MACHINE LEARNING Introduction

Trainer : Sujata Mohite
Email: sujata.mohite@sunbeaminfo.com

Course Contents

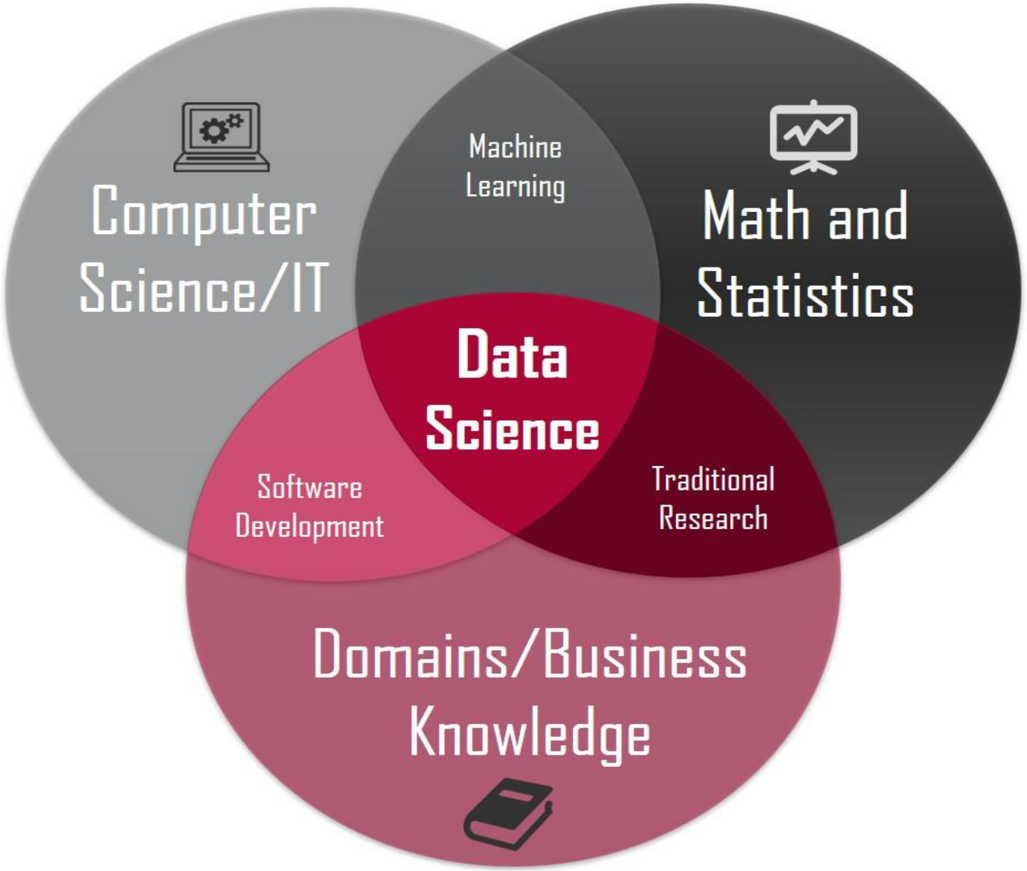
- Introduction to Machine Learning
- Preparing the Data ← preprocessing the data
- Feature Engineering, Model Selection & Tuning Training Models

  hyper parameters
- Classification ✓
- Support Vector Machine
- Decision Tree ←
- Ensemble Learning

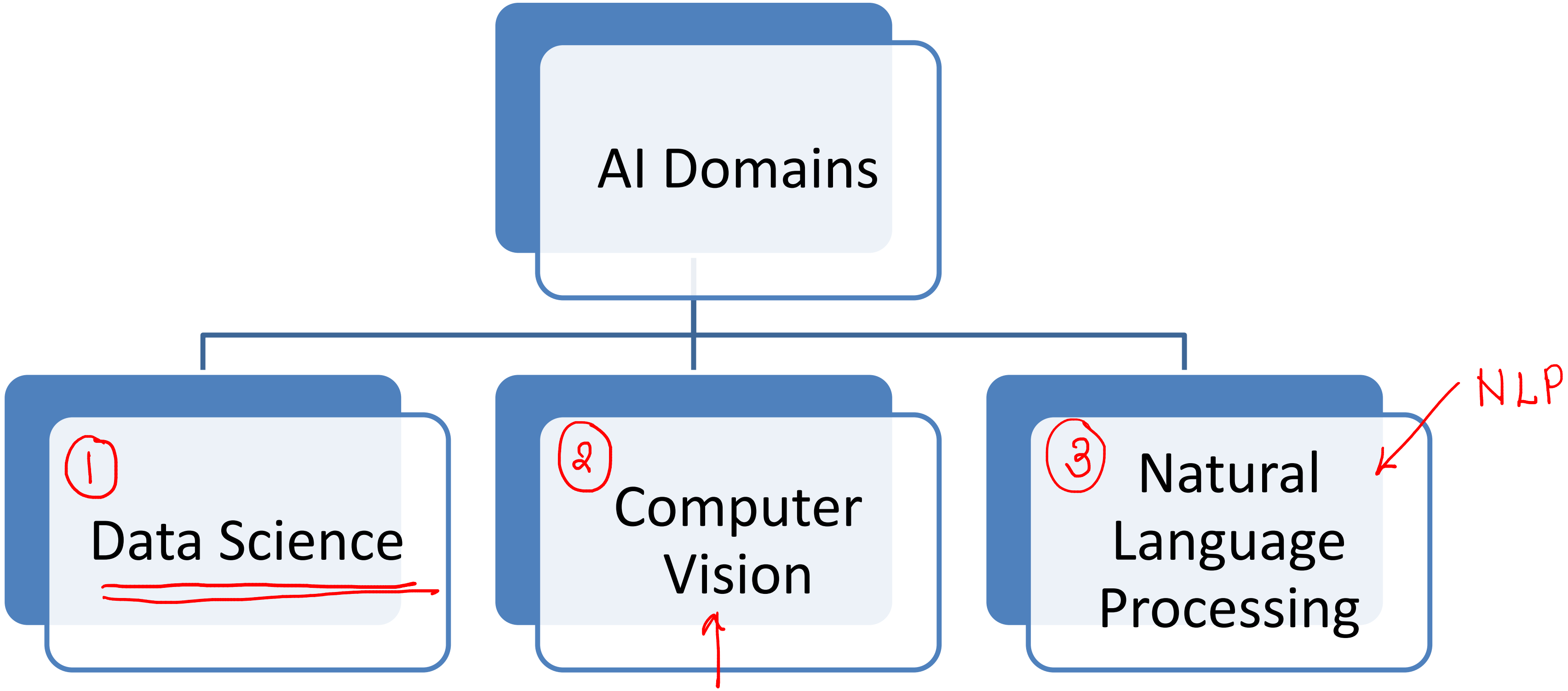
  Bagging Boosting Stacking

  Bagging → Random forest

Data Science



Main Domains of AI Technology



AI Domains

1 Data Science

2 Computer Vision

3 Natural Language Processing

NLP

1. What is Data Science ?

- Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data.

image, video

algo

- Data science is related to data mining, machine learning and big data

- It processes the numeric /alphanumeric data with data systems to extract information to make decision or predict the outcome for the problem statement.
- Example:

Petrol / Gold / housing

1. Recommendations

2. Price Predicting

3. Weather Forecasting

Netflix
spotify
amazon

Data Science Techniques

Data Collection

- Cannot be analyzed straight away
- It is untouched data that is accumulated and stored on the server
- Also known as raw facts or primary data
- Can be collected by various techniques like
    - ✓ Survey
    - ✓ Automated tools

Data Preprocessing

- This process tries to fix the problem that has occurred while data gathering
- Before processing with data analysis, it is important to remove the wrong data
- Techniques
  - Class Labeling
    - Labeling the data to the correct data types
    - e.g. numeric and categorical
  - Data cleansing
    - Deal with inconsistent data
    - Also known as data cleaning or data scrubbing
  - Dealing with missing values
  - Data balancing
  - Data shuffling
    - Prevents unwanted patterns
    - Improves predictive performance

Analyzing Data

- Once the data is cleaned and formatted, it can be analyzed for various reasons
- It explains past performance
- It can answer simple questions like
  - What happened ?
  - When did it happen?
- Or it can answer complex questions like
  - How did marketing team performed last quarter in terms of revenue
  - How does that compare to the performance in the same quarter last year

- Frequently used terms
- Metric
  - used to gauge the business performance or progress
  - metric = measure + business meaning
- Key Performance Indicators (KPI):
  - Key: related to the business goals
  - Performance: how successfully you have

performed within a specified timeframe

  - Indicators: shows values indicates somethings about the business
- dashboards

Predictive Analytics - Traditional

- After the analysis is over, the next logical step is analytics
- It can be performed traditional statistical modelling like
  - Regression
    - A model used for quantifying causal relationships among the different variables included in your analysis
    - Mostly used for predicting future values
  - Clustering
    - Creating different clusters (groups) by understanding data
  - Factor Analysis
  - Time Series analysis

