Chapter 4 Machine Learning

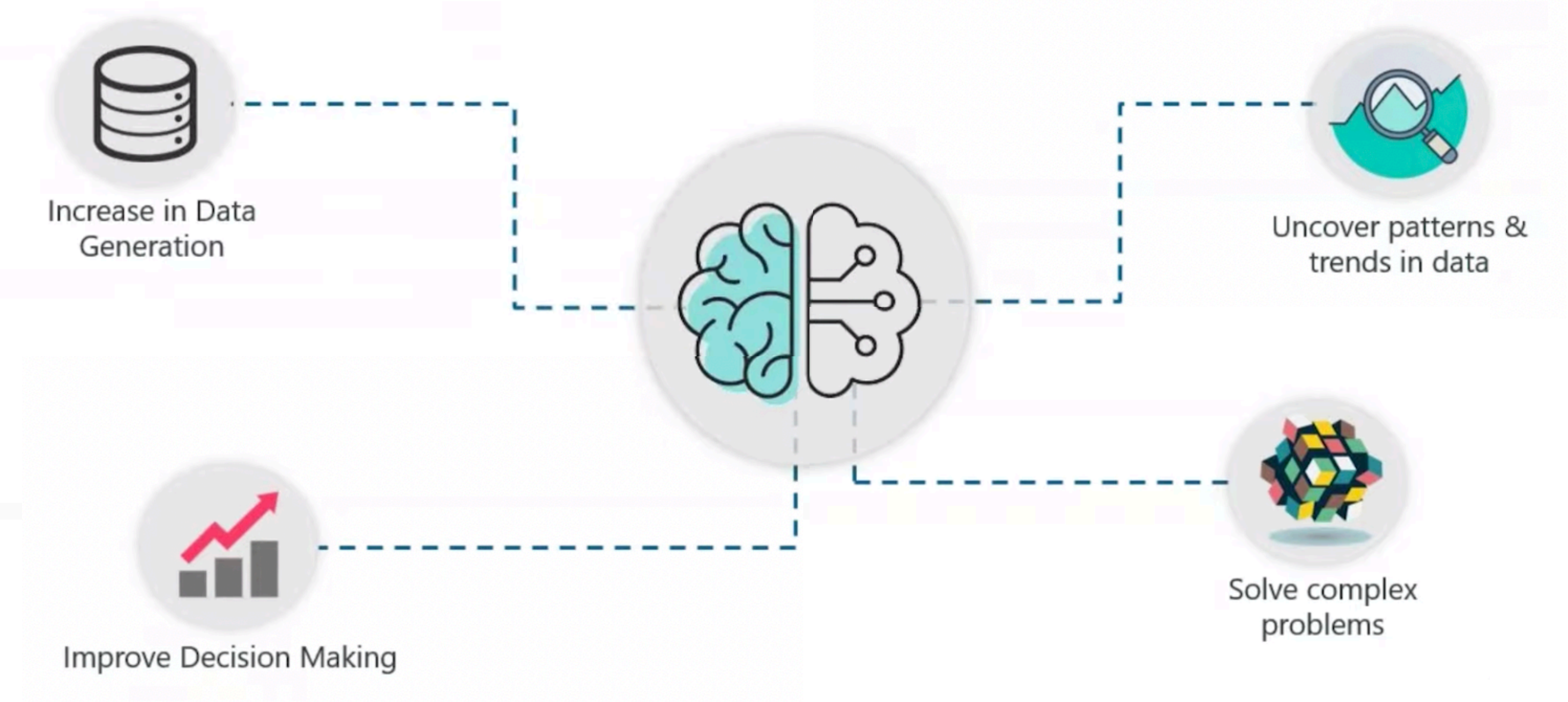
COMP 6721 Introduction of AI

Russell & Norvig – Section 18.1 & 18.2

Why Machine Learning?

Over 2.5 quintillion bytes of data are created every single day, and it is only going to grow from there. By 2020, it is estimated that 1.7MB of data will be created every second for every person on earth.

Why Machine Learning?



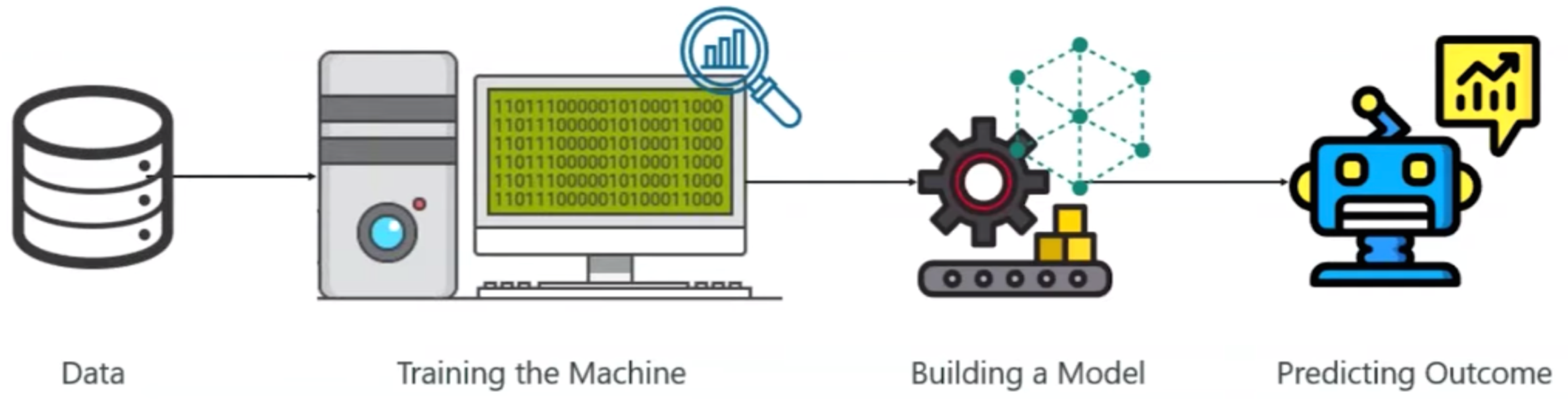
What is Machine Learning

- In 1959, Arthur Samuel first proposed the concept Machine Learning.

- Machine Learning is a subset of Artificial Intelligence which provides machines the ability to learn automatically & improve from experience without being explicitly programmed.

What is Machine Learning

- "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."

