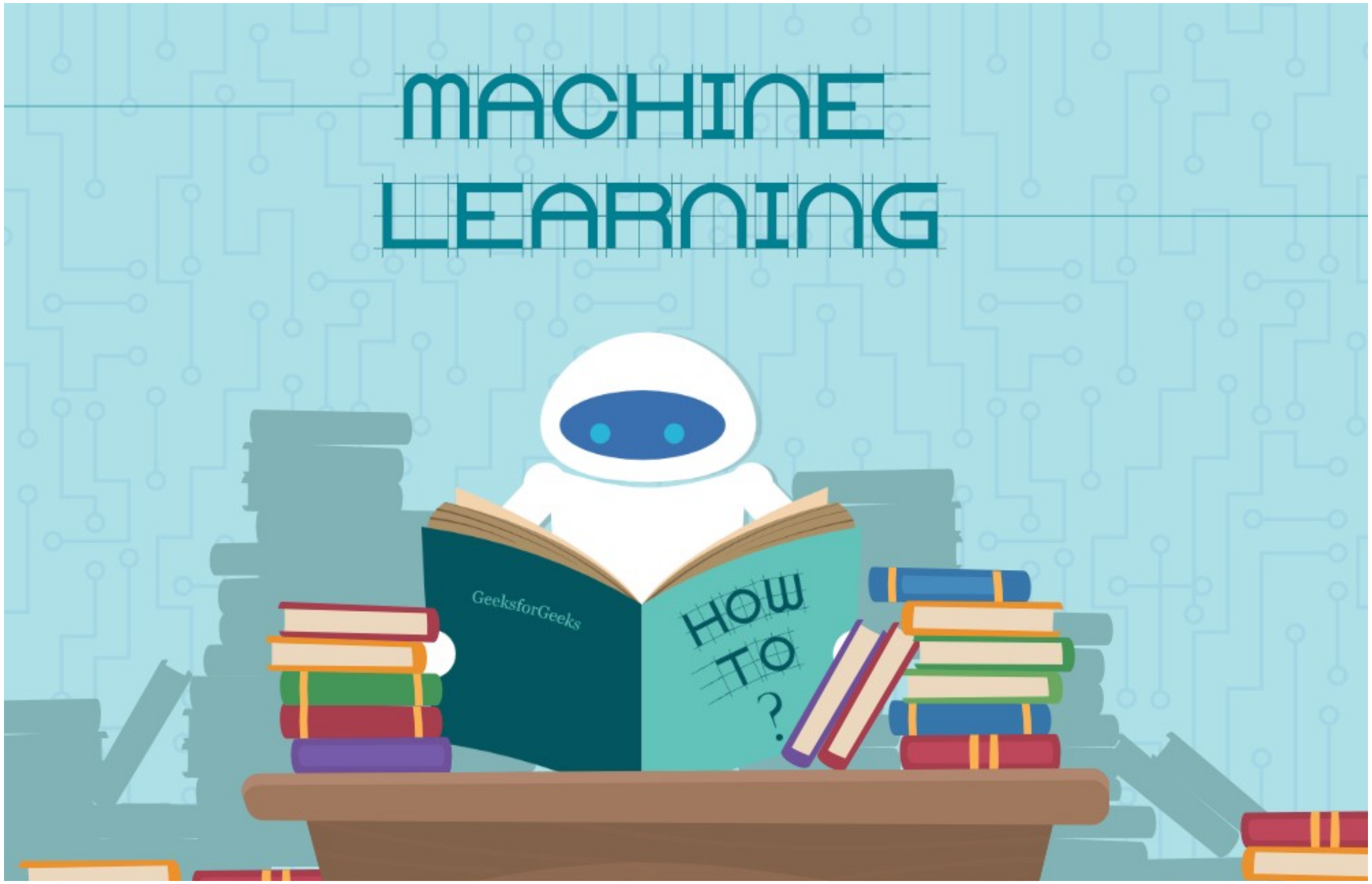
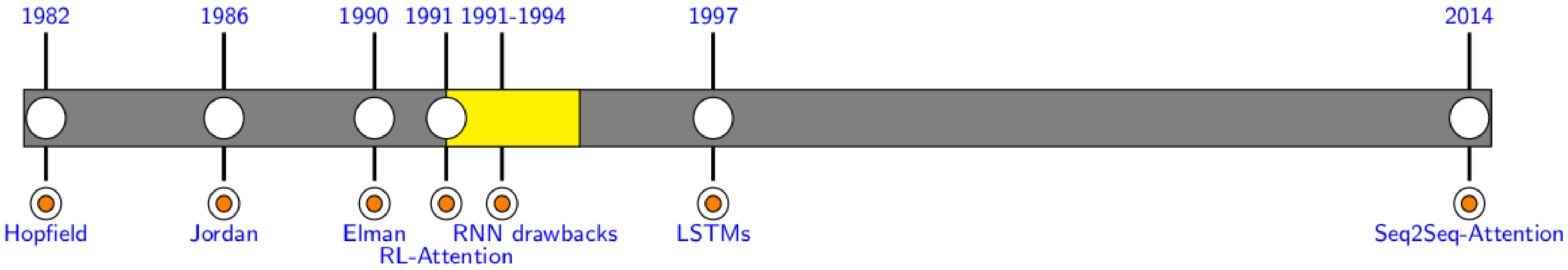
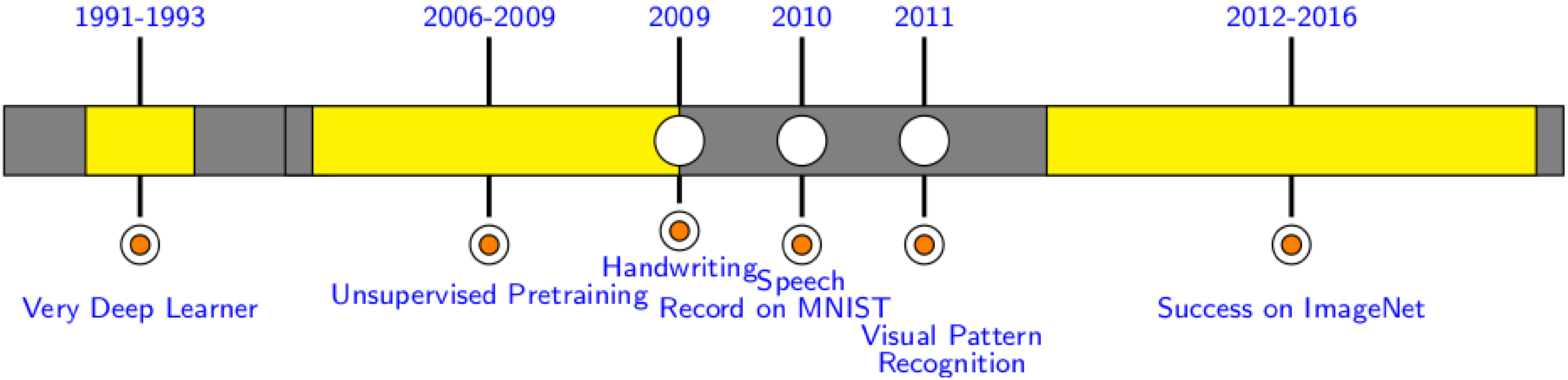


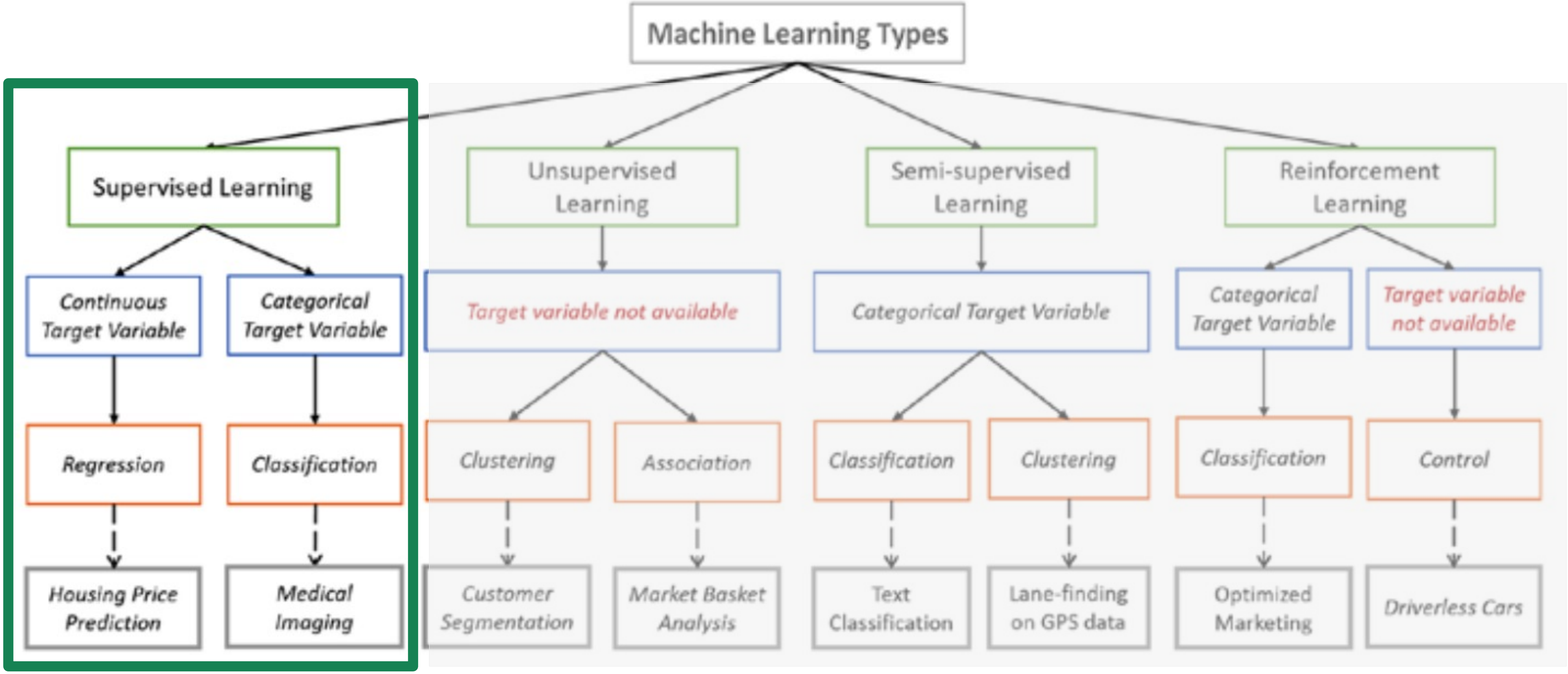
Image Courtesy : GeeksforGeeks

Is ML a New Field ?

