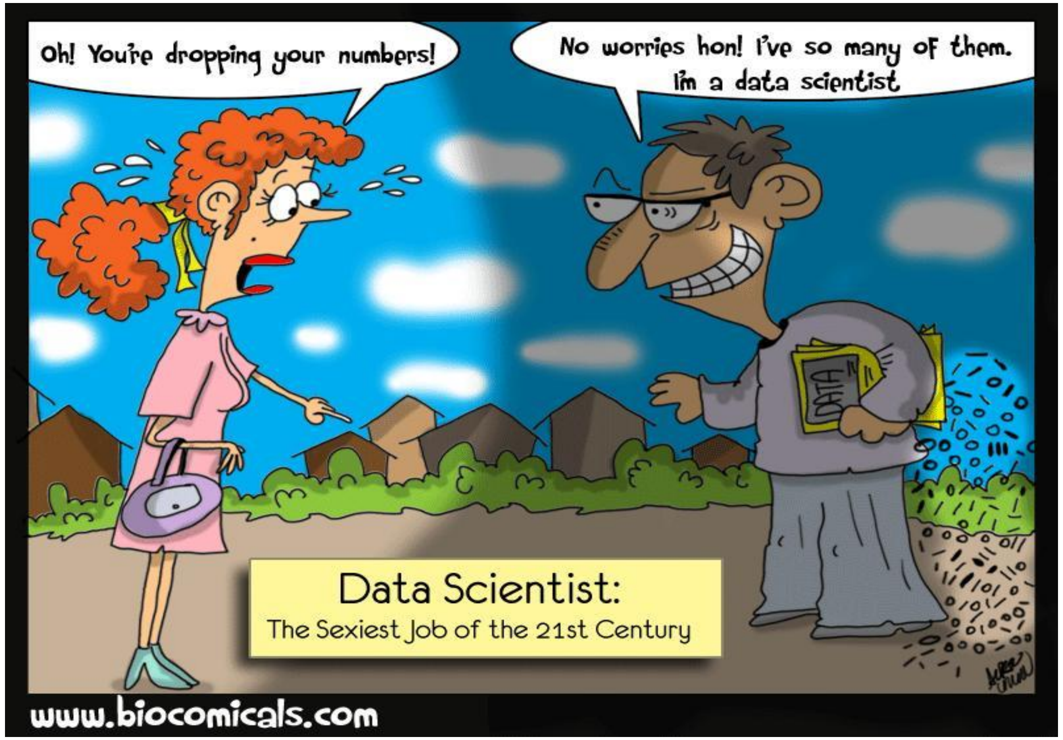
- Data Scientists are analytical experts who are responsible for finding insights and patterns in the data.
- Responsible for handling raw data, analyzing the data, implementing various statistical procedures, visualizing the data and generating insights from it.
- Must have knowledge of various tools like Hadoop, R, Python,.
- A Data Architect is responsible for implementing the blueprints of a company's data platform in terms of delineates various models, policies, rules that govern the storage of data as well as its use in the organizations.
- Tools used by a Data Architect are XML, Hive, SQL, Spark and Pig.

- A Data Engineer is responsible for building big data pipelines and models for the data scientists to work on.
- Must be well versed with both structured as well as unstructured data.
- Responsible for building data models, maintaining, managing and testing it.
- Responsible for modeling large scale processing systems using tools like SQL, Hive, Pig, Python, Java etc.

Job Roles Data Science



- A Statistician is responsible for implementing A/B testing, harvesting data, describing data, developing inferential statistical tools and performing hypothesis testing.
- Tools used by statisticians are R, SAS, SPSS, Matlab, Python, Stata, SQL etc.
- A Machine Learning Engineer is responsible for tailoring machine learning models for performing classification and regression tasks.
- It is an advanced field and people are required to possess analytical aptitude skills to develop machine learning algorithms.
- Some of the popular tools used by the machine learning engineers are TensorFlow, Keras, PyTorch, scikit-learn,
- Decision Scientists help the company to make business decisions with the help of tools like Artificial Intelligence and Machine Learning.
- It is a part of data science that extends to design thinking and behavioral sciences to better understand the clients.


Oh! You're dropping your numbers!
No worries hon! I've so many of them. I'm a data scientist
DATA
Data Scientist:
The Sexiest Job of the 21st Century
www.biocomicals.com