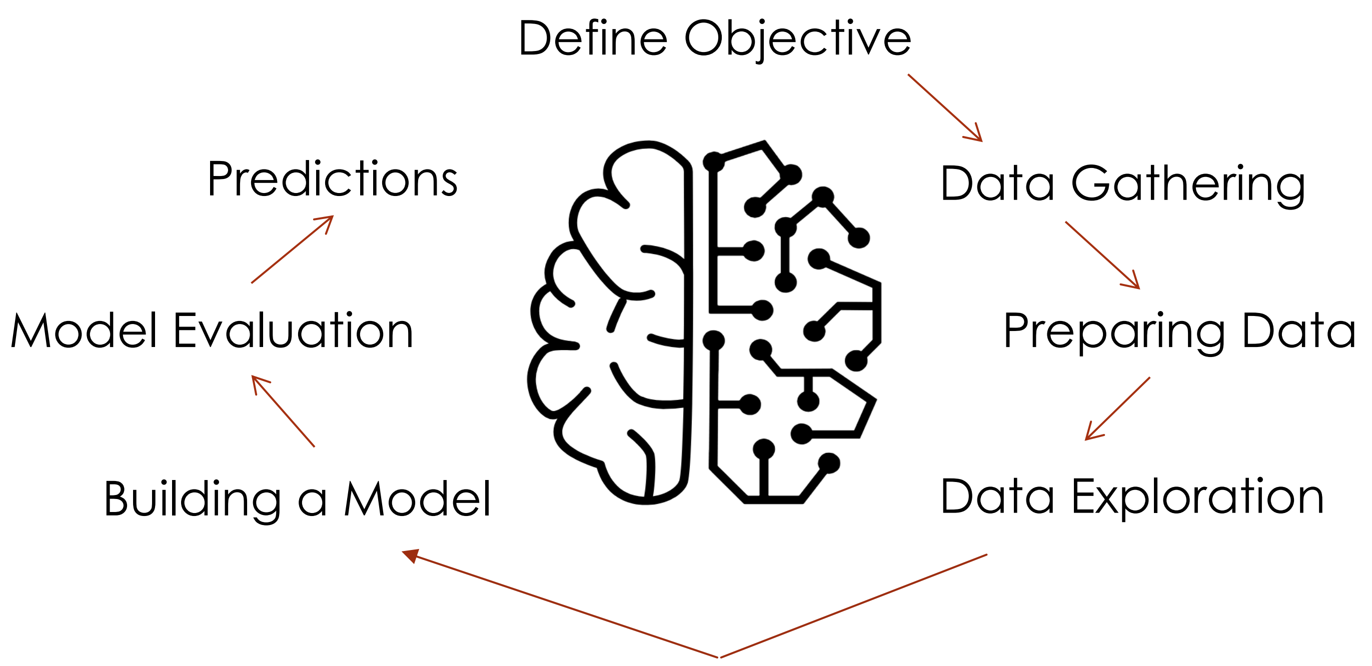


Definitions

- **Algorithms:** A set of rules and statistical techniques used to learn patterns from data.
- **Model:** A model is trained by using a machine learning algorithm.
- **Predictor Variable:** It is a features(s) of the data that can be used to predict the output.
- **Response Variable:** It is the feature or the output variable that needs to be predicted by suing the predictor variable(s).
- **Training Data:** The Machine Learning model is built using the training data.
- **Testing Data:** The Machine Learning model evaluated using the testing data.

Machine Learning Process

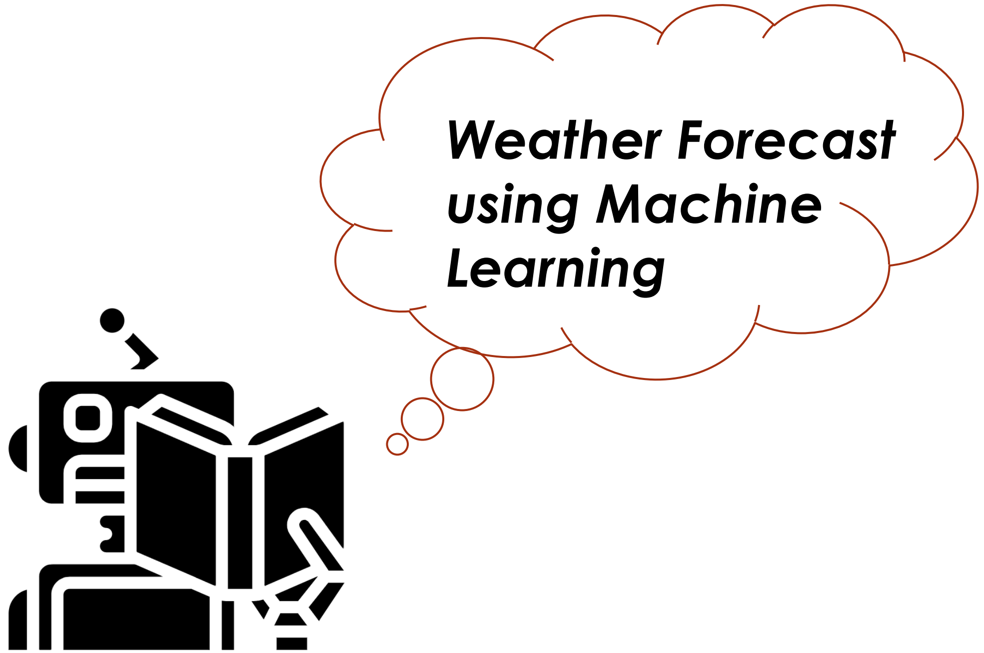
- **Machine Learning Process** involves building **a Predictive model** that can be used to find **a solution** for a **Problem Statement.**



Machine Learning Process

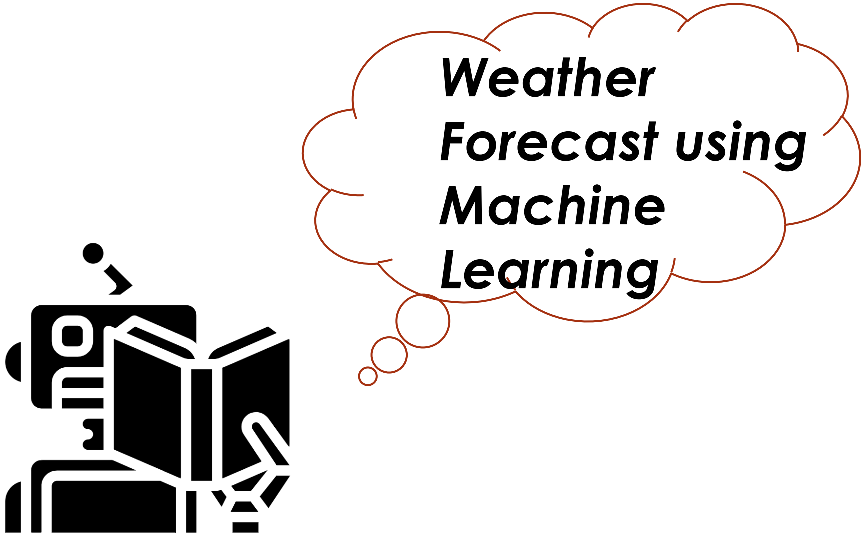
- Step 1: Define the objective of the problem

To predict the possibility of rain by studying the weather conditions



Machine Learning Process

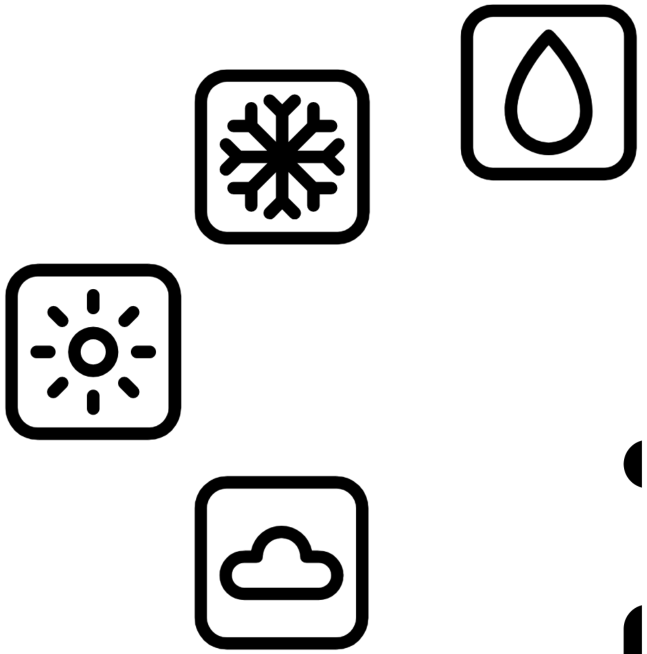
- What we are trying to predict?
- What are the target features?
- What is the input data?
- What kind of problem are we facing? Binary classification? Clustering?