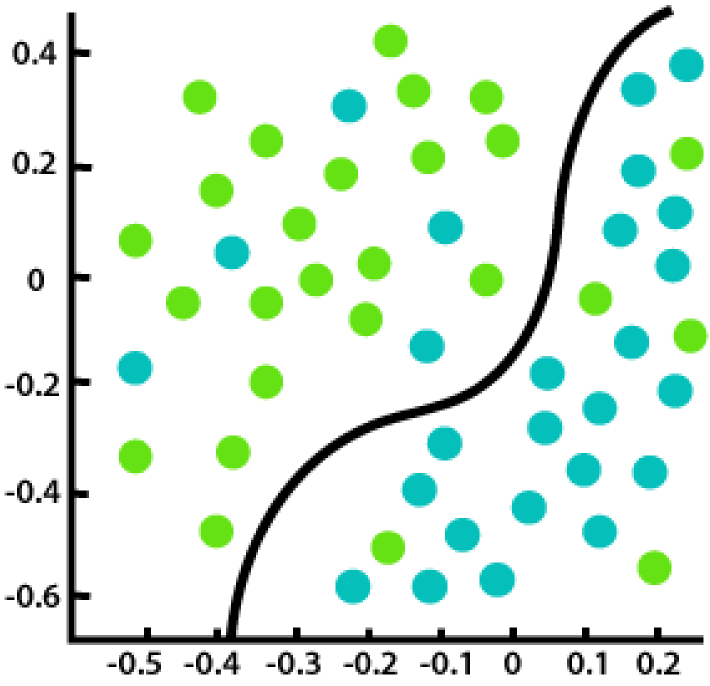


Not Really

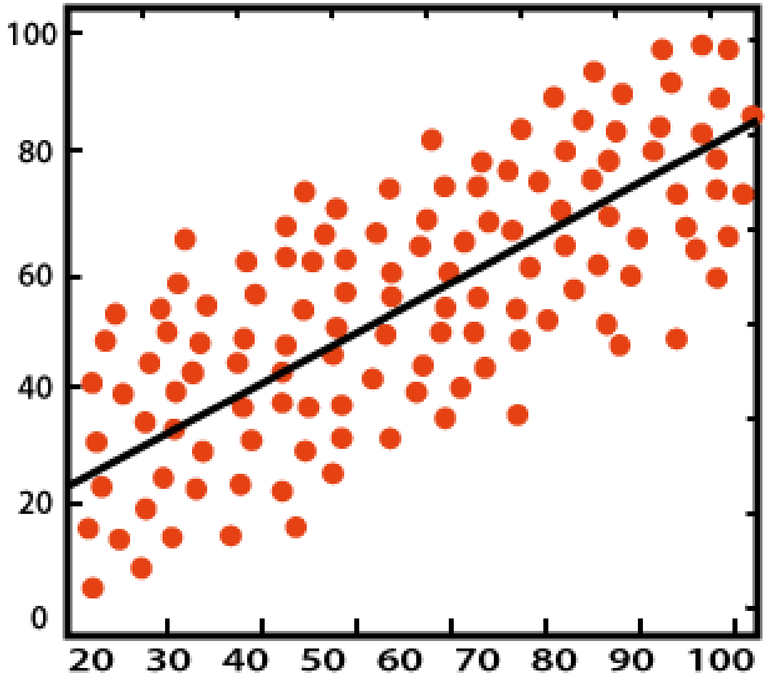



Machine Learning Types
Supervised Learning
Unsupervised Learning
Semi-supervised Learning
Reinforcement Learning
Continuous Target Variable
Categorical Target Variable
Target variable not available
Categorical Target Variable
Categorical Target Variable
Target variable not available
Regression
Classification
Clustering
Association
Classification
Clustering
Classification
Control
Housing Price Prediction
Medical Imaging
Customer Segmentation
Market Basket Analysis
Text Classification
Lane-finding on GPS data
Optimized Marketing
Driverless Cars

Supervised Learning



Classification



Regression

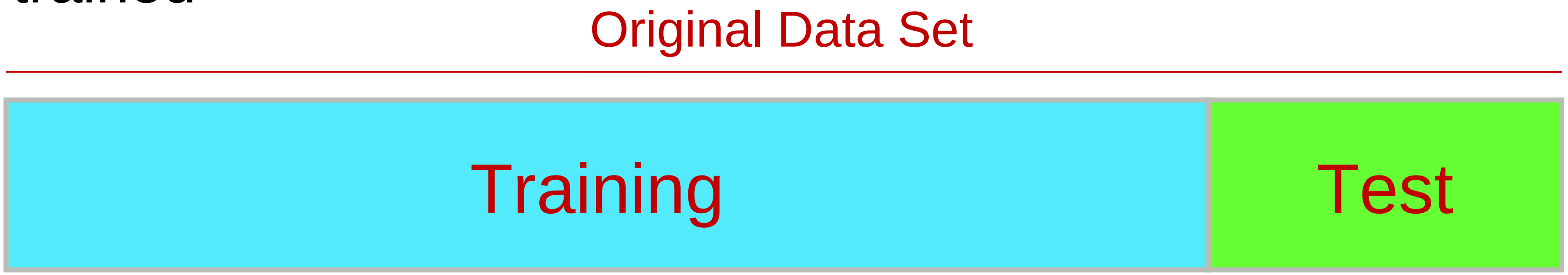
Prepared by: Prof. Siddharth P. Shah & Prof. Hariom A. Pandya

Classification Or Regression ?

1) Predicting Tomorrow’s Temprature
2) Identify the Vehicle Type
3) Detecting the Face region from Image
4) Decide the year when a song is released based on the description of a song
5) Categorise the movie based on the review (best,good,worst)
6) Spam email Detection
7) Estimate the relationship between the stock market and economic
8) Given some description of a person, predict their income
9) Predict GATE score based on the CPI

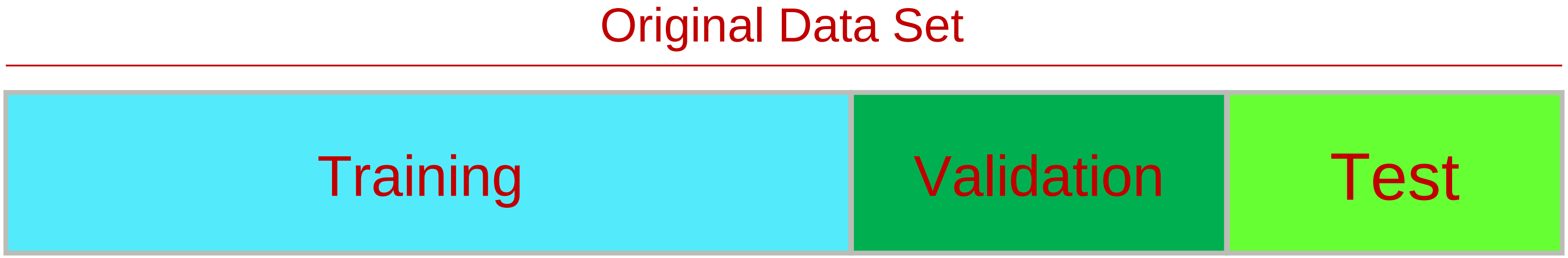
Data Set Partition

- Generally the data set is partitioned in two independent sets
- **Training Set :** is the subsection of a dataset from which the machine learning algorithm uncovers, or “learns,” relationships between the features and the target variable.
- **Test Set (Holdout dataset):** is the subset of the original dataset (independent of the training set) which is used to estimate the model’s performance after it has been trained



Data Set Partition

- Sometimes the data set is partitioned in three independent sets rather than two.
- Apart from Training and Test data set the third partition called Validation data set is also created.
- **Validation Set:** is a set of data used for parameter tuning and/or model selection