Predictive Analytics – Machine Learning

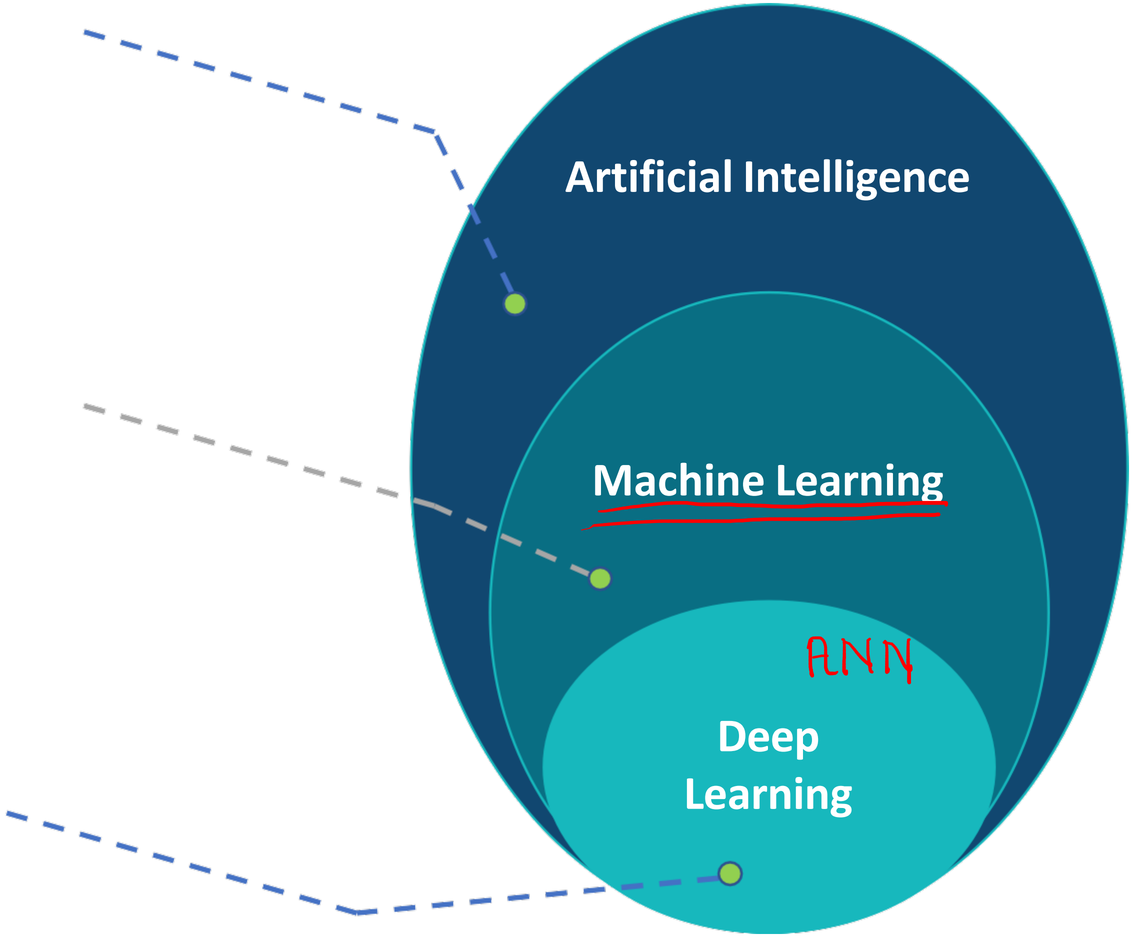
- Utilizes artificial intelligence to behavior in unprecedented ways
- There are different techniques
  - ① ✓ Supervised Learning
  - ② Unsupervised Learning
  - ③ Reinforcement Learning

AI, ML, DL

Artificial Intelligence:
- A technique which enables machine to mimic human behavior

Machine Learning:
- Subset of AI which uses statistical methods to enable machines to improve the experience

Deep Learning:
- Subset of ML which makes the computation of multi-layer neural network feasible



Artificial Intelligence

What is AI?

- Artificial Intelligence is an attempt to make a computer, a robot, or other piece of technology ‘think’ and process data in the same way as we humans do.

- AI is a branch of science which deals with helping machines finds solutions to complex problems in a more human-like fashion.

- AI therefore has to study how the human brain ‘thinks’, learns, and makes decisions when it tries to solve problems or execute a task.

- The aim of AI is to improve technology by adding functionality related to the human acts of reasoning, learning, and problem-solving.

- Example : Home Automation Systems, Cortana is example of a voice controlled intelligent system

Why are we talking about it now ?

More Computational Power

More Data

Better algorithms

Broad investment

AI applications

- ✓ Google's search engine
- JPMorgan Chase's Contract Intelligence (COiN) platform uses AI, machine learning and image
  recognition software to analyse legal documents
- IBM Watson: Healthcare organizations use IBM AI (Watson) technology for medial diagnosis
- Google's AI Eye Doctor can examine retina scans and identify a condition called as diabetic retinopathy which can cause blindness
- Facebook uses ML and DL to detect facial features and tag your friends
- Twitter uses AI to identify hate speech and terroristic language in the tweets
- Smart Assistants: Siri, Google Assistant, Alexa, Cortana (NLP)
- ✓ Tesla automated cars
- ✓ Netflix uses AI for movie recommendations
- ✓ Spam filtering

Machine Learning

What is machine learning ?

Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed.
Machine Learning is the science (and art) of programming computers so they can learn from data.

Machine Learning (ML): Involves training algorithms to recognize patterns in data and make decisions or predictions without being explicitly programmed.

Deep Learning(DL) : A subset of machine learning that uses neural networks with many layers to analyze large amounts of data for highly accurate predictions.

Examples of Applications

- Analyzing images of products on a production line to automatically classify them
  - This is image classification, typically performed using convolutional neural networks

CNN

- Detecting tumors in brain scans (MRI)
  - This is semantic segmentation, where each pixel in the image is classified (typically use CNNs)

cricket education

political sports

- Automatically classifying news articles ←
  - This is natural language processing (NLP), and more specifically text classification
- Automatically flagging offensive comments on discussion forums
  - This is also text classification, using the same NLP tools
- Forecasting your company's revenue next year, based on many performance metrics
  - This is a regression task (i.e., predicting values) that may be tackled using any regression model

Examples of Applications

- Making your app react to voice commands
    - This is speech recognition, which requires processing audio samples: since they are long and complex sequences, they are typically processed using RNNs, CNNs, or Transformers
- Detecting credit card fraud
    - This is anomaly detection example
- Segmenting clients based on their purchases so that you can design a different marketing strategy for each segment
    - This is clustering example
- Representing a complex, high-dimensional dataset in a clear and insightful diagram
    - This is data visualization, often involving dimensionality reduction techniques

Examples of Applications

- Recommending a product that a client may be interested in, based on past purchases
  - This is a recommender system

- Building an intelligent bot for a game
  - This is often tackled using Reinforcement Learning

Take an Example

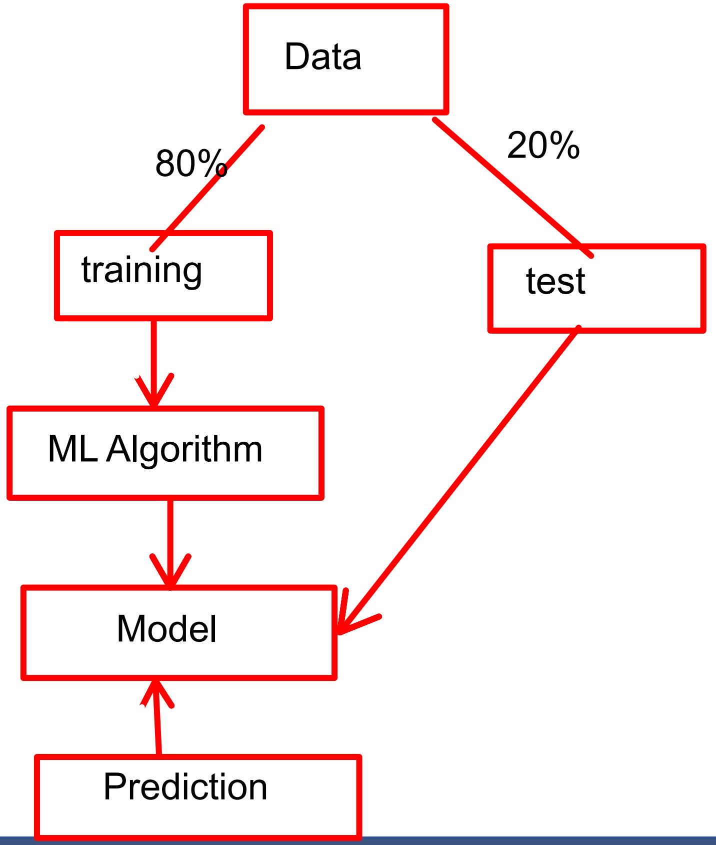
x	Y
1	1
2	4
3	9
4	16
5	25
8	??