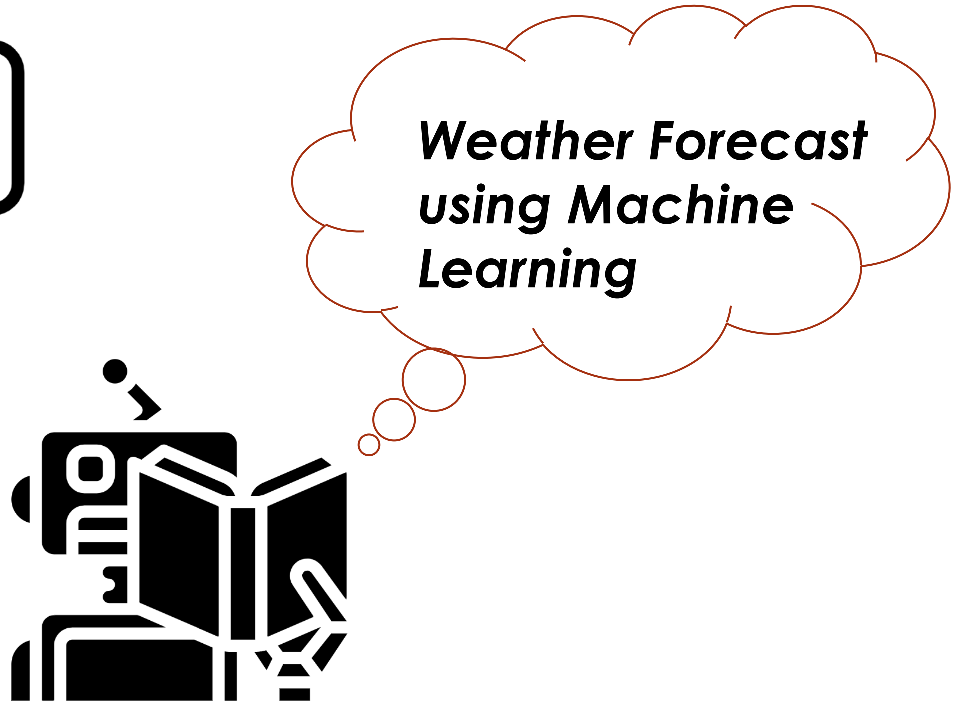


Machine Learning Process

- Step 2: Data Gathering

Data such as weather conditions, humidity level, temperature, pressure etc. are either collected manually or scarped from the web.





Machine Learning Process

- Data Open Sources
- Google Public Data Explorer

https://www.google.com/publicdata/directory

- Registry of Open Data on AWS (RODA)

https://registry.opendata.aws/

- Kaggle

https://www.kaggle.com/datasets

- Dbpedia

https://wiki.dbpedia.org/

Machine Learning Process

- Step 3: Preparing Data

Data Cleaning involves getting rid of inconsistencies in data such as missing values or redundant variables.

- Transform data into desired format
- Data Cleaning
    - Missing values
    - Corrupted data
    - Remove unnecessary data

Machine Learning Process

- Step 4: Exploratory Data Analysis

Data Exploration involves understanding the patterns and trends in the data. At this stage all the useful insights are drawn and correlations between the variables are understood.

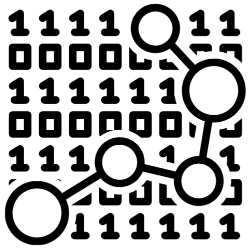
Machine Learning Process

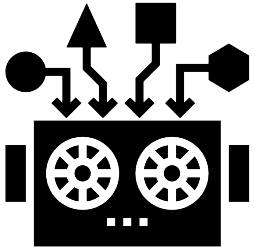
- Step 5: Building a Machine Learning Model

At this stage a Predictive Model is built by using Machine Learning Algorithms such as Linear Regression, Decision Tress, etc.

- Machine Learning model is built by using the training data set.
- The model is the Machine Learning algorithm that predicts the output by using the data fed to it.

Training Data





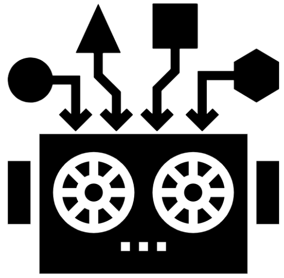
Machine Learning Model

Machine Learning Process

- Step 6: Model Evaluation & Optimization

The efficiency of the model is evaluated and any further improvement in the model are implemented.

- Machine Learning model is evaluated by using the testing data set.
- The accuracy of the model is calculated
- Further improvement in the model are done by using techniques like parameter tuning.