Original Data Set

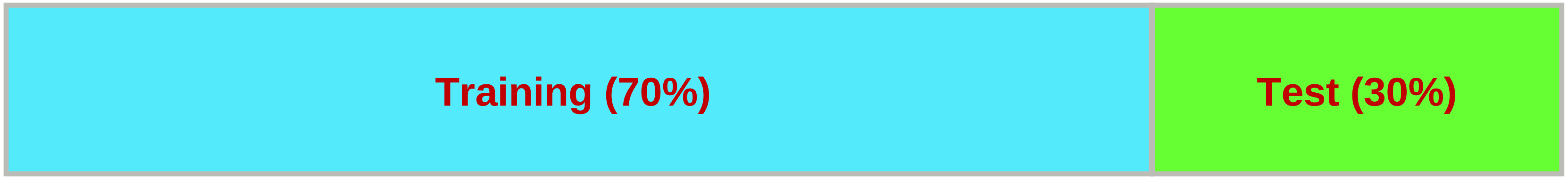
Training

Validation

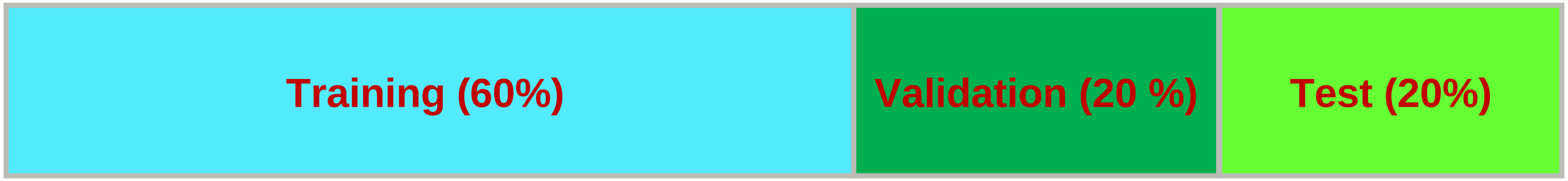
Test

Data Set Partition

- No hard and fast rule for the proportion of each of the partition.
- In case of two partitions it can be70-30



- In case of three partitions it can be 60-20-20



Introduction to Classification

Classification

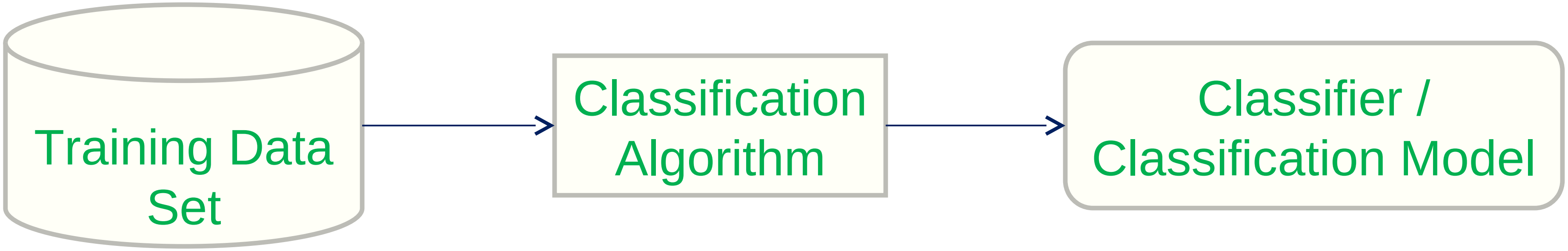
- Classification is the process of predicting the class of given data instance.
- Examples
  - To predict that the email is spam or ham (not spam)
  - To predict the sentiment orientation (positive or negative)
- Classes are sometimes called as targets, labels or categories.
- Predictive modelling for classification is the task of approximating a mapping function (f) from input variables (x) to discrete output variables (y).
- Mapping function (f) is also called hypothesis

Applications of Classification

- Speech Recognition
- Handwriting Recognition
- Biometric Identification
- Document classification
- Sentiment analysis
- Author Identification
- Face detection
- Video actor extractor
- Vehicle identification for Video
- …..

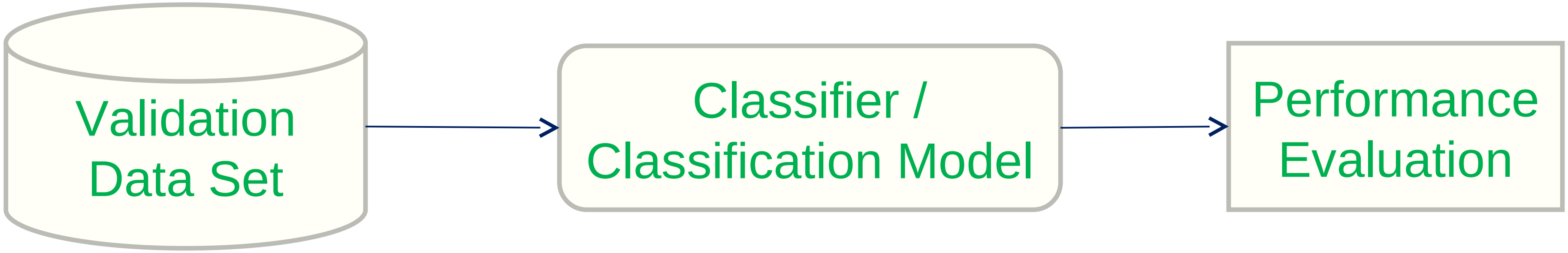
Classification – A Three-step Process

1. **Training Phase:**
   - In this step the classification algorithms build the classifier / classification model.
   - The classifier is built from the training set made up of database tuples and their associated class labels.
   - These tuples can also be referred to as sample, object, instance or data points.



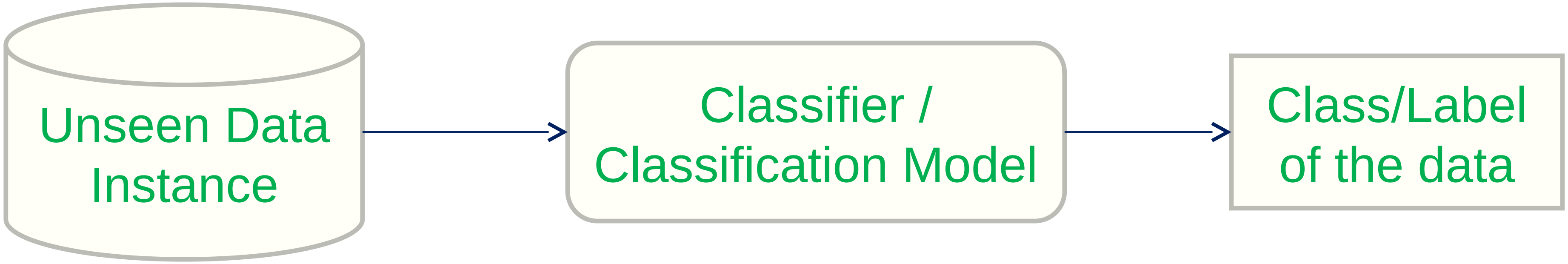
Classification – A Three-step Process

2. **Validation Phase:**
   - In this step, the validaiton data is used to evaluate the performance of classification model.
   - The model evaluation can be performed using different metrics depending on the nature of the application under consideration.
   - **Some of the evaluation metrics are :** Accuracy, Error Rate, Sensitivity (recall), Specificity, Precision, F1-score etc..



Classification – A Three-step Process

3. **Classification/Test Phase:**
   - In this step, developed model is applied to the real-world data instance to predict the class / label for it.