$$y = x^2$$

Formula / Model / Function

$= 8^{2}$

$= 64$



independent/
input

output/dependent

x	y
1	1
2	4
3	9
4	16
5	25
6	36

Dataset

y = x^2

formula/model/function

Input

x = 3
y = x^2
= 3^2
= 9

x = 7
y = x^2
= 7^2
= 49

prediction

x	y
1	1
2	4
3	9
4	16
5	25
6	36

training dataset

x	y
8	64
9	81

test

for x =2
seen data

y = 4

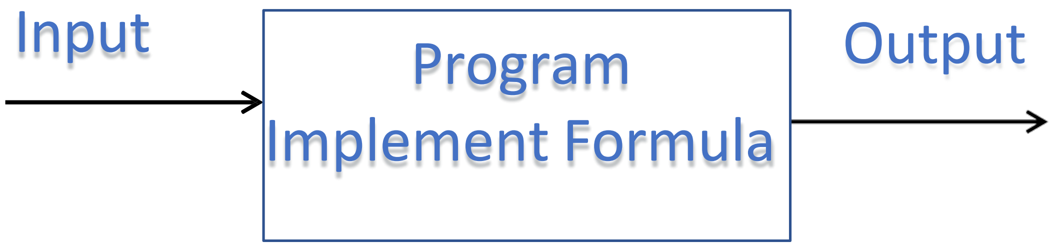
for x = 8
unseen
data
y = 64

prediction

Model = good
Accuracy = Good

ML Program

Non ML Programming

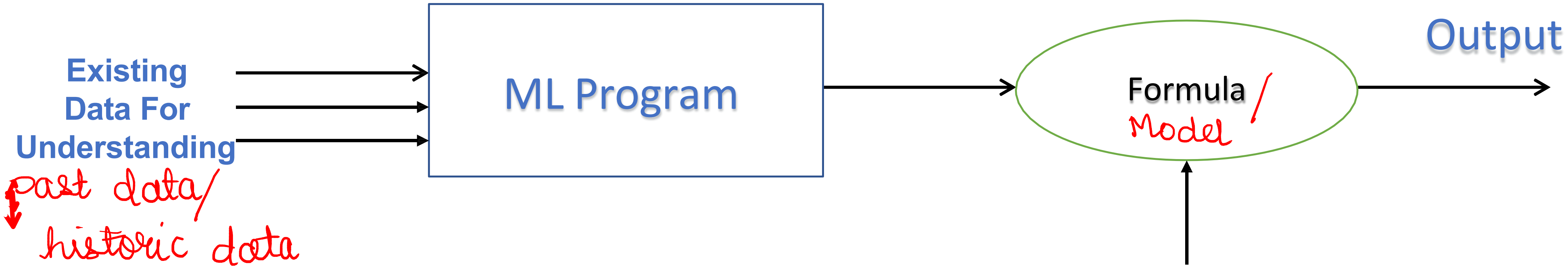


Input

Program
Implement Formula

Output

ML Programming



Existing
Data For
Understanding

past data/
historic data

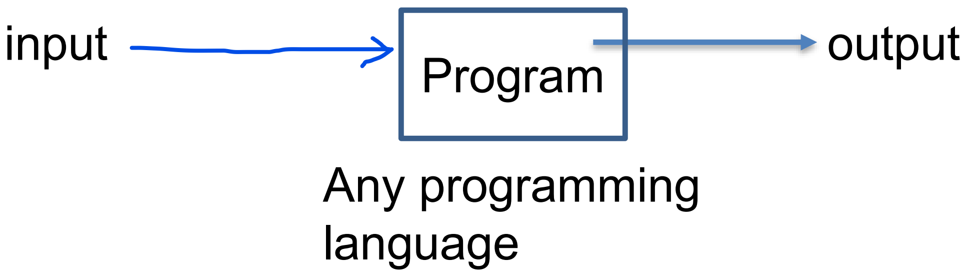
ML Program

Formula / Model

Output

ML Program

Traditional Appraoch	ML Approach
1. Algorithm[Formula] is already known	1. Formula is unknown
2. Language is used to implement the algorithm	2. Language is used to find the model
3. No past data required	3. A past data is mandatory

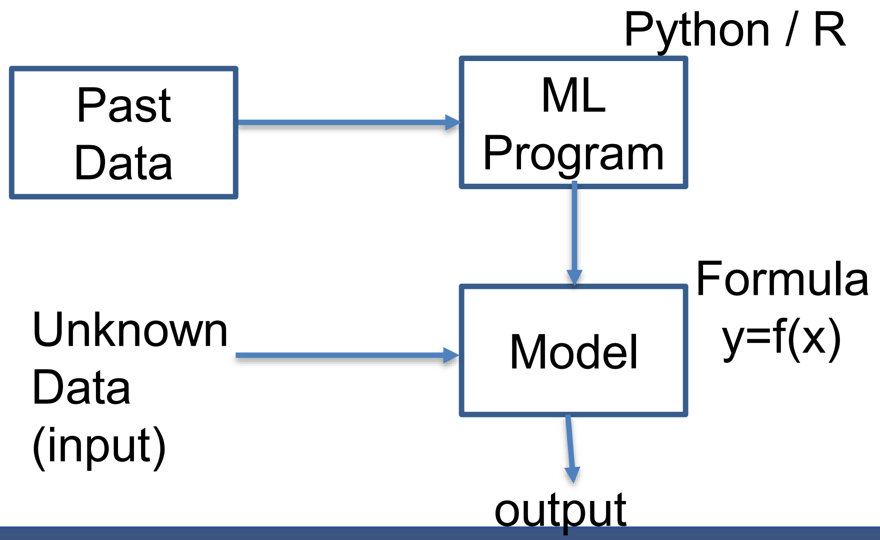


input

Program

output

Any programming language



Python / R

Past Data

ML Program

Formula y=f(x)

Unknown Data (input)

Model

output

Where to use Machine Learning ?

- Problems for which existing solutions require a lot of fine-tuning or long lists of rules:
  - one Machine Learning algorithm can often simplify code and perform better than the traditional approach

- Complex problems for which using a traditional approach yields no good solution:
  - the best Machine Learning techniques can perhaps find a solution

- Fluctuating environments:
  - a Machine Learning system can adapt to new data

- Getting insights about complex problems and large amounts of data

Types

Types of Machine Learning

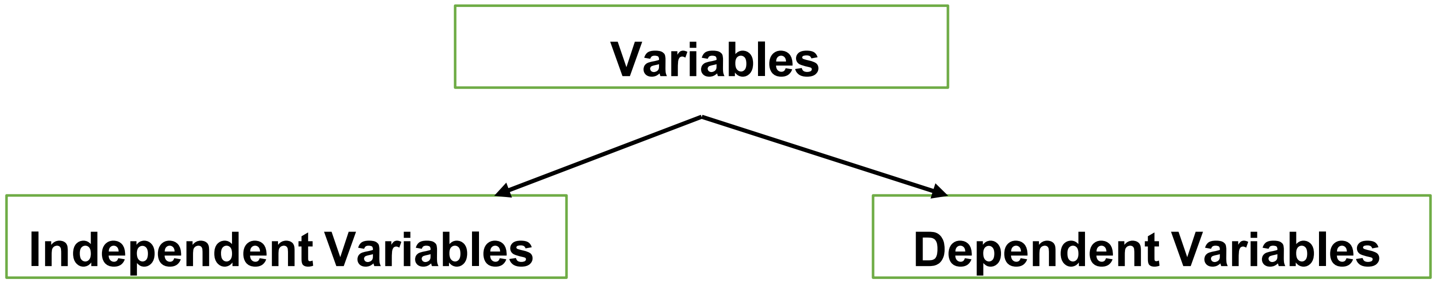
- There are so many different types of Machine Learning systems that it is useful to classify them in broad categories, based on the following criteria
  - Whether or not they are trained with human supervision
    - supervised, unsupervised, and Reinforcement Learning
  - Whether or not they can learn incrementally on the fly ← at runtime
    - online versus batch learning
  - Whether they work by simply comparing new data points to known data points, or instead by detecting patterns in the training data and building a predictive model, much like scientists do
    - instance-based versus model-based learning

Variable / Column / Feature

No.	Location	Area	Price
1	Pune 2BHK	Abc	70
2	Mumbai 2BHK	Pqr	90

Known Data
Or
Labelled Data

No. column not required
Location and Area are Independent variables
Price is Dependent Variable



Variables

Independent Variables

Dependent Variables

One or many
Input variables eg. Location and area

One and Only one Output
Eg. Price

Supervised

prediction

Unsupervised

No prediction (EDA)