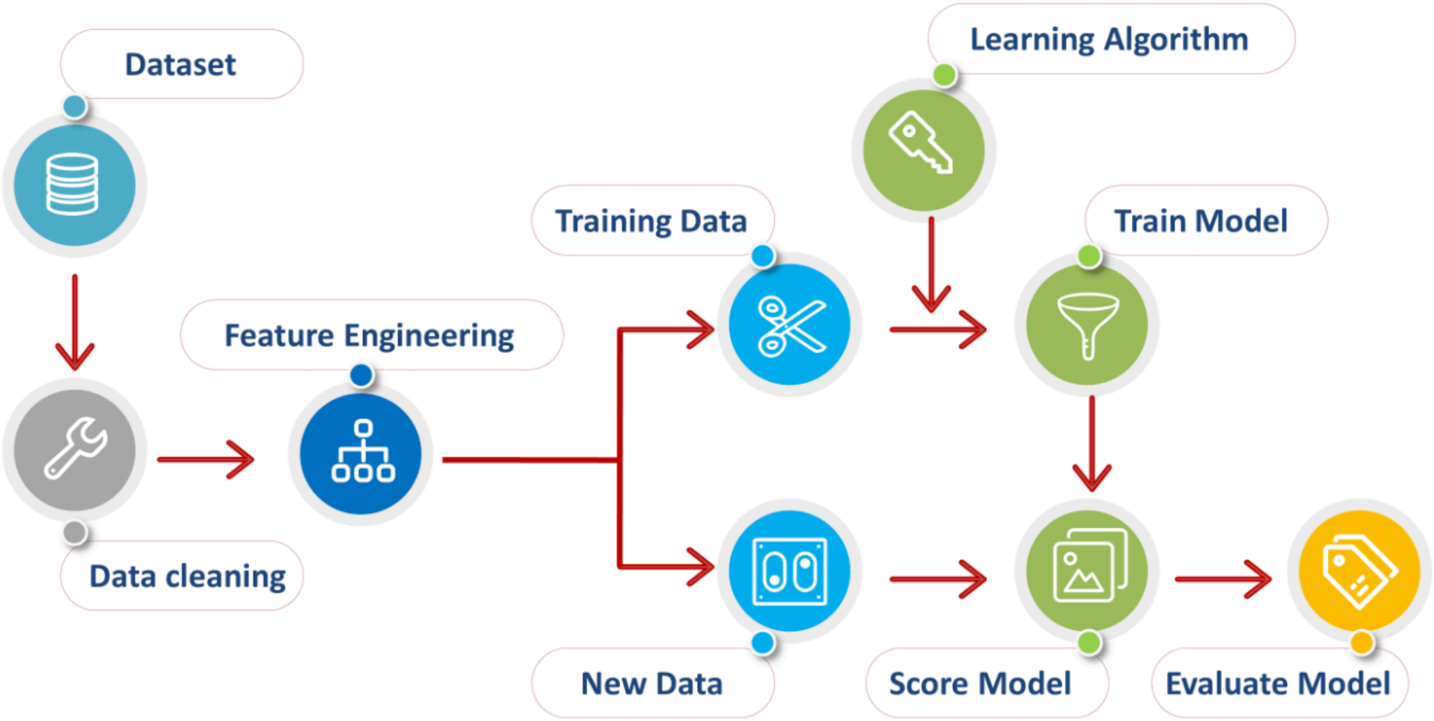


Machine Learning Model

Machine Learning Process

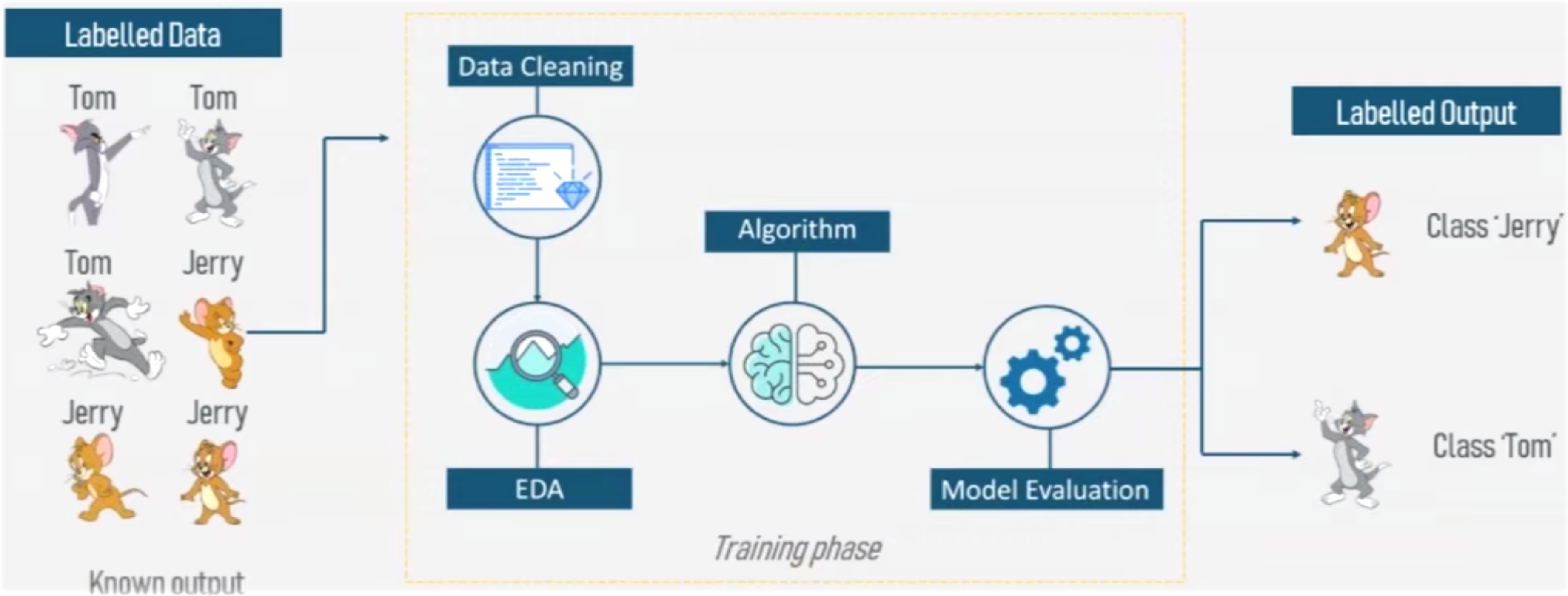
- Step 7: Predictions

The final outcome is predicted after performing parameter tuning and improving the accuracy of the model.



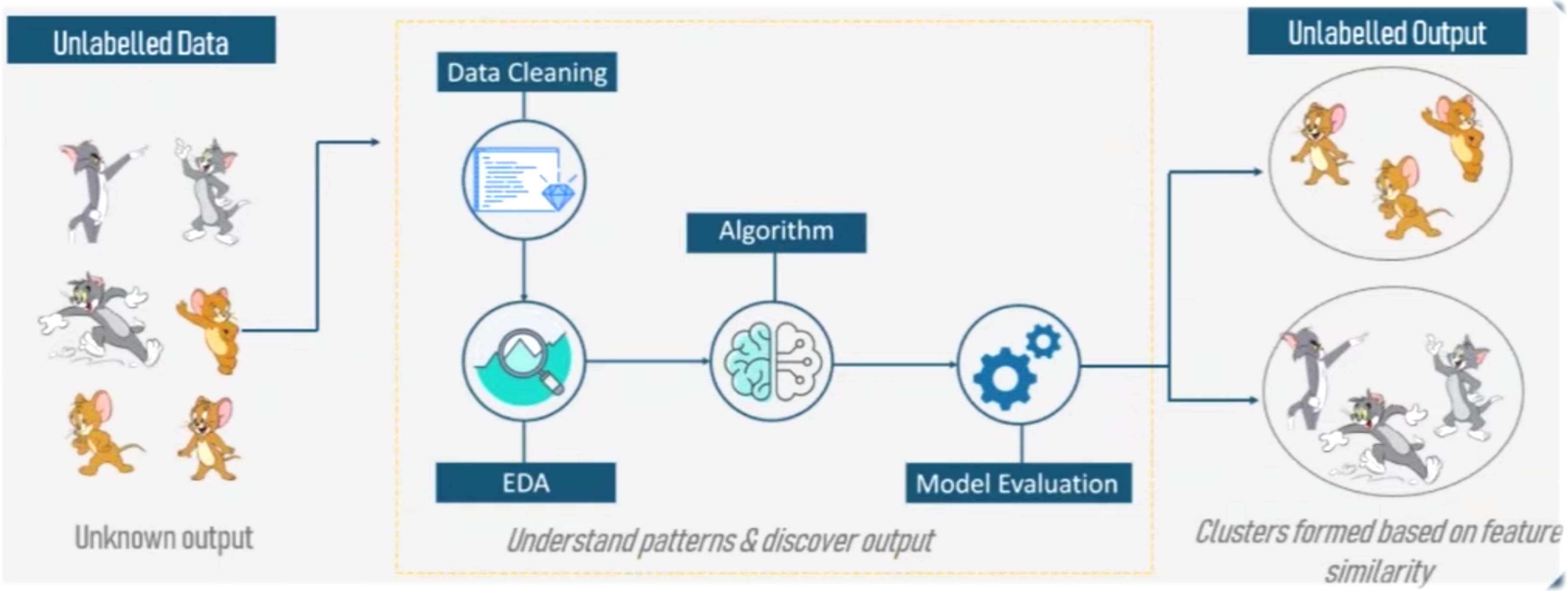
Types of Machine Learning

- ***Supervised Learning*** is a technique in which we teach or train the machine using data which is well labelled.