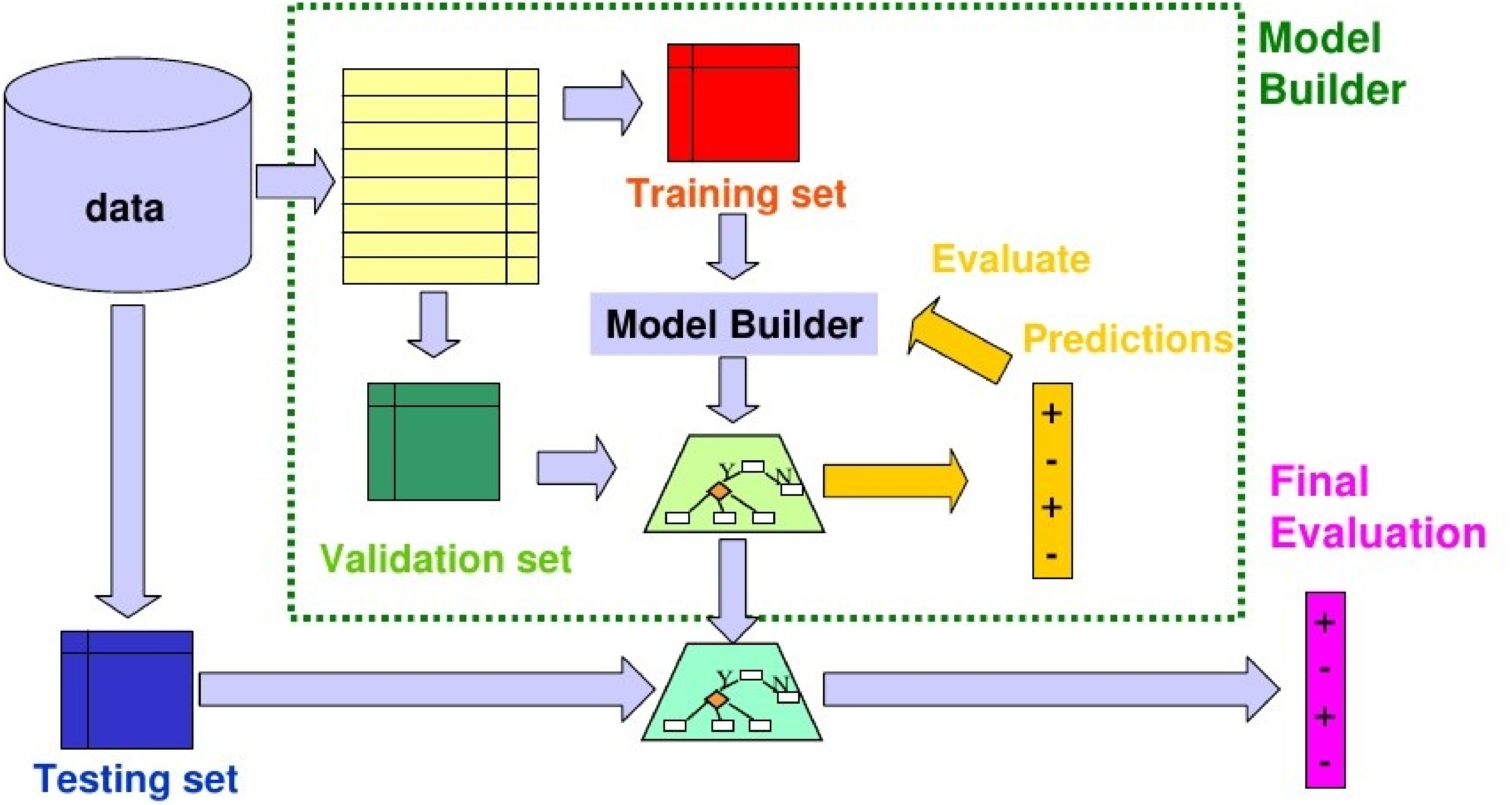


Unseen Data Instance

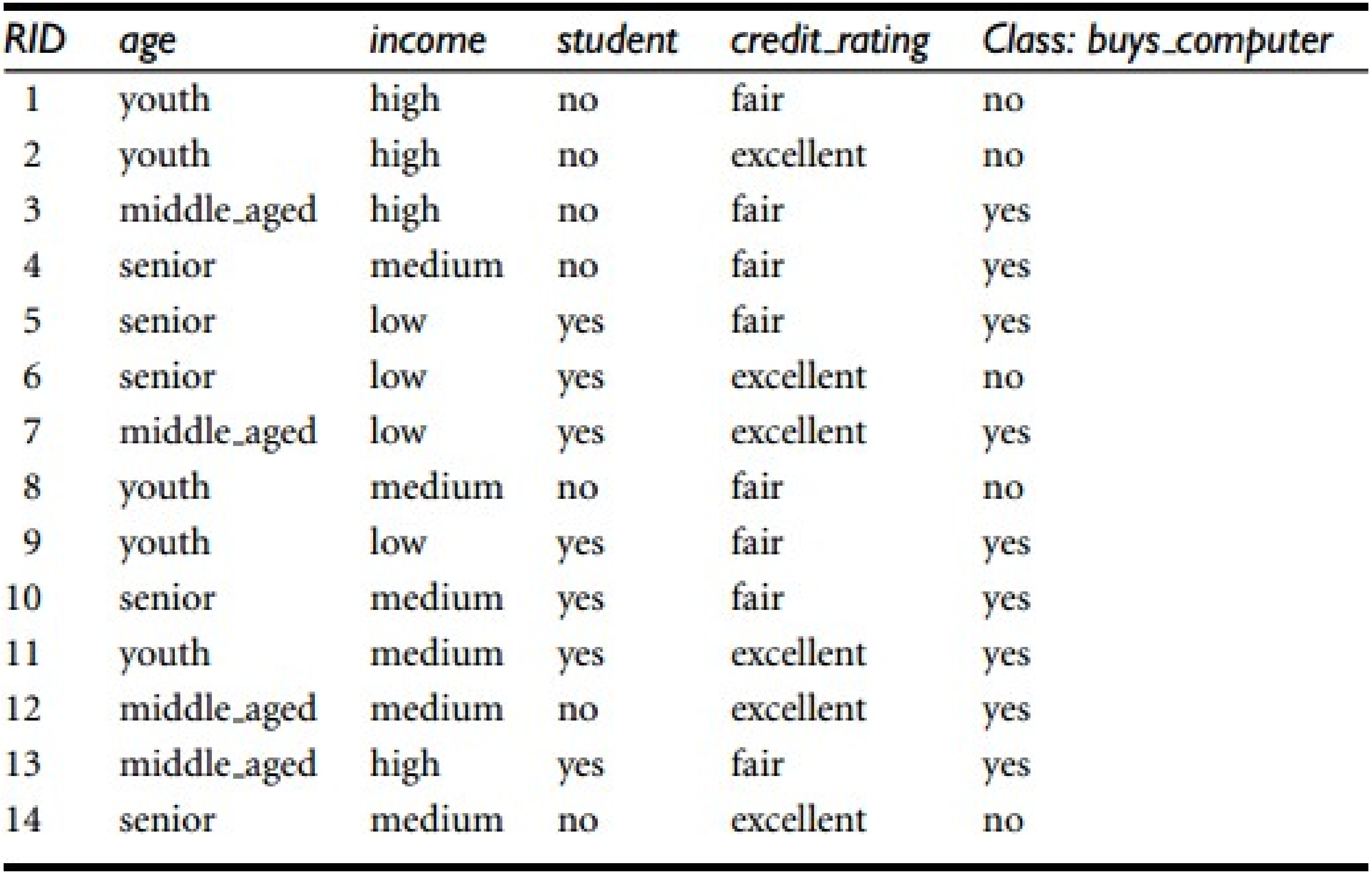
Classifier / Classification Model

Class/Label of the data

Classification Process



Data Set - Example



RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Source : A book “**Data Mining : Concepts and Techniques by Han, Kamber & Pei”**

Types of Classification

- **Binary Classification:** Classification with only 2 distinct classes or with 2 possible outcomes
  - Examples:
    - Male or Female
    - Spam or Ham email
    - Positive or Negative Sentiment

- **Multi-class Classification:** Classification with more than two distinct classes.
  - Examples:
    - Classification of types of crops
    - Classification of mood/feelings in songs/music

Classification Algorithms:

Generative V/s Discriminative

Learning and Prediction

- Boy:
  - Draw the image of lion based on what he saw inside the zoo.
  - He compared both the images with the toy
  - Answered based on the *closest match* of image & toy
  - "Explicitely models the actual distribution of each class" : Generative Model
- Girl:
  - knows only the differences,
  - based on different properties learned
  - Answer is based on learned features
  - "Decision boundary between the classes" : Descriminative Model

Background

- An **event** is a set of outcomes(one or more) from an experiment.
  - Ex. Getting a *Tail* when tossing a coin
  - 1) Independent : Tossing a coin two times
  - 2) Dependent : Drawing 2 Cards from a Deck
  - 3) Mutually Exclusive (Never happen at same time): play football & rugby at the same time
- **Probability** is the likelihood of an event occurring
- **Joint Probability**: likelihood of more than one event occurring at the same time P(A ∩ B)= P(A,B)=P(A).P(B)
  - Ex. probability that a card is a four and red =
    - p(four and red) = 2/52=1/26
      - p(four): 4/52 and p(red):1/2

Joint Probability Requirement

- events A and B must happen at the same time
- events A and B must be independent of each other

- What if events are dependent ?

  Event X is the probability there are clouds in the sky
  Event Y is the probability that it rain

- Joint probability cannot be used to determine how much the occurrence of one event influences the occurrence of another event.
- Therefore the joint probability of X and Y (two dependent events) will be P(Y)

Conditional Probability

- The conditional probability of an event B is the probability that the event will occur given the knowledge that an event A has already occurred.
- It is denoted by P(B|A)
- The joint probability of two dependent events becomes
  - P(A and B) = P(A)P(B|A)

Generative and Discriminative Classifiers

- **Generative Classifier:** tries to learn the model that generates the data by estimating the assumptions and distributions of the model.
  - Examples: Naive Bayes Classifier, Bayesian Networks, Hidden Markov Models (HMM)
- **Discriminative Classifier:** tries to model by just depending on the observed data. It makes fewer assumptions on the distributions but depends heavily on the quality of the data
  - Examples : Decision Tree, Logistic Regression, SVM, ANN, K-NN,
- **Example:**
  - software engineer lives in Silicon Valley (X),
  - what is the probabilityP(Y|X), he makes over 100K salary (Y)