Reinforcement Learning

self learning
experience

Supervised Learning

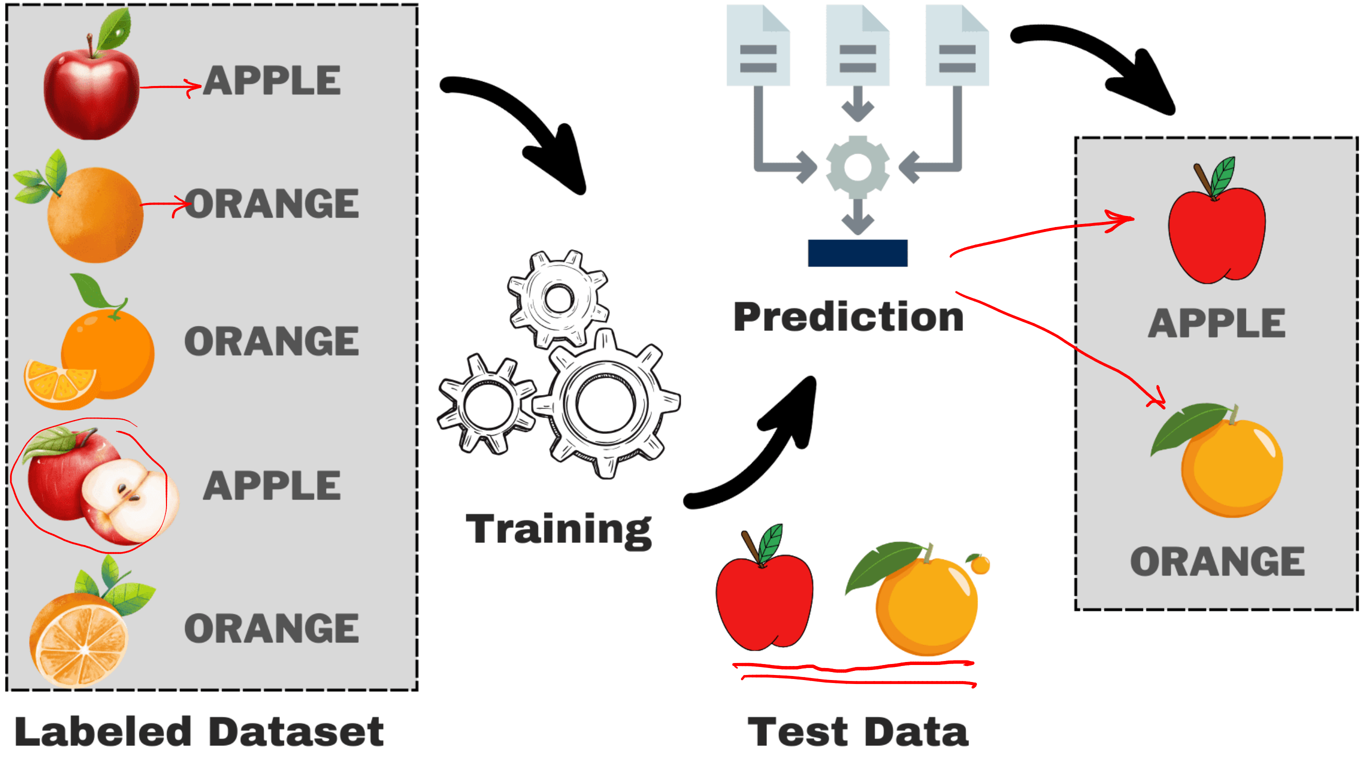
- The majority of practical machine learning uses supervised learning
- Supervised learning is where you have **input variables (x)** and an **output variable (Y)** and you use an algorithm to learn the mapping function from the input to the output

Y = f(X) → model / formula ✓ ,

- The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.
- It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process
- We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher
- Learning stops when the algorithm achieves an acceptable level of performance (measured in terms of **accuracy**)

Supervised Learning

- Output variable is already known for each input variable
- Algorithm learns to map input and output
- Model learns to associate features in the images with these predefine categories.



Supervised Learning – Problems

- ✓ **Regression**
  - Related to predicting future values
  - E.g.
    - Population growth prediction
    - Expecting life expectancy
    - Market forecasting/prediction
    - Advertising Popularity prediction
    - Stock prediction
  - Algorithms
    - ✓ Linear and multi-linear regression
    - ✓ Logistic regression
    - Naïve Bayes
    - ✓ Support Vector Machine

Supervised Learning – Problems

- **Classification** ← decision based prediction
  - Related to classify the records
  - not imp. imp

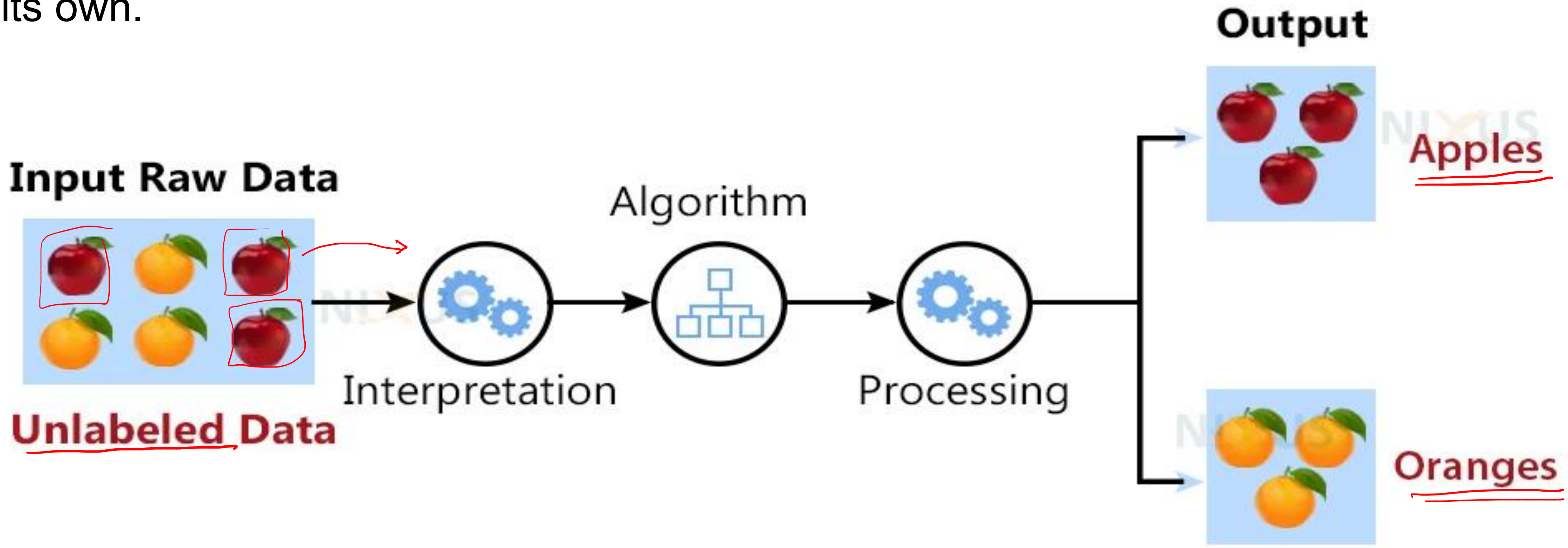
    Based on class / labels (eg. Email : Spam / Ham , Gender : Male / Female , Loan : Yes / No)
  - E.g.
    - Find whether an email received is a spam or ham
    - Identify customer segments
    - Find if a bank loan is granted
    - Identify if a kid will pass or fail in an examination
  - Algorithms
    - ✓ Logistic Regression
    - ✓ Decision Tree
    - ✓ Random Forest
    - ✓ Support Vector Machine
    - ✓ K-nearest neighbor

Unsupervised Learning

- Unsupervised learning is where you only have input data (X) and no corresponding output variables
- The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data
- These are called unsupervised learning because unlike supervised learning above there is no correct answers and there is no teacher
- Algorithms are left to their own devises to discover and present the interesting **structure** in the data
- Structure in the form of GROUPS / CLUSTERS / ASSOCIATION
- Mostly used for EDA (Exploratory Data Analysis)
- Is to understand the data's structure, find patterns, identify anomalies or outliers, and check assumptions before formal modeling or hypothesis testing. EDA is a crucial first step in data analysis and machine learning projects to gain initial insights and ensure data quality

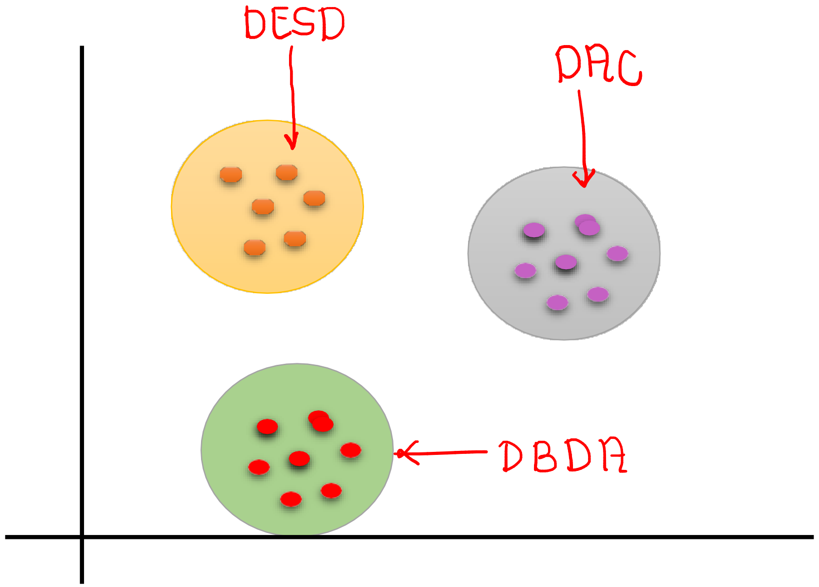
Unsupervised Learning

- Analyzes and clusters unlabeled datasets.
- These algorithms find hidden patterns and data without any human intervention.
- The training model has only input parameter values and discovers the groups or patterns on its own.