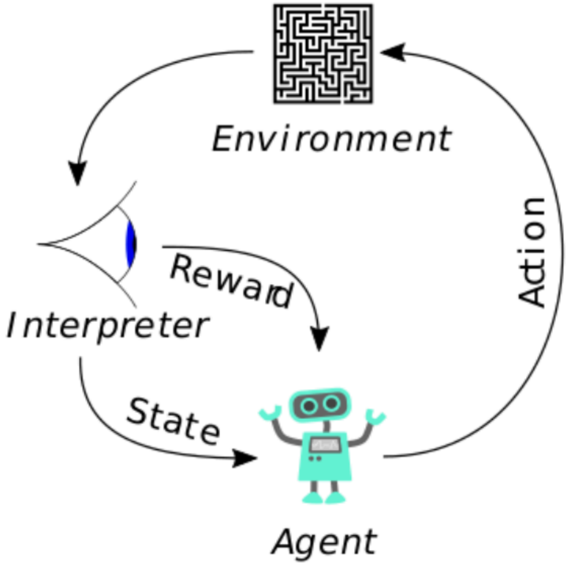
Types of Machine Learning

- ***Unsupervised Learning*** is the training of machine using information that is unlabeled and allowing the algorithm to act on that information without guidance.



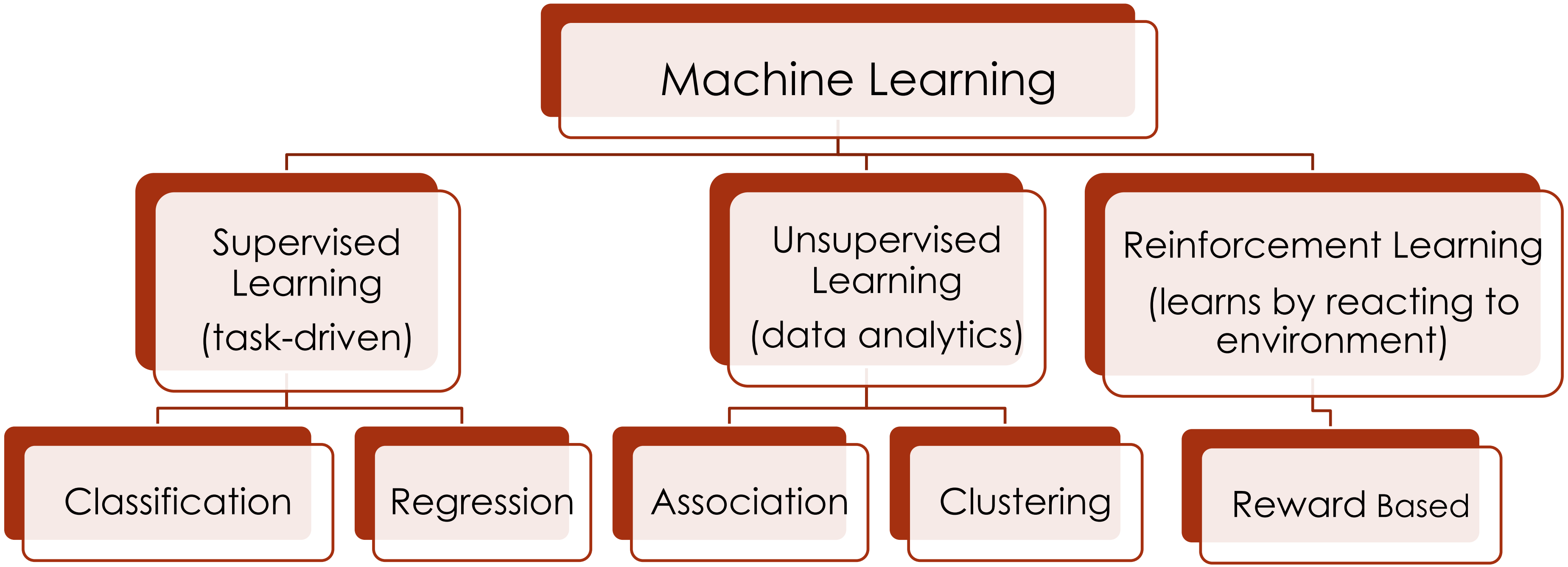
Types of Machine Learning

- ***Reinforcement Learning*** is a part of Machine learning where an agent is put in an environment and he learns to behave in this environment by performing certain actions and observing the rewards which it gets from those actions.
- e.g., self-driving cars, Alpha GO



The typical RL scenario: an agent takes actions in an environment, which is interpreted into a reward and a representation of the state, which are fed back into the agent.

Types of Machine Learning



Types of Machine Learning

- In **Supervised** learning
  - We are given a training set of (X, f(X)) pairs

big nose	big teeth	big eyes	no moustache	f(X) = not person
small nose	small teeth	small eyes	no moustache	f(X) = person

small nose	big teeth	small eyes	moustache	f(X) = ?

Types of Machine Learning

- In **Unsupervised** learning
  - We are only given the Xs - not the corresponding f(X)

big nose	big teeth	big eyes	no moustache	*not given*
small nose	small teeth	small eyes	no moustache	*not given*

small nose	big teeth	small eyes	moustache	f(X) = ?

  - No *teacher* involved / Goal: find regularities among the Xs (clustering)
  - Data mining

Note on Data Mining

- Other names:
- Unsupervised Machine Learning
- Clustering
- Knowledge Discovery
- Example: predict if a customer is likely to purchase certain goods according to history of shopping activities.

Types of Machine Learning

- In **Reinforcement** learning
  - We are not given the (X, f(X)) pairs

small nose	big teeth	small eyes	moustache	f(X) = ?

  - But somehow we are told whether our learned f(X) is right or wrong
  - Goal: maximize the objective of right answers

Types of Machine Learning

	Supervised Learning	Unsupervised Learning	Reinforcement Learning
Definition	The machine learns by using labelled data	The machine is trained on unlabeled data without any guidance	An agent interacts with its environment by producing actions & discovers errors and rewards
Types of problems	Regression &Classification	Association & Clustering	Reward based
Type of data	Labelled data	Unlabelled data	No pre-defined data
Training	External supervision	No supervision	No supervision
Approach	Map labelled input to known output	Understand patterns and discover output	Follow trail and error method
Popular Algorithms	Linear Regression, Logistic Regression, KNN, etc	K-means, C-means, etc	Q-learning, etc

Types of Problems

Regression

- Supervised Learning
- Output is a continuous quantity
- Main aim is to forecast or predict
- Eg: Predict stock market price
- Algorithm: Linear Regression

Classification

- Supervised Learning
- Output is a categorical quantity
- Main aim is to compute the category of the data
- Eg: Classify emails as spam or non-spam
- Algorithm: Logistic Regression

Clustering

- Unsupervised Learning
- Assigns data points into clusters
- Main aim is to group similar items clusters
- Eg: Find all transactions which are fraudulent in nature
- Algorithm: K-means

Logical Inference

- Inference: process of deriving new facts from a set of premises.
- Types of logical inference:
  - Deduction
  - Abduction
  - Induction

Deduction

- also known as Natural Deduction
- Conclusion follows necessary from the premises.
- From $A \Rightarrow B$ and A, we conclude that B
- We conclude from the general case to a specific example of the general case
- Example:

  All men are mortal.

  Marcus is a man.

  Marcus is mortal.

Abduction

- Conclusion is one hypothetical (most probable) explanation for the premises
- From $A \Rightarrow B$ and B, we conclude A
- Example:

  Drunk people do not walk straight.

  John does not walk straight.

  John is drunk.

- Not sound… but may be most likely explanation for B
- Used in medicine…
  - in reality… disease $\Rightarrow$ symptoms
  - patient complains about some symptoms… doctor concludes a disease

Induction

- Conclusion about all members of a class from the examination of only a few member of the class.
- From $A \wedge C \Rightarrow B$ and $A \wedge D \Rightarrow B$, we conclude $A \Rightarrow B$
- We construct a general explanation based on a specific case.
- Example:

  All CS students in COMP 6721 are smart.

