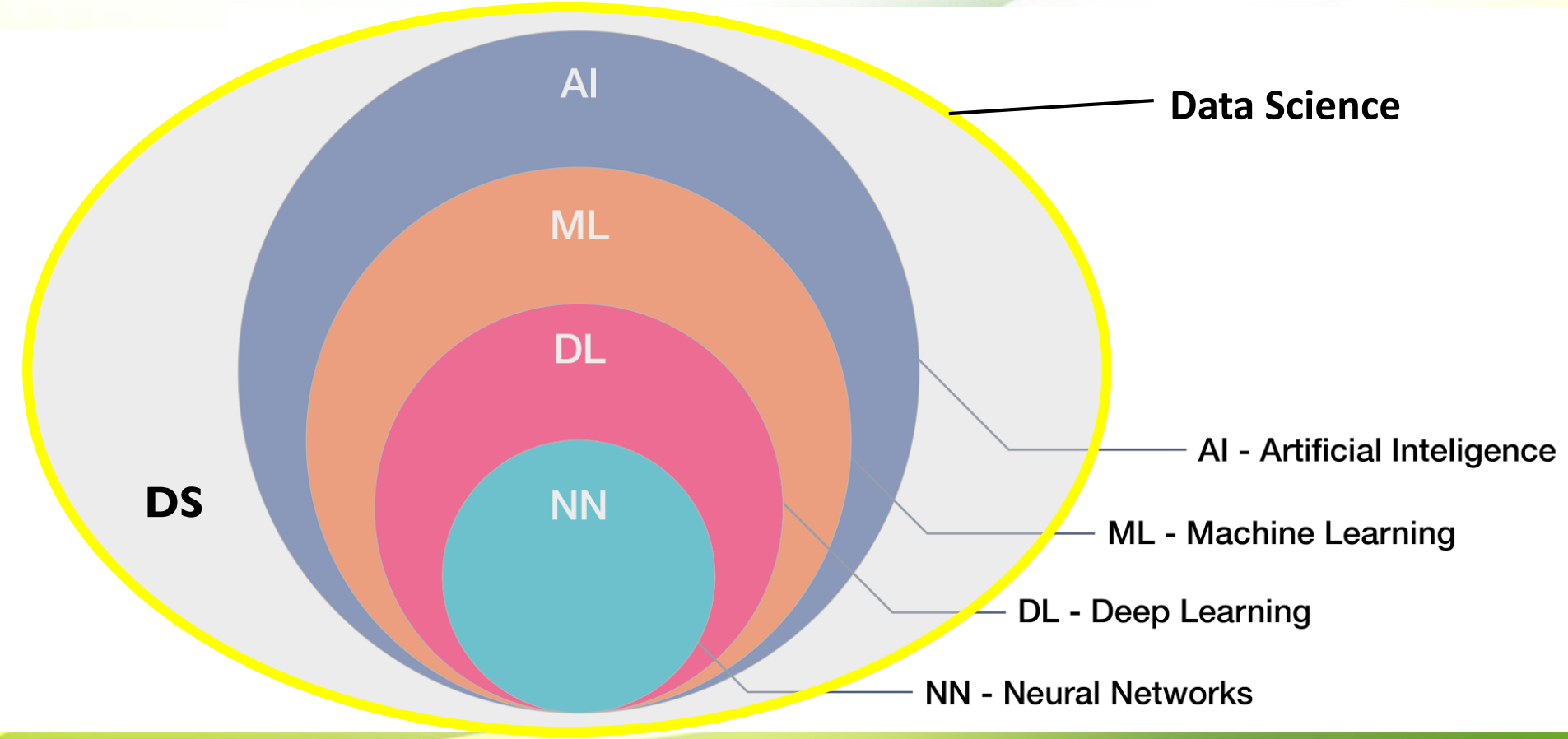
INTRODUCTION TO MACHINE LEARNING

3rd Sem, MCA

Contents

- Introduction to Machine Learning,
- Types of Learning
- Model Evaluation
- Basic Concepts in Machine Learning
  - Feature Vector
  - Hypothesis Space
  - Bias & Variance
  - Inductive Learning
  - Generalization
  - Parametric vs. Non Parametric Models
  - Dimensionality
  - Model Selection
  - No free lunch theories
- Probability Theory & Distributions;
  - Brief review of Probability Theory,
  - Common Discrete Distributions;
  - Common Continuous Distributions,
  - Joint Probability Distributions;
- Transformation of random variables,
- Monte Carlo Approximation.