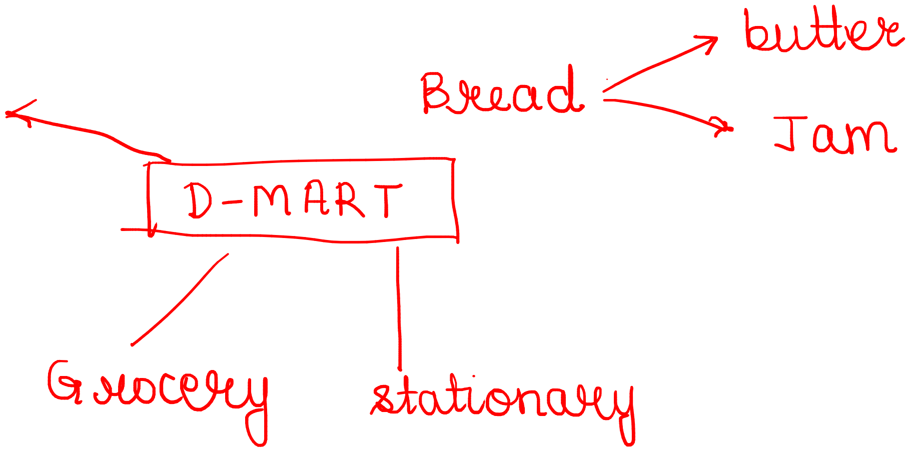
Unsupervised Learning - Problems

- **Clustering**
  - discover the inherent groupings in the data, such as grouping customers by purchasing behaviour
  - E.g.
    - Batsman vs bowler
    - Customer spending more money vs less money
  - Algorithms
    - ✓ K-means clustering
    - ✓ Hierarchical clustering



Unsupervised Learning - Problems

- **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
  - E.g.
    - Market basket analysis



- Algorithms
  - Apriori
  - Eclat

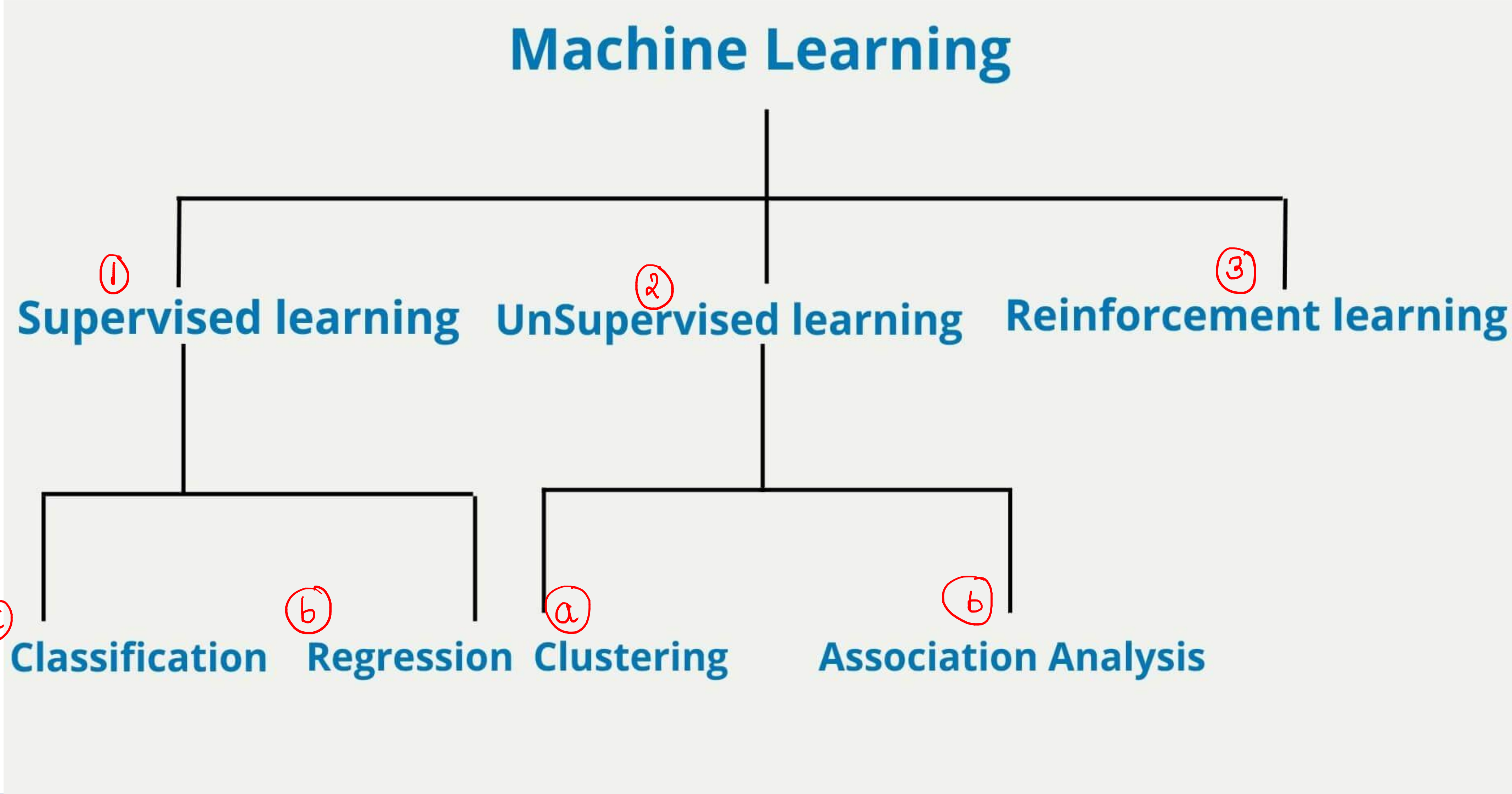
Stationary
Cloth store
Electronic gadgets

Reinforcement Learning

feedback / review

- It is about taking suitable action to maximize reward in a particular situation
- It is employed by various software and machines to find the best possible behaviour or path it should take in a specific situation
- Reinforcement learning differs from the supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task
- In the absence of training dataset, it is bound to learn from its experience.
- Examples
  - Resources management in computer clusters
  - ✓ Traffic Light Control
  - Robotics
  - Web system configuration
  - ✓ Chemistry
- Algorithms
  - ✓ Q-Learning
  - ✓ Deep Q-Learning

Machine Learning



Machine Learning

1 Supervised learning

- a Classification
- b Regression

2 UnSupervised learning

- a Clustering
- b Association Analysis

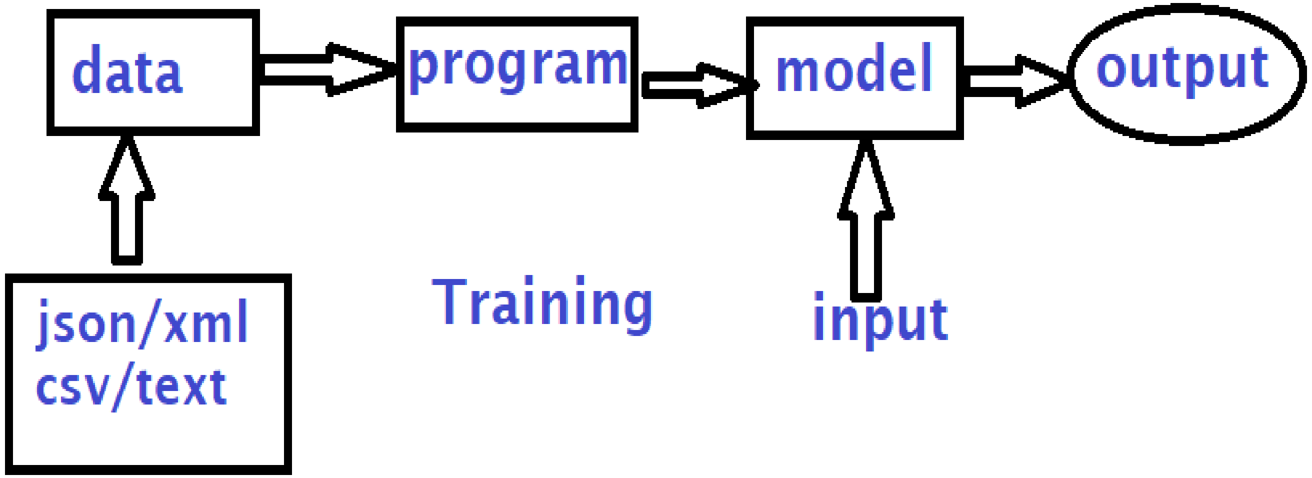
3 Reinforcement learning

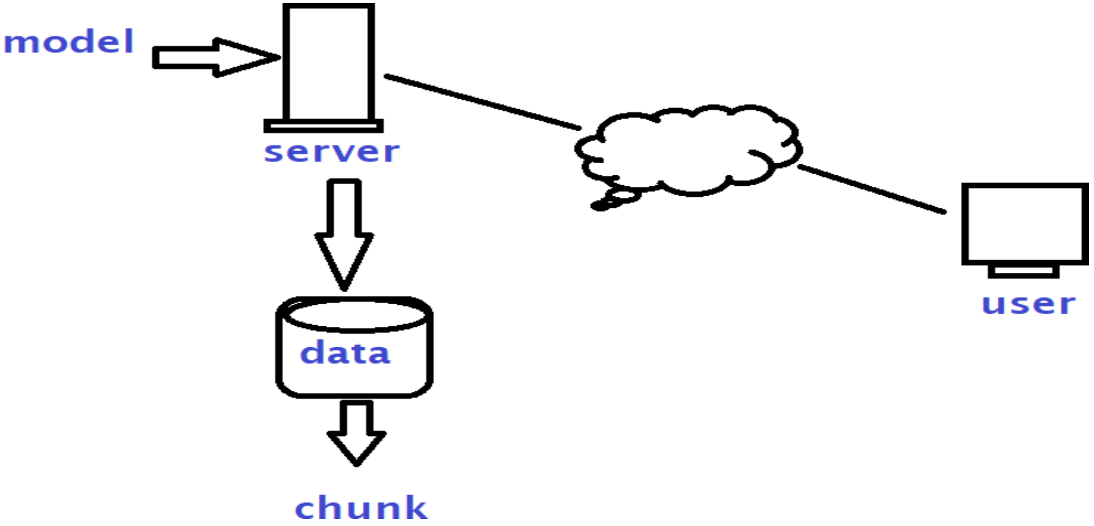
Batch/ Offline Learning

Online Learning

Batch Learning

- In batch learning, the system is incapable of learning incrementally
- it must be trained using all the available data
- This will generally take a lot of time and computing resources, so it is typically done offline
- First the system is trained, and then it is launched into production and runs without learning anymore, it just applies what it has learned
- This is also called as offline learning