  All CS students on vacation are smart.

  All CS students are smart.

- Not sound
- But, can be seen as hypothesis construction or generalisation

Inductive Learning

- = learning from examples , most work in ML
- Examples are given (positive and/or negative) to train a system in a classification (or regression) task
- Extrapolate from the training set to make accurate predictions about future examples
- Can be seen as learning a function
- Given a new instance X you have never seen
- You must find an estimate of the function f(X) where f(X) is the desired output
- Ex:

small nose	big teeth	small eyes	moustache	f(X) = ?

X

  - X = features of a face (ex. small nose, big teeth, …)
  - f(X) = function to tell if X represents a human face or not

Inductive Learning Framework

- Input data are represented by a ***vector of features***, X
- Each vector X is a list of (attribute, value) pairs.
  - Example:

  X = [nose:big, teeth:big, eyes:big, moustache:no]
- The number of attributes is fixed (positive, finite)
- Each attribute has a fixed, finite number of possible values
- Each example can be interpreted as a point in a n-dimensional feature space
  - where n is the number of attributes

Note: *attribute == feature*

Example 0

has-hair?	has-scales?	has-feathers?	flies?	lives in water?	lays eggs?	
1	0	0	0	0	0	Dog
1	0	0	0	0	0	Cat
1	0	0	1	0	0	Bat
1	0	0	0	1	0	Whale
0	0	1	1	0	1	Canary
0	0	1	1	0	1	Robin
0	0	1	1	0	1	Ostrich
0	1	0	0	0	1	Snake
0	1	0	0	0	1	Lizard
0	1	0	0	1	1	Alligator

Real ML applications typically require hundreds, thousands or millions of examples

Example 1

- Problem Statement: To study the House Sales dataset and build a Machine Learning model that predicts the house pricing index.

```
> str(data)
'data.frame':   21613 obs. of  21 variables:
 $ id           : num  7.13e+09 6.41e+09 5.63e+09 2.49e+09 1.95e+09 ...
 $ date         : Factor w/ 372 levels "20140502T000000",..: 165 221 291 221 284 11 57
 $ price        : num  221900 538000 180000 604000 510000 ...
 $ bedrooms     : int  3 3 2 4 3 4 3 3 3 3 ...
 $ bathrooms    : num  1 2.25 1 3 2 4.5 2.25 1.5 1 2.5 ...
 $ sqft_living  : int  1180 2570 770 1960 1680 5420 1715 1060 1780 1890 ...
 $ sqft_lot     : int  5650 7242 10000 5000 8080 101930 6819 9711 7470 6560 ...
 $ floors       : num  1 2 1 1 1 1 2 1 1 2 ...
 $ waterfront   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ view         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ condition    : int  3 3 3 5 3 3 3 3 3 3 ...
 $ grade        : int  7 7 6 7 8 11 7 7 7 7 ...
 $ sqft_above   : int  1180 2170 770 1050 1680 3890 1715 1060 1050 1890 ...
 $ sqft_basement: int  0 400 0 910 0 1530 0 0 730 0 ...
 $ yr_built     : int  1955 1951 1933 1965 1987 2001 1995 1963 1960 2003 ...
 $ yr_renovated : int  0 1991 0 0 0 0 0 0 0 0 ...
 $ zipcode      : int  98178 98125 98028 98136 98074 98053 98003 98198 98146 98038 ...
 $ lat          : num  47.5 47.7 47.7 47.5 47.6 ...
 $ long         : num  -122 -122 -122 -122 -122 ...
 $ sqft_living15: int  1340 1690 2720 1360 1800 4760 2238 1650 1780 2390 ...
 $ sqft_lot15   : int  5650 7639 8062 5000 7503 101930 6819 9711 8113 7570 ...
```

→ Linear Regression Algorithm

↓

Predict the house pricing index

Regression

Example 2

- Problem Statement: To study a bank credit dataset and make a decision about whether to approve the loan of an applicant based on his profile.

```
$ Account.Balance                    : int  1 1 2 1 1 1 1 1 4 2 ...
$ Duration.of.Credit..month.         : int  18 9 12 12 12 10 8 6 18 24 ..
$ Payment.Status.of.Previous.Credit: int  4 4 2 4 4 4 4 4 4 2 ...
$ Purpose                            : int  2 0 9 0 0 0 0 0 3 3 ...
$ Credit.Amount                      : int  1049 2799 841 2122 2171 2241
$ Value.Savings.Stocks               : int  1 1 2 1 1 1 1 1 1 3 ...
$ Length.of.current.employment       : int  2 3 4 3 3 2 4 2 1 1 ...
$ Instalment.per.cent                : int  4 2 2 3 4 1 1 2 4 1 ...
$ Sex...Marital.Status               : int  2 3 2 3 3 3 3 3 2 2 ...
$ Guarantors                         : int  1 1 1 1 1 1 1 1 1 1 ...
$ Duration.in.Current.address        : int  4 2 4 2 4 3 4 4 4 4 ...
$ Most.valuable.available.asset      : int  2 1 1 1 2 1 1 1 3 4 ...
$ Age..years.                        : int  21 36 23 39 38 48 39 40 65 23
$ Concurrent.Credits                 : int  3 3 3 3 1 3 3 3 3 3 ...
$ Type.of.apartment                  : int  1 1 1 1 2 1 2 2 2 1 ...
$ No.of.Credits.at.this.Bank         : int  1 2 1 2 2 2 2 1 2 1 ...
$ Occupation                         : int  3 3 2 2 2 2 2 2 1 1 ...
$ No.of.dependents                   : int  1 2 1 2 1 2 1 2 1 1 ...
$ Telephone                          : int  1 1 1 1 1 1 1 1 1 1 ...
$ Foreign.Worker                     : int  1 1 1 2 2 2 2 2 1 1 ...
```

→ KNN Algorithm

Approve   Reject

Classification

Example 3

- Problem Statement: To cluster a set of movies as either good or average based on their social media out reach.

	director_facebook_likes	actor_3_facebook_likes	actor_1_facebook_likes	cast_total_facebook_likes
AvatarÂ	0	855	1000	4834
Pirates of the C...	563	1000	40000	48350
SpectreÂ	0	161	11000	11700
The Dark Knigh...	22000	23000	27000	106759
John CarterÂ	475	530	640	1873
Spider-Man 3Â	0	4000	24000	46055
TangledÂ	15	284	799	2036
Avengers: Age ...	0	19000	26000	92000
Harry Potter an...	282	10000	25000	58753
Batman v Super...	0	2000	15000	24450
Superman Retur...	0	903	18000	29991
Quantum of Sol...	395	393	451	2023
Pirates of the C...	563	1000	40000	48486

→ K-means Algorithm

Popular

Unpopular

Clustering

Supervised Learning Algorithms

- Linear Regression
- Logistic Regression
- Decision Tree
- Random Forest
- Naïve Bayes Classifier

Linear Regression

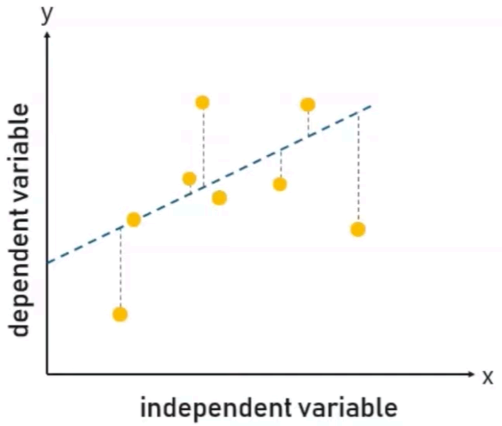
- **Linear Regression** is a method to predict dependent variable (Y) based on values of independent variables (X). It can be used for the cases where we want to predict some continuous quantity.
- Dependent variable (Y)

The response variable who's value needs to be predicted.