Data Science



Data Science

- **Data is EVERYTHING;** a new form of revenue.
- Data gives better business insights; helps to uncover *(hidden)* patterns in data.
- Example:
  - One guy order for a computer. He also purchases a mouse and keyboard.
  - Data modeling will build this pattern.
  - Companies use this patter/relation for better business policy. Predictive analytics for future buying.
  - Companies use DS to build recommendation engines. Prescriptive analytics in DS.
- Various algorithms could be applied on data to get more accurate results.
- Running these algorithms on huge datasets needs AI, ML, DL.
- ML is used in DS to make predictions by discovering hidden patterns in data.

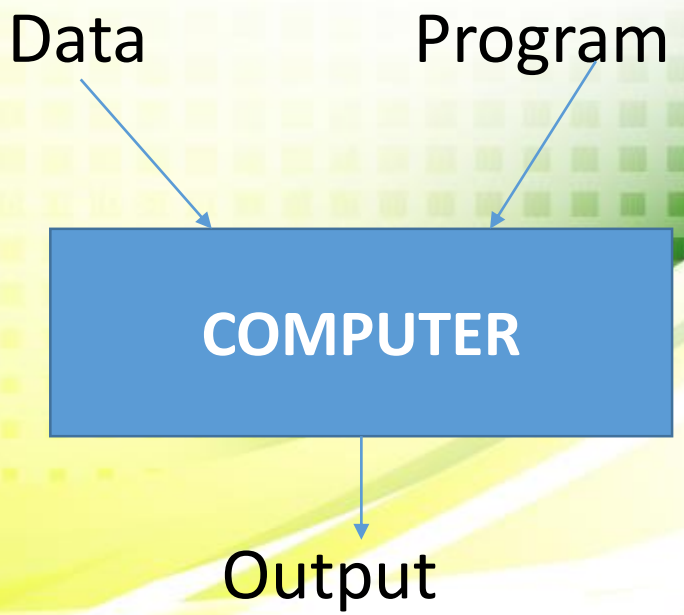
Machine Learning

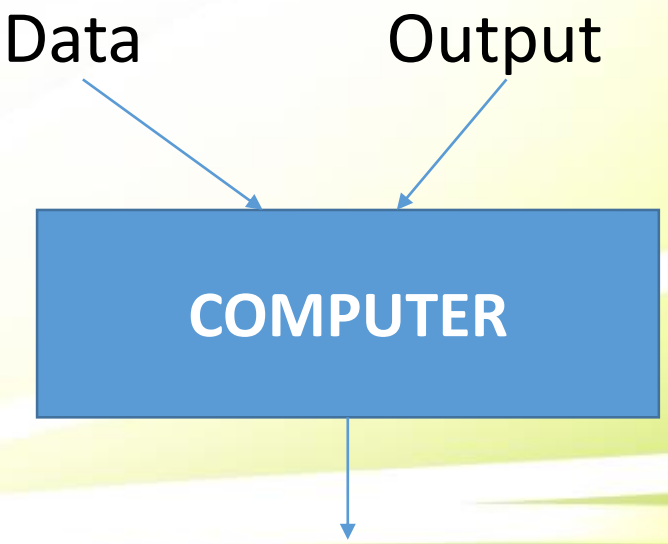
- Machine's ability to learn.
- ML is an implementation of AI.
- Establish Relationship between independent & dependent variables present in data.
- ML is used in situations where machine should learn from huge amounts of data given to it ***(training dataset)***, and then apply that knowledge on new pieces of data that streams into system.
- After learning/training phase is complete (with training data); ML model is tested on data which machine never encountered before ***(testing dataset).***
- Statistical tool to analyze data to get conclusive knowledge.