Generative Classifiers

- **To find P(Y|X):**

1) what is the probability a software engineer living in Silicon Valley **(X)** given he makes over 100K **(Y)**, i.e. together it means **P(X|Y)**
2) Then, it estimates how many software engineers make over 100K, regardless of if he is in Silicon Valley or not, i.e. **P(Y)**.
3) Ratio of number of software engineers live in Silicon Valley to the total number of software engineers will give **P(X)**
4) Get P(Y|X) using Bayes Theorem

$$P(Y|X) = \frac{P(X|Y)\,P(Y)}{P(X)}$$

Discriminative Classifiers

- **To find P(Y|X)**
  - Estimate P(Y|X) from the data

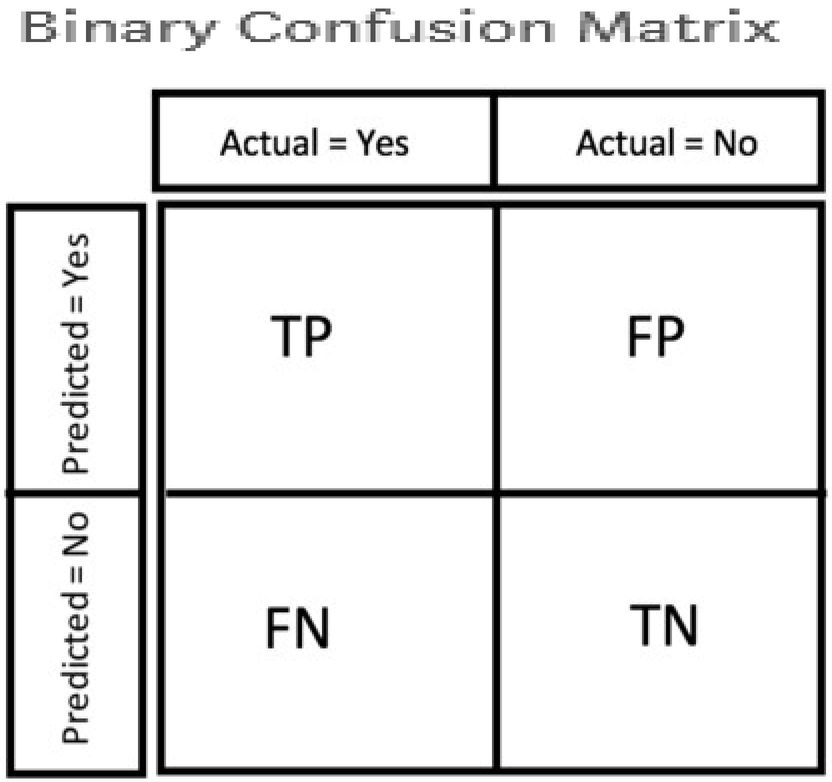
- probability of a software engineer making over 100K salary given he lives in Silicon Valley, i.e. **P(Y|X).**

Performance Measures

- Classification model performance can be evaluated using different metrics.
- Some of them are listed below:
  - Accuracy
  - Precision
  - Recall
  - F1-Score

Confusion Matrix

- It is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known.



Binary Confusion Matrix

	Actual = Yes	Actual = No
Predicted = Yes	TP	FP
Predicted = No	FN	TN

TP	:	True Positive
FP	:	False Positive
TN	:	True Negative
FN	:	False Negative

Terms Associated with Confusion Matrix

- **True Positives (TP):** The total number of accurate predictions that were “positive.” e.g. the total number of emails correctly predicted as spam.

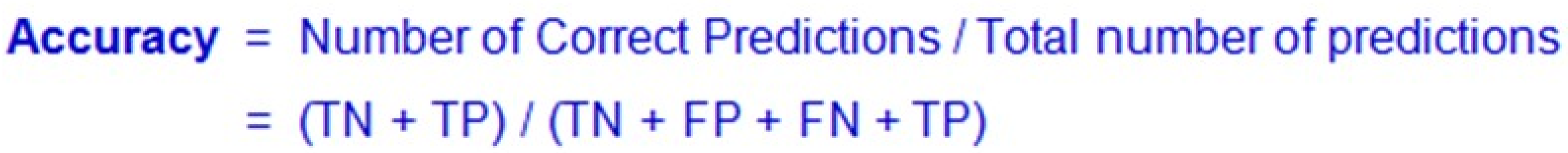
- **False Positives (FP):** The total number of inaccurate predictions that were “positive.” e.g. the total number of emails incorrectly predicted as spam.

- **True Negative (TN):** The total number of accurate predictions that were “negative.” e.g. the total number of emails correctly predicted as non-spam.

- **False Negative (FN):** The total number of inaccurate predictions that were “negative.” e.g. the total number of emails incorrectly predicted as non-spam.

Accuracy

- **Accuracy:** It is the ratio of number of correct predictions to total number of predictions.



Accuracy = Number of Correct Predictions / Total number of predictions

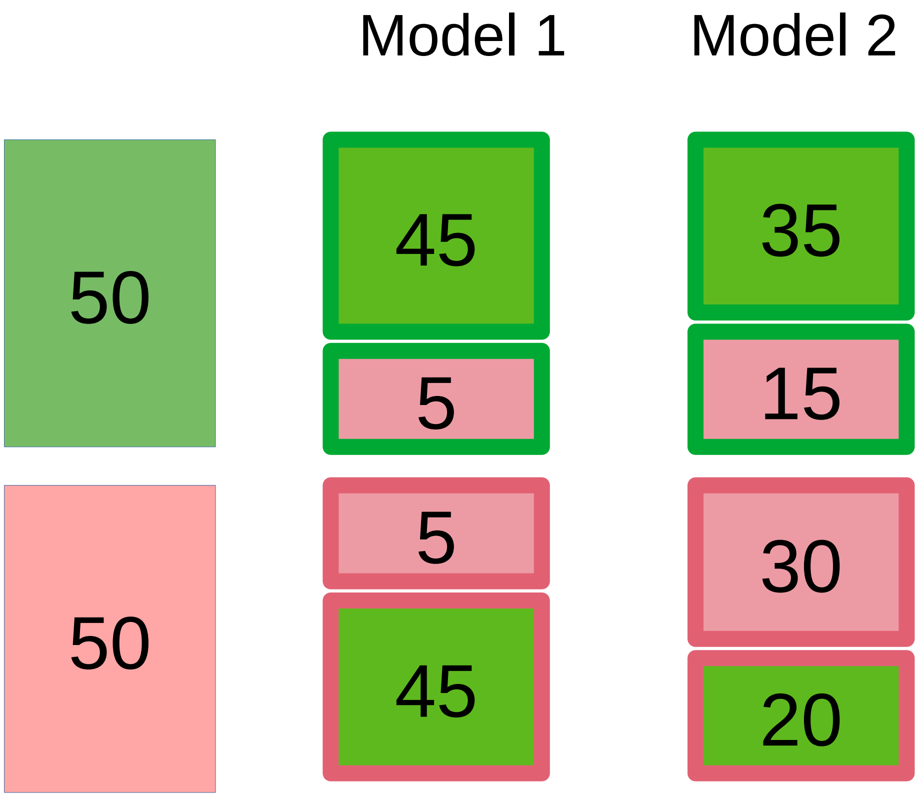
= (TN + TP) / (TN + FP + FN + TP)

- Accuracy alone doesn’t tell the full story when you’re working with a class-imbalanced data set, where there is a significant disparity between the number of positive and negative labels.

- **Precision** and **Recall** are better metrics for evaluating class-imbalanced problems.

Accuracy Example

- Test Data Size: 100 Example



Model 1

Model 2

50

45

5

5

45

35

15

30

20

50

TP: ?
TN: ?
FP: ?
FN: ?

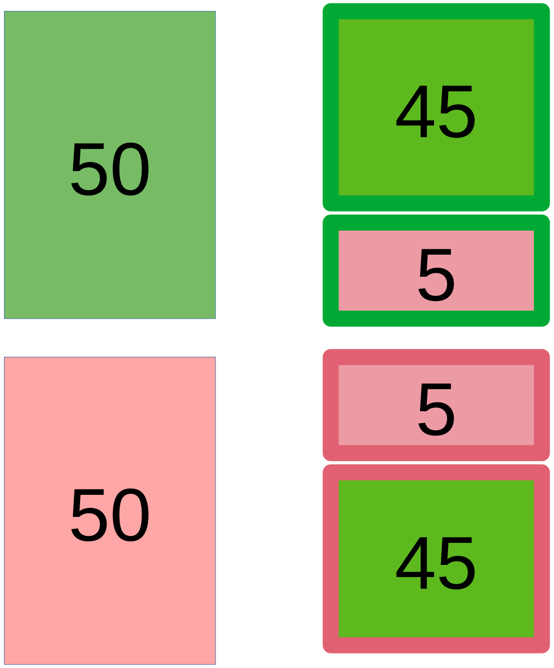
Accuracy:

Best Model?