- Independent variable (X)

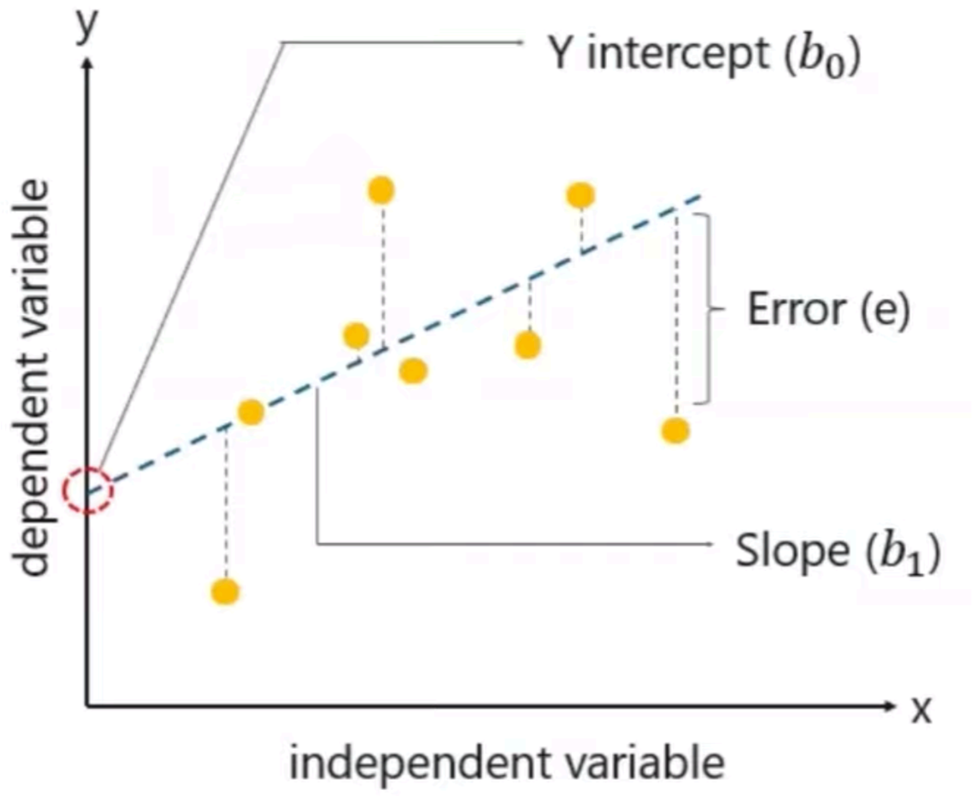
The predictor variable used to predict the response variable.

- The following equation is used to represent a linear regression model:



$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

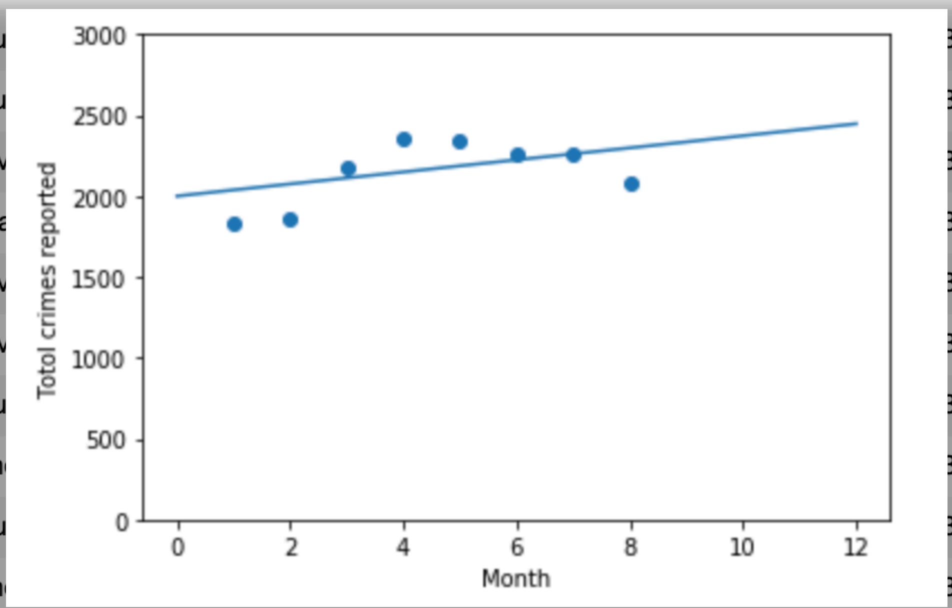
Linear Regression



$$\underset{\text{Dependent Variable}}{Y_i} = \underbrace{\underset{\text{Population Y intercept}}{\beta_0} + \underset{\text{Population Slope Coefficient}}{\beta_1} \underset{\text{Independent Variable}}{X_i}}_{\text{Linear component}} + \underbrace{\underset{\text{Random Error term}}{\varepsilon_i}}_{\text{Random Error component}}$$

Example

	CATEGORIE	DATE	QUART	PDQ	X	Y	LONGITUDE	LATITUDE	YEAR	MONTH
117	Vol dans / sur véhicule à moteur	2015-12-10	soir	46.0	299675.192008	5.050911e+06	-73.565691	45.598083	2015	12
118	Vol dans / sur véhicule à moteur	2015-12-11	jour	21.0	300642.816009	5.040606e+06	-73.553200	45.505362	2015	12
157	Introdu						3.591324	45.643140	2015	12
158	Introdu						3.565691	45.598083	2015	12
159	M						3.649992	45.561410	2015	12
160	Vols qua						3.659137	45.547889	2015	12
161	M						3.655480	45.551530	2015	12
168	M						3.655480	45.551530	2015	12
206	Introdu						3.585057	45.547853	2015	3
207	Vol dans / sur véhicule à m						3.698049	45.480716	2015	3
208	Introdu						3.549800	45.530277	2015	1
209	Vol dans / sur véhicule à m						3.689860	45.487279	2015	1
210	Vol de véhicule à moteur	2015-01-15	soir	5.0	286009.859000	5.034484e+06	-73.740224	45.450037	2015	1
211	Vol dans / sur véhicule à moteur	2015-01-16	nuit	26.0	294021.501009	5.038488e+06	-73.637887	45.486229	2015	1
212	Vol de véhicule à moteur	2015-01-29	soir	8.0	287386.553000	5.035839e+06	-73.722671	45.462264	2015	1