Batch Learning - cons

- If you want a batch learning system to know about new data, you need to train a new version of the system from scratch on the full dataset, then stop the old system and replace it with the new one
    - The whole process of training, evaluating, and launching a Machine Learning system can be automated easily

- Training using the full set of data can take many hours
    - Typically train a new system only every 24 hours or even just weekly

- Training on the full set of data requires a lot of computing resources (CPU, memory space, disk space, disk I/O, network I/O)

Online Learning

- In online learning, you train the system incrementally by feeding it data instances sequentially, either individually or in small groups called mini-batches
- Each learning step is fast and cheap, so the system can learn about new data on the fly, as it arrives
- Online learning is great for systems that receive data as a continuous flow (e.g., stock prices) and need to adapt to change rapidly or autonomously
- It is also a good option if you have limited computing resources
  - once an online learning system has learned about new data instances, it does not need them anymore, so you can discard them
  - This can save a huge amount of space
- Online learning algorithms can also be used to train systems on huge datasets that cannot fit in one machine's main memory (this is called **out-of-core learning**)

Online Learning

- One important parameter of online learning systems is how fast they should adapt to changing data: this is called the **learning rate**
- If you set a high learning rate, then your system will rapidly adapt to new data, but it will also tend to quickly forget the old data
  - you don’t want a spam filter to flag only the latest kinds of spam it was shown
- if you set a low learning rate, the system will have more inertia
  - that is, it will learn more slowly, but it will also be less sensitive to noise in the new data or to sequences of nonrepresentative data points (**outliers**)

Instance Based

Model Based

Instance Based

- The system learns the examples by heart
- Classification Based on record/ row / example
- Then generalizes to new cases by using a similarity measure to compare them to the learned examples (or a subset of them)

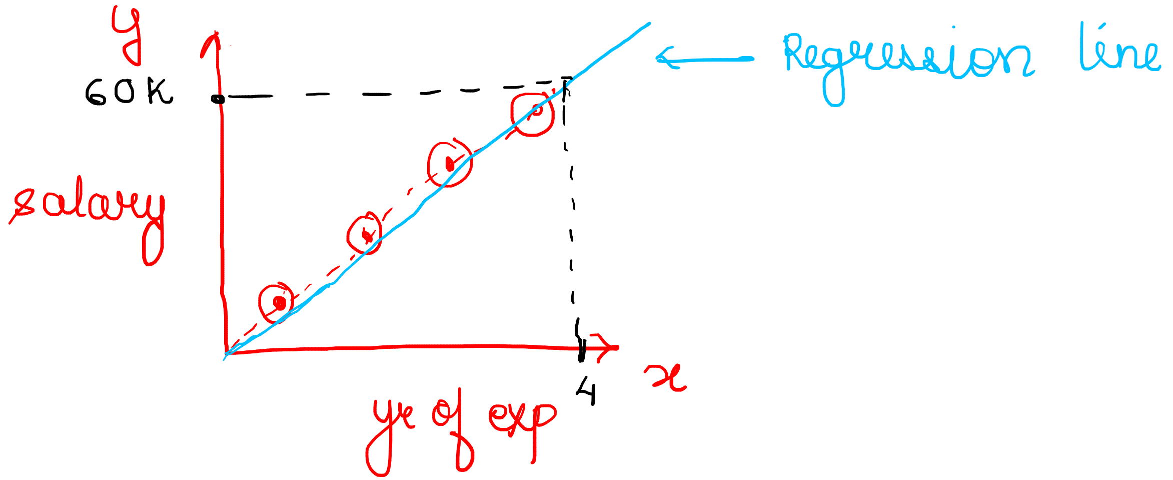
assumption

- It is called instance-based because it builds the hypotheses from the training instances
- It is also known as **memory-based learning** or **lazy-learning**
- **Advantages:**
  - Instead of estimating for the entire instance set, local approximations can be made to the target function
  - This algorithm can adapt to new data easily, one which is collected as we go
- **Disadvantages:**
  - Classification costs are high
  - Large amount of memory required to store the data, and each query involves starting the identification of a local model from scratch
- E.g :- K Nearest Neighbor (KNN)

Model Based

- Train model from training data to estimate model parameters i.e. discover patterns
- Store the built model in suitable format
- Generalize the rules of model
- Predict the unseen instance (data) using the model
- It requires a known model form
- It takes less memory compared to the instance based learning
- E.g.
  - ✓ Linear Regression

yrofexp	salary
1	20
1.5	30
¦	¦



End to End Process

Steps

- Look at the big picture ← means the actual data
- Get the data → collect/organize data ← database, csv file, image file
- Discover and visualize the data to gain insights (**Exploratory Data Analysis** (**EDA**))

head(), tail(), info(), describe() & visualize the data

convert categorical to numeric data

- Prepare the data for Machine Learning algorithms → data cleansing
- Select a model and train it → by trial and error method
- Fine-tune your model(hyper parameter tuning → optimizing model)

Grid Search( )

- Present your solution ← accuracy
- Launch, monitor, and maintain your system

Look at the Big Picture

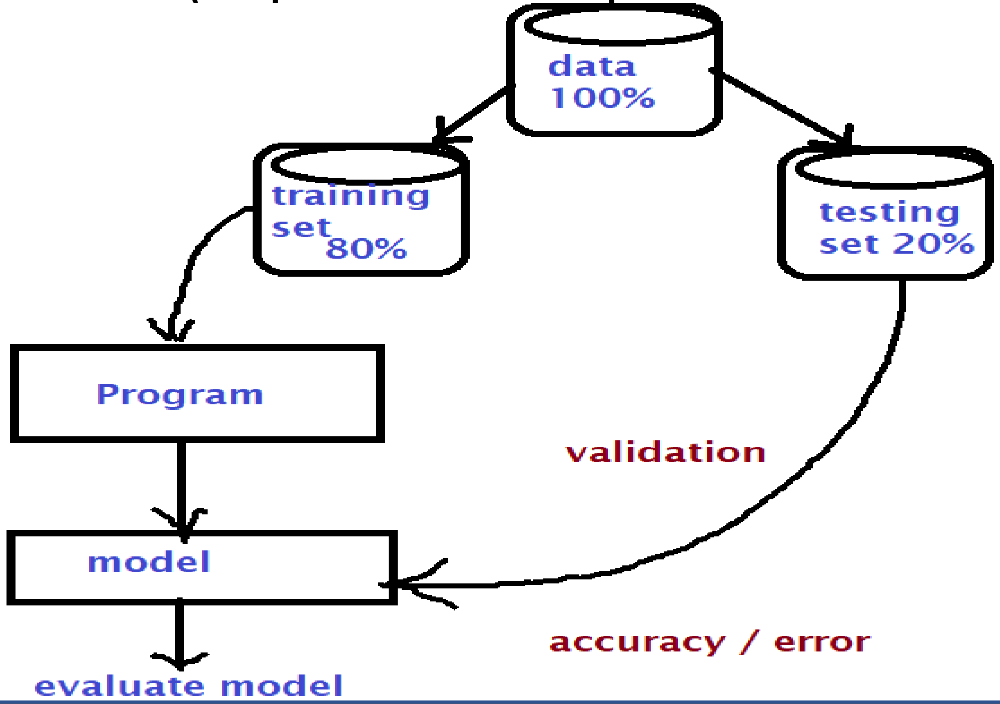
- Frame the Problem(Domain Knowledge)
  - The first question to ask your boss is what exactly the business objective is Building a model is probably not the end goal
  - How does the company expect to use and benefit from this model?
  - Knowing the objective is important because it will determine
    - how you frame the problem
    - which algorithm you will select
    - which performance measure you will use to evaluate your model
    - how much effort you will spend tweaking it

evaluation metrics

- Select a Performance Measure
  - Your next step is to select a performance measure
  - A typical performance measure for **regression** problems is the **Root Mean Square Error (RMSE) and MAE(Mean Absolute Error)**
  - It gives an idea of how much error the system typically makes in its predictions, with a higher weight for large errors