Accuracy Example

- Test Data Size: 100 Example

Model 1



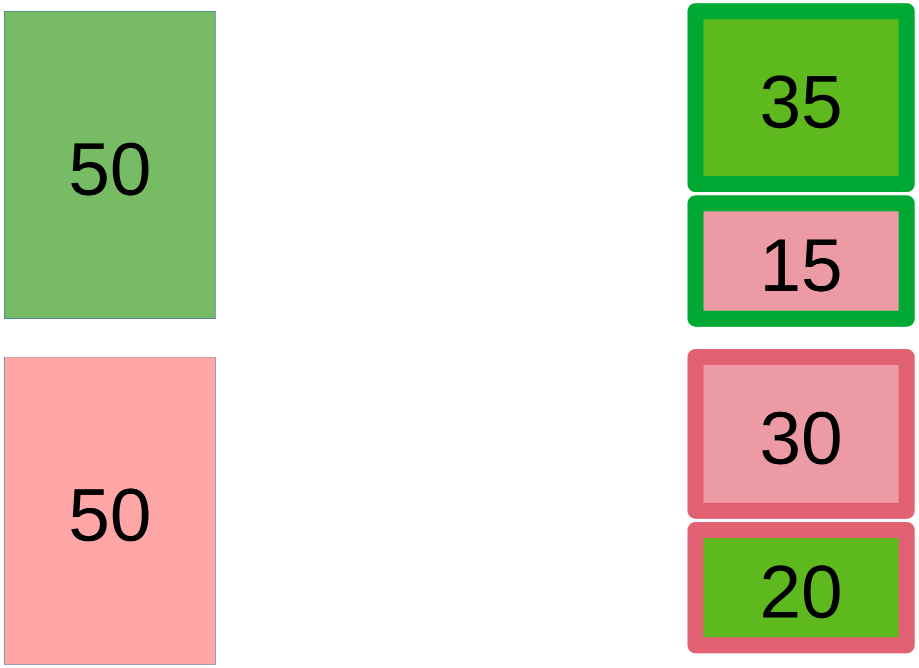
TP: 45
TN: 5
FP: 5
FN: 45

Accuracy: 50%

Accuracy Example

- Test Data Size: 100 Example

Model 2



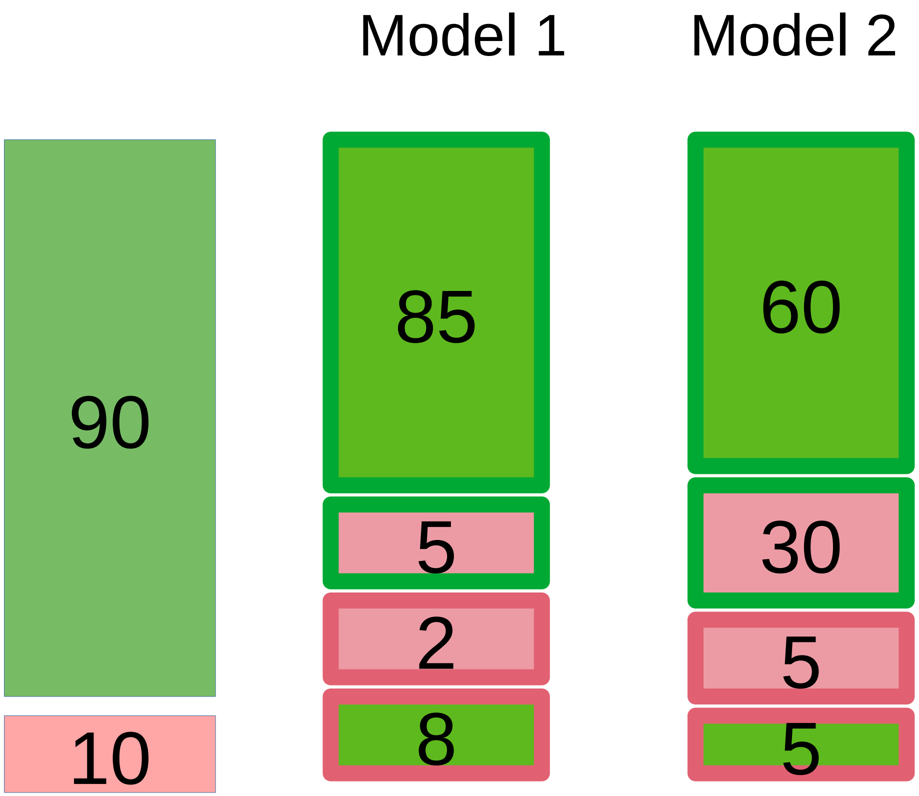
TP: 35
TN: 30
FP: 15
FN: 20

Accuracy: 75%

Best Model?

Accuracy Example

- Test Data Size: 100 Example



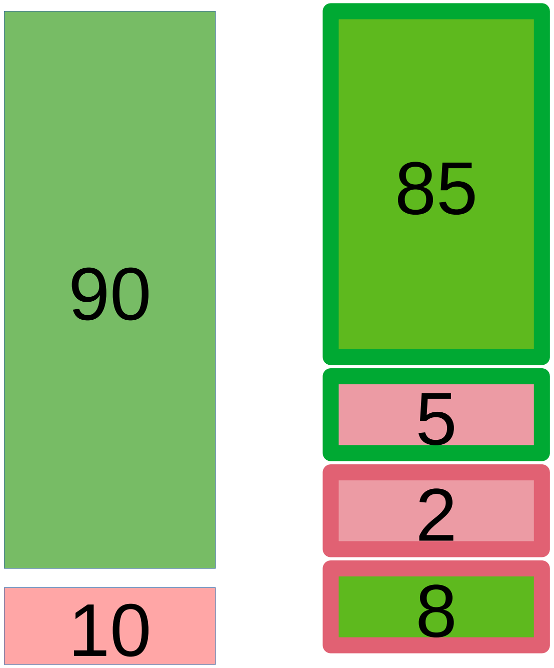
TP: ?
TN: ?
FP: ?
FN: ?

Accuracy:

Accuracy Example

- Test Data Size: 100 Example

Model 1



90

10

85

5

2

8

TP: 85

TN: 2

FP: 5

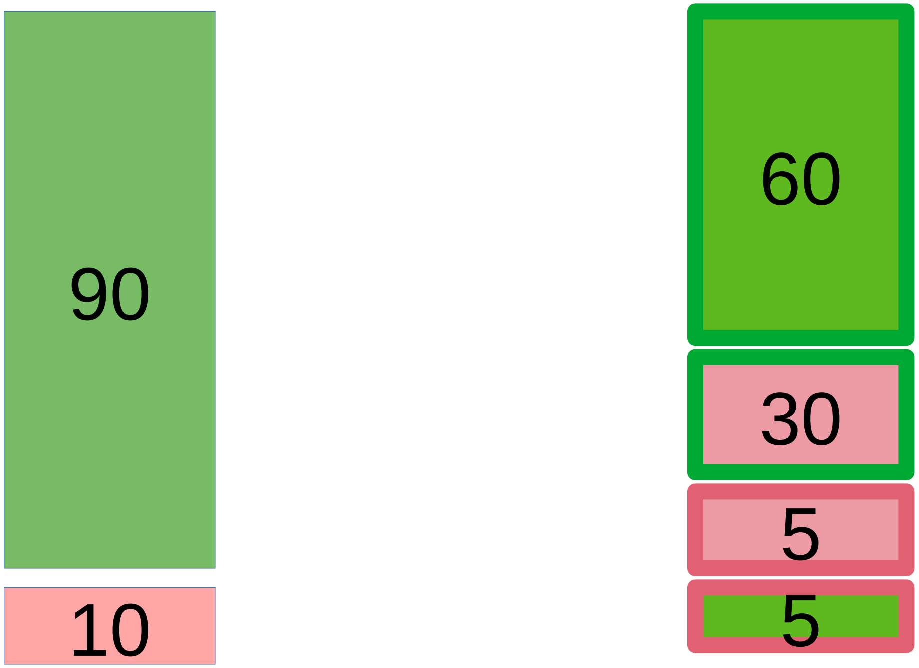
FN: 8

Accuracy:87 %

Accuracy Example

- Test Data Size: 100 Example

Model 2



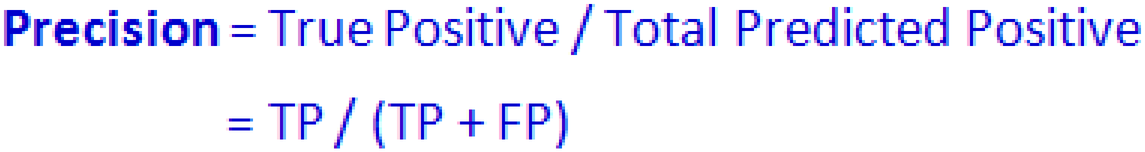
TP: 60
TN: 5
FP: 30
FN: 5

Accuracy: 65 %

Best Model?

Precision

- It is the number of correct positive results divided by the number of positive results predicted by the classifier.
- It is about being precise.
- Precision should be as high as possible.



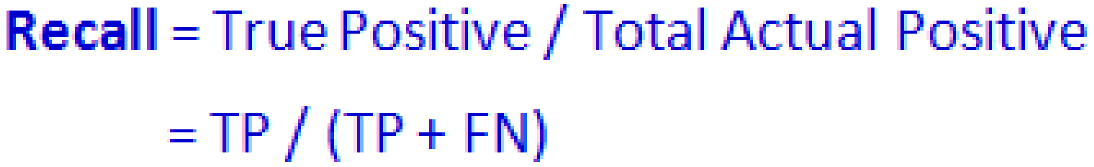
Precision = True Positive / Total Predicted Positive

= TP / (TP + FP)

- Precision is a good measure to determine, when the costs of False Positive is high.
- For instance, email spam detection. In email spam detection, a false positive means that an email that is non-spam (actual negative) has been identified as spam (predicted spam). The email user might lose important emails if the precision is not high for the spam detection model.

Recall

- It is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).
- It is also called **sensitivity** or **true positive rate (TPR)**.
- Recall should be as high as possible.



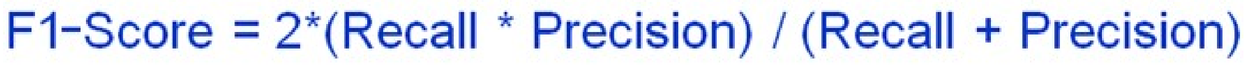
Recall = True Positive / Total Actual Positive

= TP / (TP + FN)

- Recall is a good measure to determine, when the costs of False Negative is high.
- For instance, in fraud detection. If a fraudulent transaction (Actual Positive) is predicted as non-fraudulent (Predicted Negative), the consequence can be very bad for the bank.

F1 - Score

- It is the weighted average (harmonic mean) of Precision and Recall.
- Therefore, this score takes both false positives and false negatives into account.



F1-Score = 2*(Recall * Precision) / (Recall + Precision)

- Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution.
- Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall.

Other Metrics

- **Specificity:** Number of items correctly identified as negative out of total negatives. It is also called true negative rate **(TNR)**.

$$\textbf{Specificity} = TN / (TN + FP)$$

- **False Positive Rate or Type I Error:** Number of items wrongly identified as positive out of total true negatives.

$$\textbf{FPR} = FP / (FP + TN)$$

- **False Negative Rate or Type II Error:** Number of items wrongly identified as negative out of total true positives.