Supervised Learning Algorithms

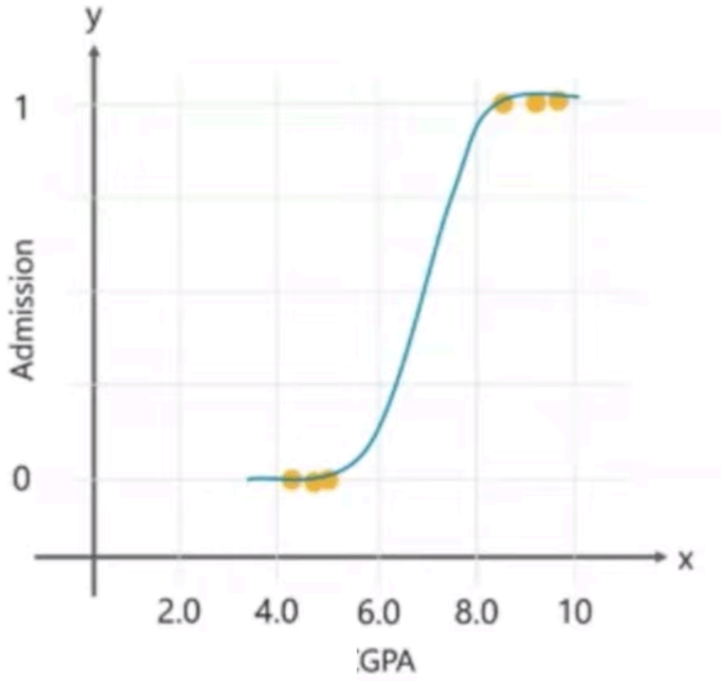
- ***Linear Regression***
- Logistic Regression
- Decision Tree
- Random Forest
- Naïve Bayes Classifier

Logistic Regression

- **Spam Detection** : Predicting if an email is Spam or not
- **Credit Card Fraud** : Predicting if a given credit card transaction is fraud or not
- **Health** : Predicting if a given mass of tissue is benign or malignant
- **Marketing** : Predicting if a given user will buy an insurance product or not
- **Banking** : Predicting if a customer will default on a loan.

Logistic Regression

- ***Logistic Regression*** is a method used to predict a dependent variable, given a set of independent variables, such that the dependent variable is **categorical**.
- Logistic Regression is used for classification.



Logistic Regression

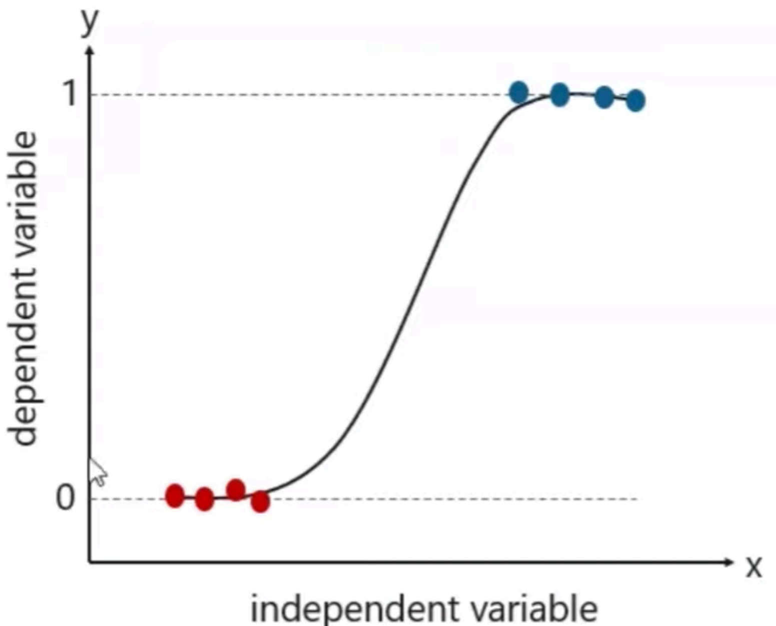
- Linear Regression equation: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Representing a relationship between *p(X) = Pr(Y=1 | X)* and X ?

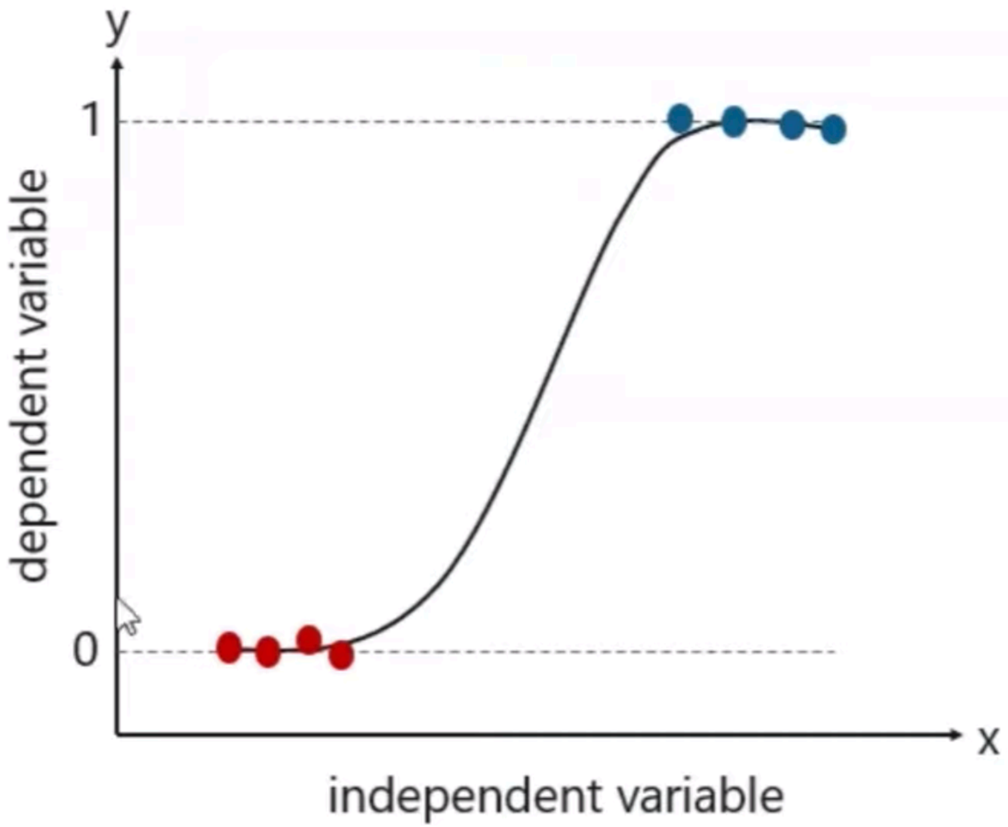
  - Take the exponent of the equation, since the exponential of any value is a positive number.
  - Secondly, a number divided by itself + 1 will always be less than 1.

Hence, the formula :

$$P(X) = \frac{e^{(\beta_0+\beta_1 x)}}{e^{(\beta_0+\beta x)} + 1}$$



Logistic Regression



$$P(X) = \frac{e^{(\beta_0+\beta_1 x)}}{e^{(\beta_0+\beta x)} + 1}$$

$$\Longrightarrow \quad p(e^{(\beta_0+\beta x)}+1) = e^{(\beta_0+\beta_1 x)}$$

$$\Longrightarrow \quad p.\,e^{(\beta_0+\beta x)} + p = e^{(\beta_0+\beta_1 x)}$$

$$\Longrightarrow \quad p = e^{(\beta_0+\beta_1 x)} - p.\,e^{(\beta_0+\beta x)}$$

$$\Longrightarrow \quad p = e^{(\beta_0+\beta_1 x)}\,(1-p)$$

$$\Longrightarrow \quad \frac{p}{(1-p)} = e^{(\beta_0+\beta_1 x)}$$

$$\Longrightarrow \quad \ln\left[\frac{p}{(1-p)}\right] = (\beta_0 + \beta_1 x)$$

Supervised Learning Algorithms

- ***Linear Regression***
- ***Logistic Regression***
- Decision Tree
- Random Forest
- Naïve Bayes Classifier

The End