Get the data

- Decide the data source (file / database / online /api)
- Download the data and make it available for the further learning
- Take a Quick Look at the Data Structure
  - Understand the data set (numeric / textual / categorical etc) and understand its features /columns /variables
  - Evaluate the features and decide which one(s) are needed(dependent/independent→ correlation analysis)
- **Create a Test Set**
  - Keep some records aside for testing and validation



Discover and Visualize the Data to Gain Insights

- Visualize the data
  - Use libraries like matplotlib or seaborn
  - Understand the pattern and relationship
- Look for correlation
- Experiment with attribute combinations

Prepare the Data for Machine Learning Algorithms

- Data Cleaning
  - Process of cleaning the data set to prepare it for ML algorithm
  - Steps
    - Check for the missing data(NA values)
    - Check for wrong data types
    - Add features if needed
    - Remove unwanted features
- Feature Scaling
  - ML algorithms don’t perform well when the input numerical attributes have very different scale
  - Scale the features to bring all of them to a single scale
- Handle categorical / text data
  - Use transformers to convert categorical to numerical (eg. Label encoding / ordinal encoding/ one-hot encoding, etc)

Select and Train a Model

- Training the model using train data set
  - Create a model using selected algorithm
  - Save the model for future use
- Evaluation the model
  - Evaluate the model to see if there is any chance to improve the accuracy
  - Techniques
    - Cross Validation

Fine-Tune Your Model

- **Grid Search**
  - One option would be to fiddle with the hyperparameters manually, until you find a great combination of hyper parameter values(configuration of algorithms)
  - This would be very tedious work, and you may not have time to explore many combinations
  - You can also automate this process using libraries like sci-kit
- **Randomized Search**
  - The grid search approach is fine when you are exploring relatively few combinations
  - But when the hyperparameter search space is large, it is often preferable to use randomized search
- **Ensemble Methods**

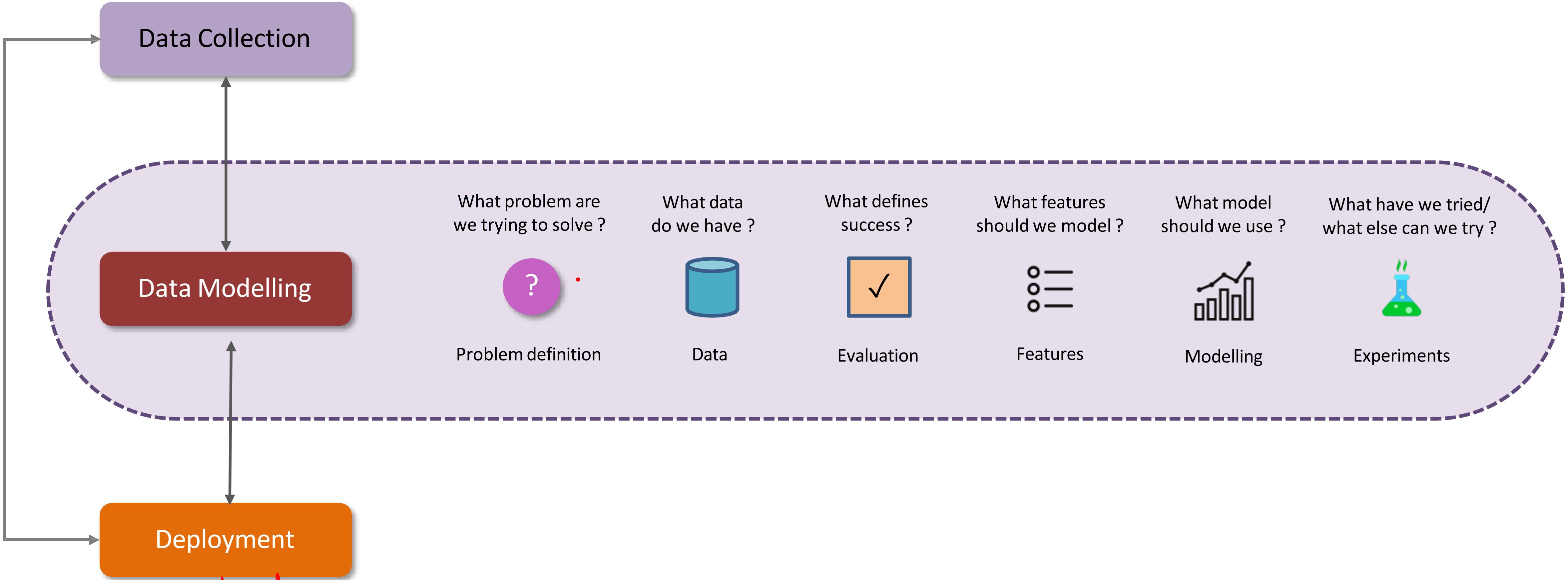
  bagging
  boosting
  stacking

  - Another way to fine-tune your system is to try to combine the models that perform best
  - The group (or “ensemble”) will often perform better than the best individual model, especially if the individual models make very different types of errors.
- Analyse the Best Models and Their Errors
- Evaluate Your System on the Test Set

Launch, Monitor, and Maintain Your System

- Deploy the application for the end users → production
- Monitor the application’s performance
- If the data keeps evolving, update your datasets and retrain your model regularly
- You should probably automate the whole process as much as possible
  - Collect fresh data regularly and label it
  - Write a script to train the model and fine-tune the hyperparameters automatically. This script could run automatically, for example every day or every week, depending on your needs
  - Write another script that will evaluate both the new model and the previous model on the updated test set, and deploy the model to production if the performance has not decreased (if it did, make sure you investigate why)

Summary



production

Salary_data.csv

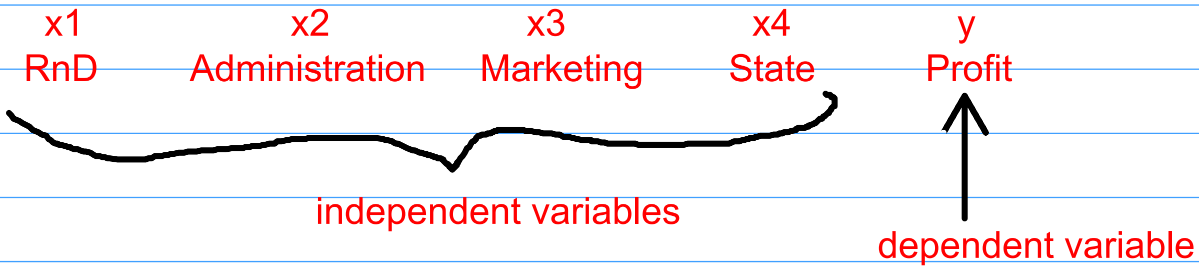
x YrsOfExp independent

y Salary dependent

Is salary dependent on yrsofexp ? Yes

Is yrsofexp dependent on salary ? No

50 startups dataset



x1 RnD

x2 Administration

x3 Marketing

x4 State

y Profit

independent variables

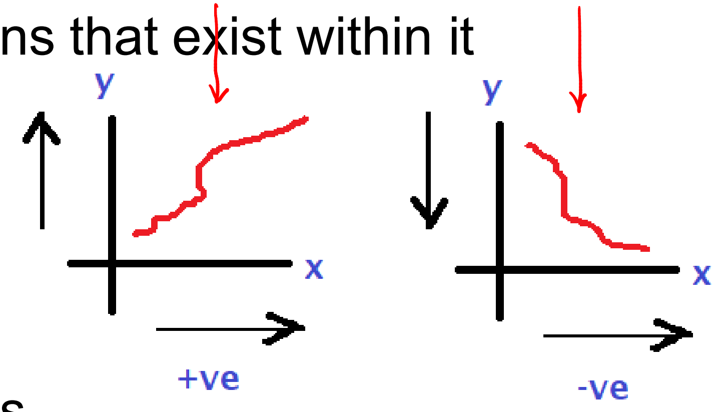
dependent variable

Measures of Correlation

Terminology

marks = [10, 11, 15, 17, 18]

- **Univariate**
  - This type of data consists of only one variable (eg. Temperature value)
  - It does not deal with causes or relationships
  - the main purpose of the analysis is to describe the data and find patterns that exist within it



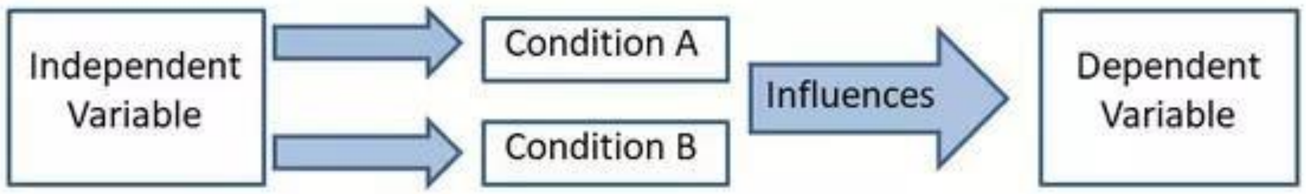
- **Bivariate**
  - This type of data involves two different variables
  - The analysis of this type of data deals with causes and relationships
  - the analysis is done to find out the relationship among the two variables

- **Multivariate**
  - When the data involves three or more variables
  - It is similar to bivariate but contains more than one independent variable