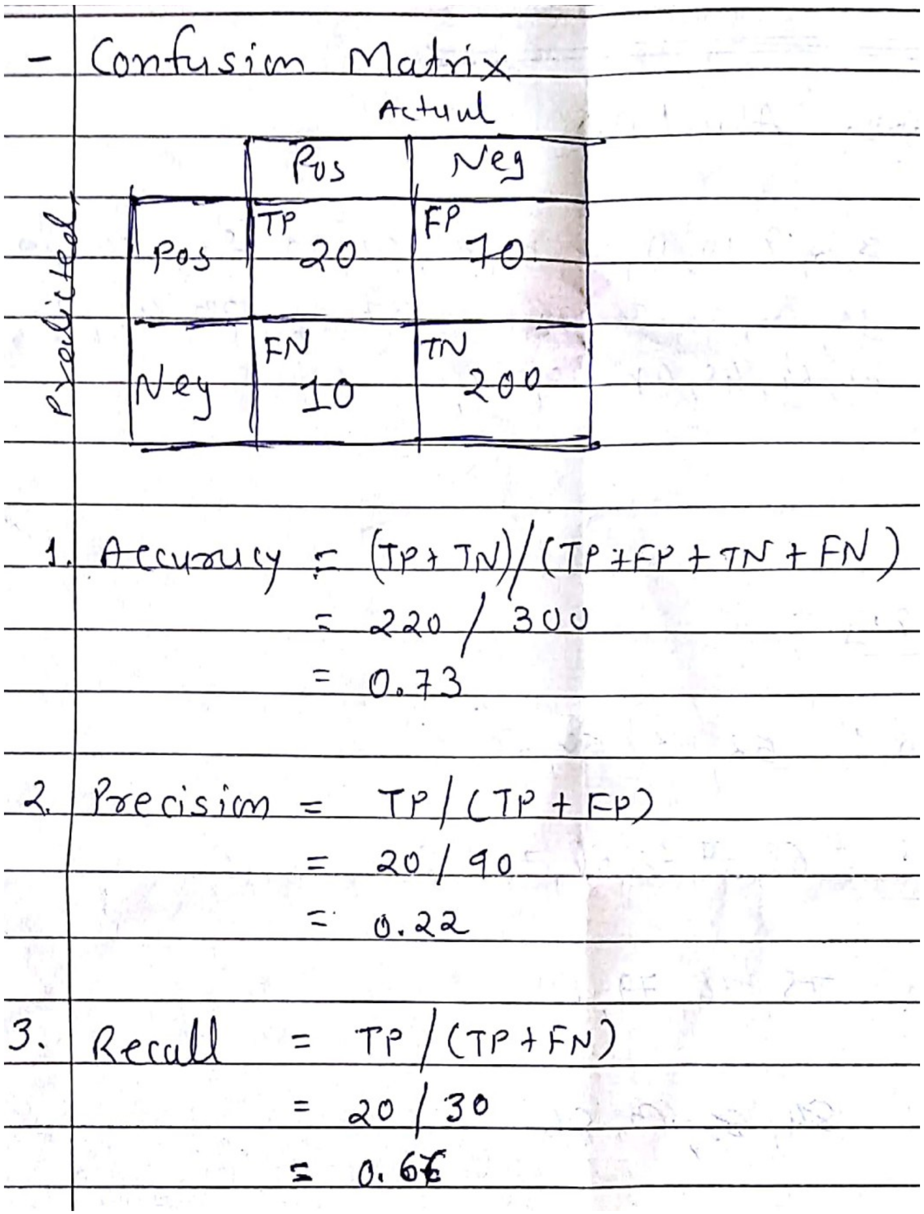
$$\textbf{FNR} = FN / (FN + TP)$$

Example - 1

A cancer prediction system is given 300 data instances to predict whether a patient has a cancer or not. Out of all, actually 30 patients have cancer. The system predicts cancer in 90 cases out of which only 20 predictions are correct .

- Find confusion matrix for the above data.
- Find value for the performance metrics:
  - Accuracy
  - Precision
  - Recall
  - F1-Score
  - Specificity
  - FPR
  - FNR

Example 1 - Solution



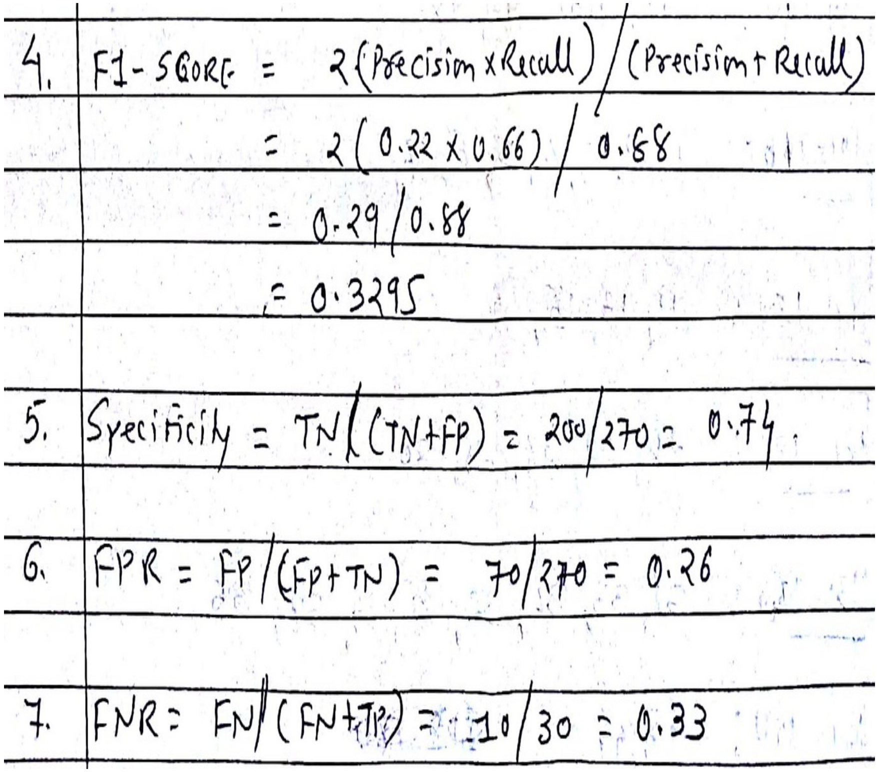
- Confusion Matrix

		Actual	
		Pos	Neg
Predicted	Pos	TP 20	FP 70
	Neg	FN 10	TN 200

1. Accuracy = (TP+TN)/(TP+FP+TN+FN)
= 220 / 300
= 0.73

2. Precision = TP/(TP+FP)
= 20/90
= 0.22

3. Recall = TP/(TP+FN)
= 20/30
= 0.67



4. F1-SCORE = 2(Precision x Recall)/(Precision + Recall)
= 2(0.22 x 0.66) / 0.88
= 0.29/0.88
= 0.3295

5. Specificity = TN/(TN+FP) = 200/270 = 0.74.

6. FPR = FP/(FP+TN) = 70/270 = 0.26

7. FNR = FN/(FN+TP) = 10/30 = 0.33

Example - 2

A cancer prediction system is given 300 data instances to predict whether a patient has a cancer or not. Out of all, actually 30 patients have cancer. The system predicts cancer in all cases.

- Find confusion matrix for the above data.
- Find value for the performance metrics:
  - Accuracy
  - Precision
  - Recall
  - F1-Score
  - Specificity
  - FPR
  - FNR

Mean Absolute Error

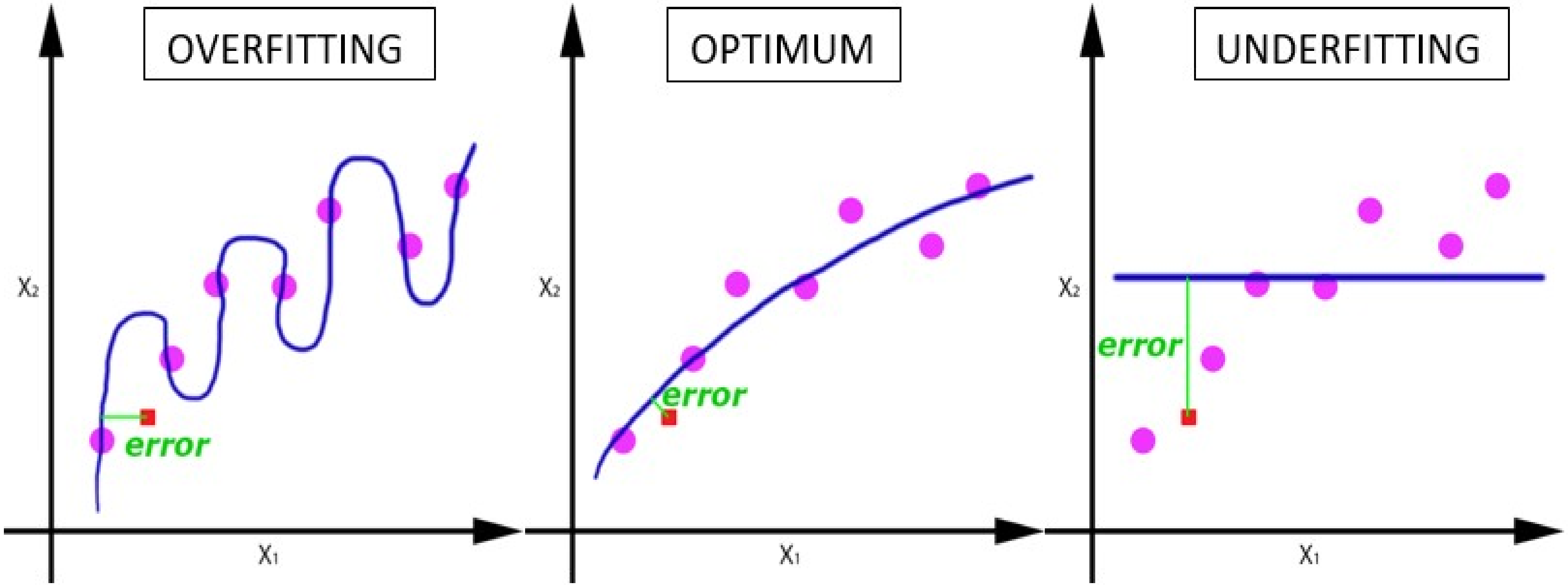
- Mean Absolute Error is the average of the difference between the Original Values and the Predicted Values.
- It gives us the measure of how far the predictions were from the actual output.
- However, they don’t gives us any idea of the direction of the error i.e. whether we are under predicting the data or over predicting the data.
- Mathematically, it is represented as :