Terminology

- **Independent variable(s)**
  - A variable that represents a quantity that is being manipulated in an experiment
  - **Represents input**
  - Also known as regressors in a statistical context.
  - x is often the variable used to represent the independent variable in an equation

- **Dependent variable**
  - A quantity whose value *depends* on how the independent variable is manipulated
  - **Represents output**
  - y is often the variable used to represent the independent variable in an equation

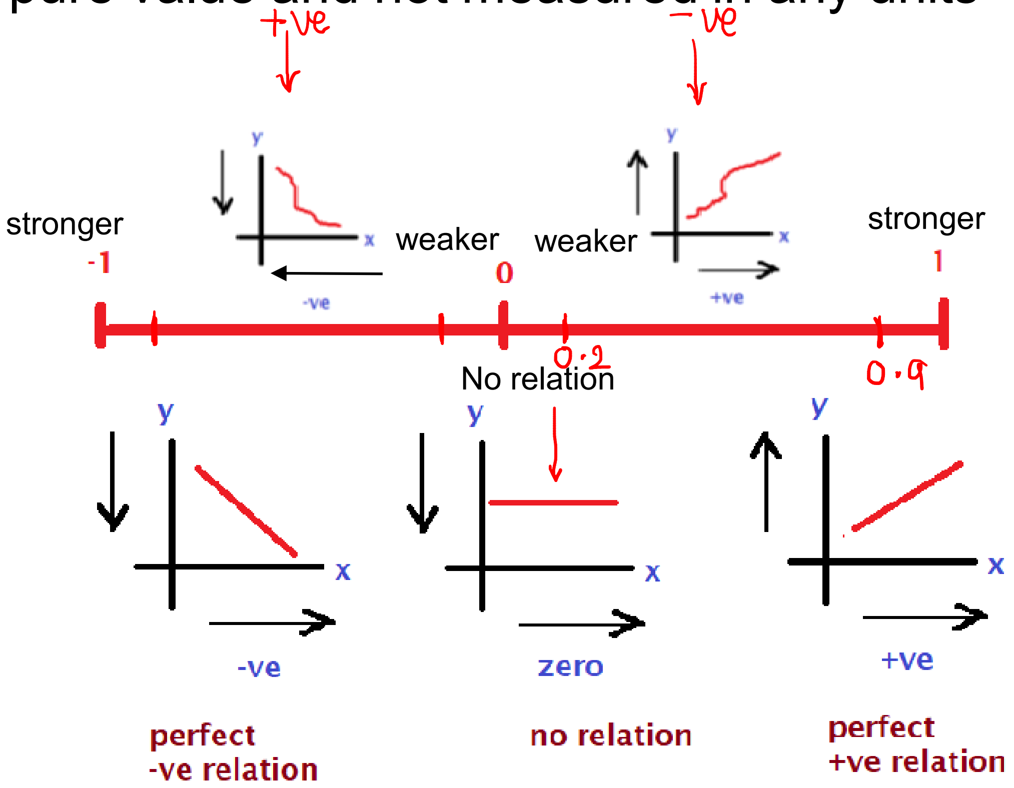


Correlation

+ve / −ve

- Measures the strength of the relationship between two or more variables
- Correlation is the scaled measure of covariance
- It is dimensionless: the correlation coefficient is always a pure value and not measured in any units

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$



- Where
  - ρ(X,Y) – the correlation between the variables X and Y
  - cov(X,Y) – the covariance between the variables X and Y
  - σ_X – the standard deviation of the X-variable
  - σ_Y – the standard deviation of the Y-variable

Correlation

- Even a high degree of correlation does not necessarily mean that the relationship of cause and effect exists between the variables.

- The explanation of significant degree of correlation may be one or both of the following-
  - The correlation may be due to pure chance, especially in a small sample-
  - Both the correlated variables may be influenced by one or more other variables
  - Both the correlated variables are mutually influencing each other.

Correlation coefficient

- Following are the ways to calculate correlation coefficient
  - Karl Pearson’s coefficient of correlation
  - Spearman’s Rank correlation
  - Scatter diagram
  - Coefficient of concurrent duration

- Correlation (r) → corr(x , y) = corr(y , x)
  - The coefficient of correlation lies between -1 and +1, symbolically -1 <= r <= 1
  - r = 1 (perfect correlation)
  - r = -1 (perfect negative correlation)
  - r > 0 (positive correlation)
  - r < 0 (negative correlation)
  - r = 0 (no correlation)

Covariance vs Correlation

- Both primarily assess the relationship between variables
- The closest analogy to the relationship between them is the relationship between the variance and standard deviation
- Covariance measures the total variation of two random variables from their expected values while correlation measures the strength of the relationship between variables
- Using covariance, we can only gauge the direction of the relationship while correlation is the scaled measure of covariance

Discrete Variable

- A discrete variable is a type of variable that can only take on specific, distinct values.
- These values are typically whole numbers or integers.
- Discrete variables often represent counts or categories.
- eg: Number of students in a classroom:
- It is a discrete variable because it can only take on whole
- number values (e.g., 25 students, 30 students).

- Outcomes of rolling a six-sided die:
- The output will be (1, 2, 3, 4, 5, or 6) which are discrete because
- they consist of distinct, separate categories.

- Number of books on a shelf:
- The number of books is discrete because it cannot take on
- fractional or continuous values
- (e.g., 5 books, 10 books, 15 books).

Continuous Variable

- Continuous variable is a type of variable that can take on any value within a given range.
- continuous variables can represent an infinite number of possible values, including fractional and decimal values.
- Continuous variables often represent measurements or quantities.
- Eg:
- Height: Height is a continuous variable because it can take on any value within a range (e.g., 150.5 cm, 162.3 cm, 175.9 cm).

- Weight: Weight is continuous because it can be measured with precision and can take on any value within a range
- (e.g., 55.3 kg, 68.7 kg, 72.1 kg).

- Time: Time can be measured with precision, and it can take on any value (e.g., 10:30:15.5 AM, 10:45:30.75 AM).

Categorical Data

- Categorical data refers to variables that belong to distinct categories such as labels, names or types.

label encoder/
one hot encoder

- Since most machine learning algorithms require numerical inputs, encoding categorical data to numerical data becomes important.
- Proper encoding ensures that models can interpret categorical variables effectively, leading to improved predictive accuracy and reduced bias.
- .

Types of Categorical Data

1. Nominal Data: Nominal data consists of categories without any inherent order or ranking. These are simple labels used to classify data.
Eg: 'Red', 'Blue', 'Green' (Car Color).
Encoding Options: One-Hot Encoding or Label Encoding, depending on the model's needs.

2. Ordinal Data: Ordinal data includes categories with a defined order or ranking, where the relationship between values is important.
Eg: 'Low', 'Medium', 'High' (Car Engine Power).
Encoding Options: Ordinal Encoding.

Using the right encoding techniques, we can effectively transform categorical data for machine learning models which improves their performance and predictive capabilities

X	Y
1	100
2	200
3	300
4	400
5	500
6	600
7	700
8	800
9	900
10	1000

Dataset

x_train	y_train
1	100
2	200
3	300
4	400
5	500
6	600
7	700
8	800

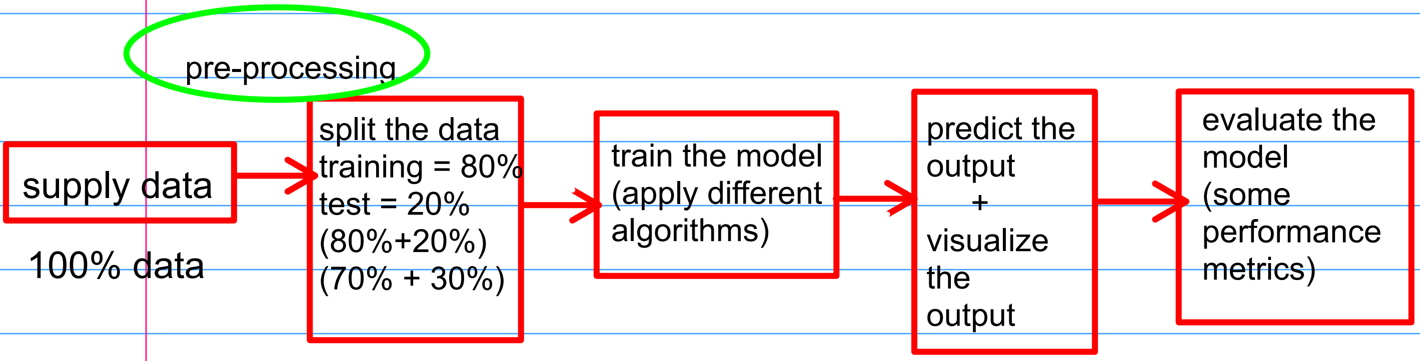
training = 80%

x_test	y_test
9	900
10	1000

test data = 20%

11 = ? new data

```
x_train,x_test,y_train,y_test  =  train_test_split (x,y,train_size=0.8)
```



pre-processing

supply data

100% data

split the data
training = 80%
test = 20%
(80%+20%)
(70% + 30%)

train the model
(apply different
algorithms)

predict the
output
+
visualize
the
output

evaluate the
model
(some
performance
metrics)

Thank You!!