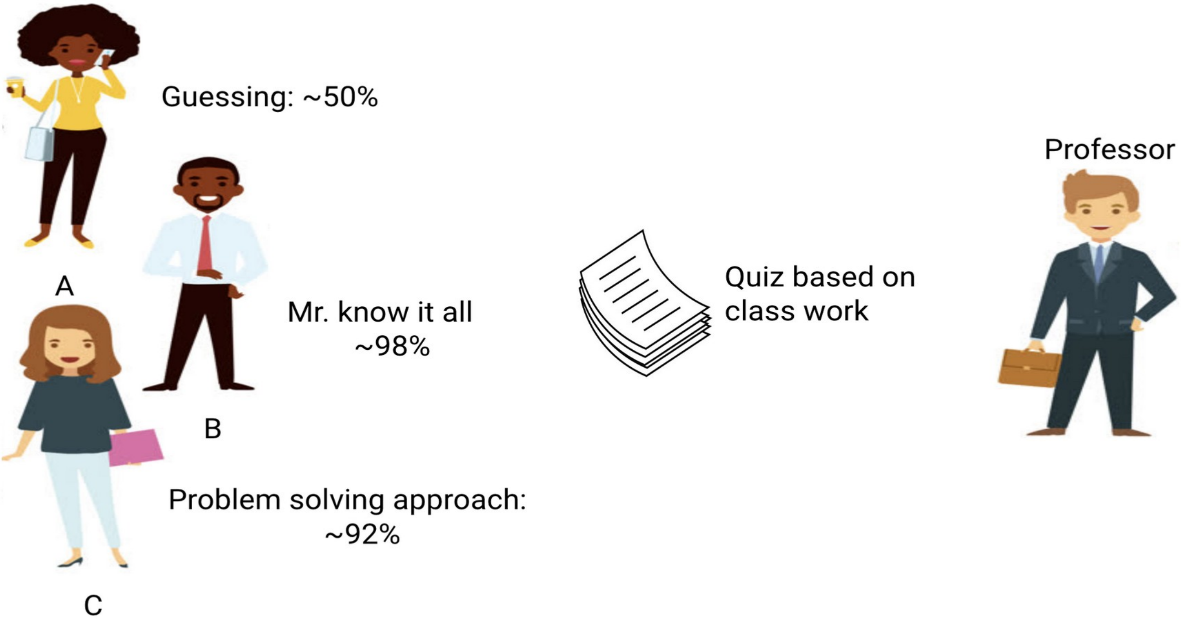
$$MeanAbsoluteError = \frac{1}{N}\sum_{j=1}^{N}|y_j - \hat{y}_j|$$

Mean Squared Error (MSE)

- Mean Squared Error(MSE) is quite similar to Mean Absolute Error, the only difference being that MSE takes the average of the square of the difference between the original values and the predicted values.
- The advantage of MSE being that it is easier to compute the gradient, whereas Mean Absolute Error requires complicated linear programming tools to compute the gradient.
- As, we take square of the error, the effect of larger errors become more pronounced then smaller error, hence the model can now focus more on the larger errors.
- Mathematically, it is represented as :

$$MeanSquaredError = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$$





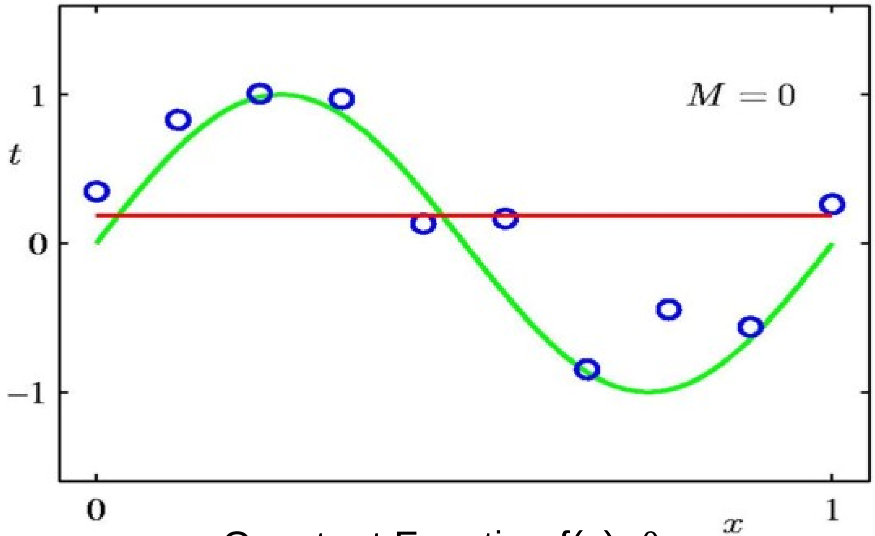
Guessing: ~50%

Mr. know it all
~98%

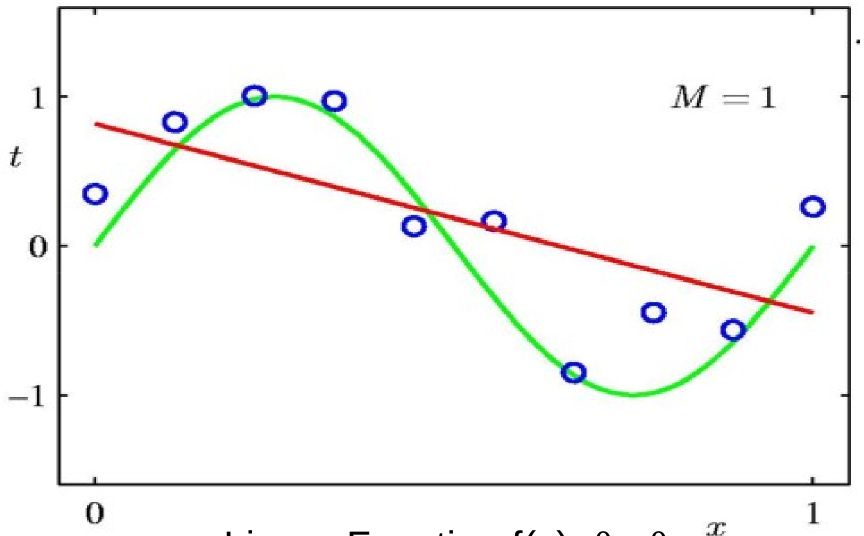
Problem solving approach:
~92%

Quiz based on
class work

Professor

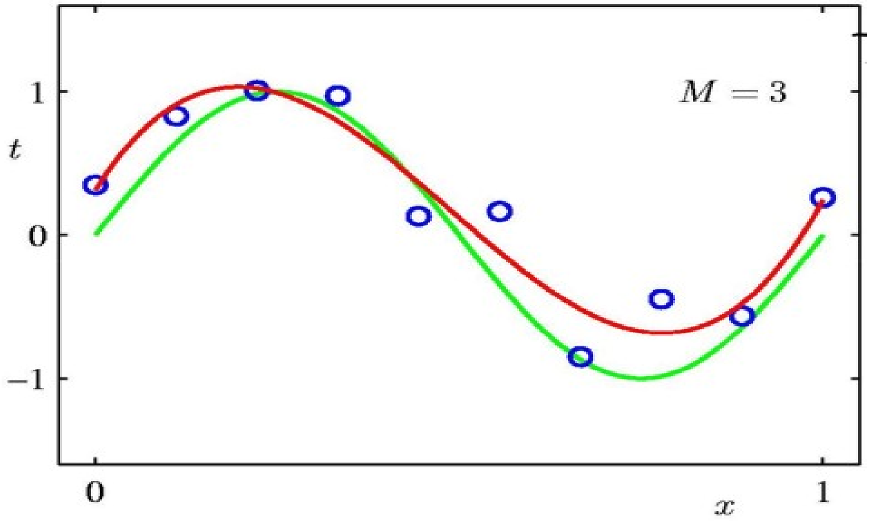


Constant Function $f(x)=\theta_0$



Linear Function $f(x)=\theta_0+\theta_1 x_1$

Cubic Function $f(x)=\theta_0+\theta_1 x_1+\theta_2 x_1^2+\theta_3 x_1^3$



9th Degree Function $f(x)=\theta_0+\theta_1 x_1+ \ldots. +\theta_9 x_1^9$