Machine Learning

- Field of study that gives computers the **ability to learn without being explicitly programmed**.
- Subfield of artificial intelligence → capability of machine to imitate intelligent human behavior.
    - Artificial intelligence systems are used to perform complex tasks in a way that is similar to how humans solve problems.
- Use and development of computer systems that are able to learn and adapt *without following explicit instructions*, ***by using algorithms and statistical models to analyse and draw inferences from patterns in data.***

Programmatic Vs. Machine Learning Solution

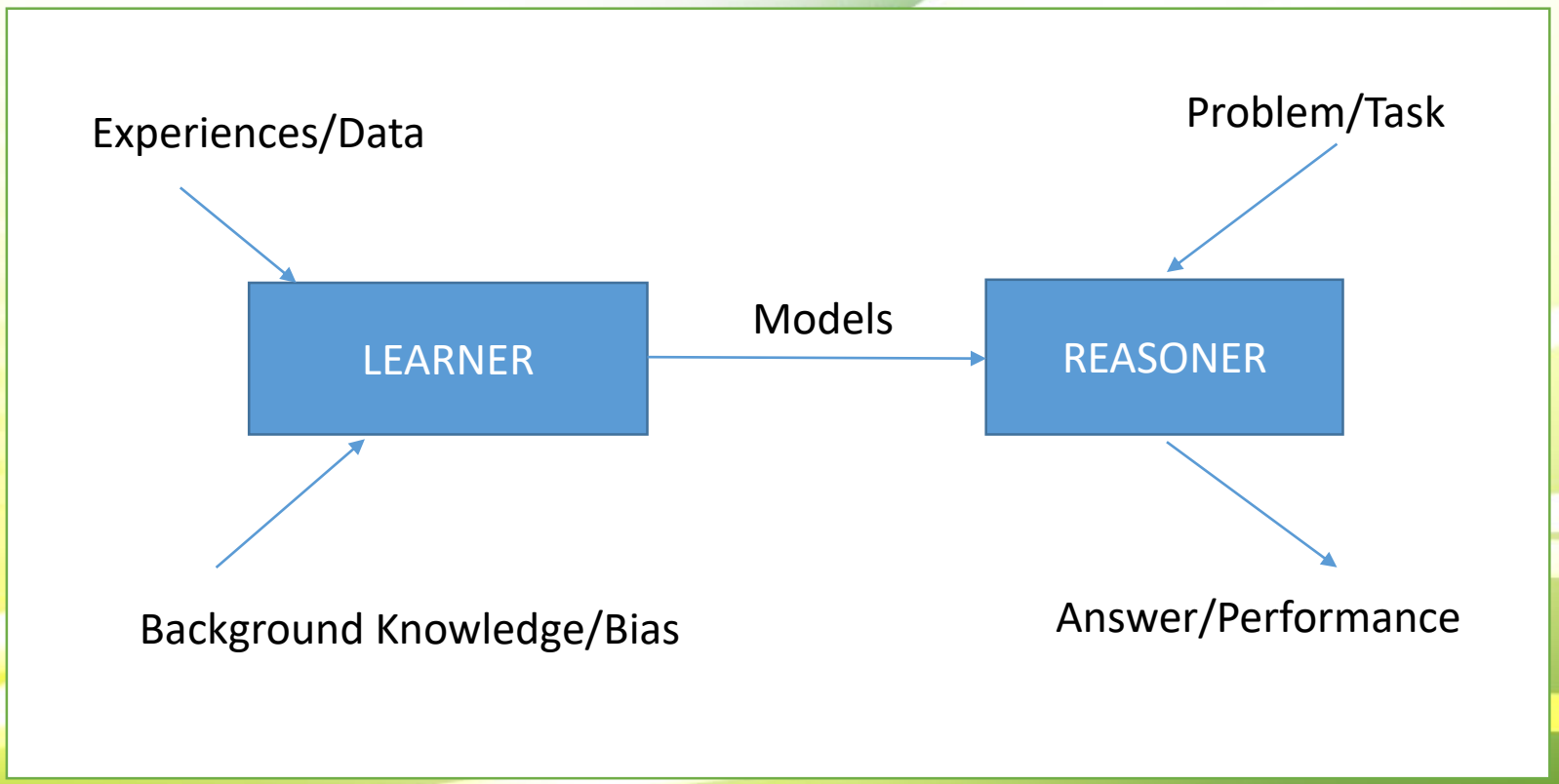




Program or Learning

to solve <u>SUBSEQUENT TASKS</u>

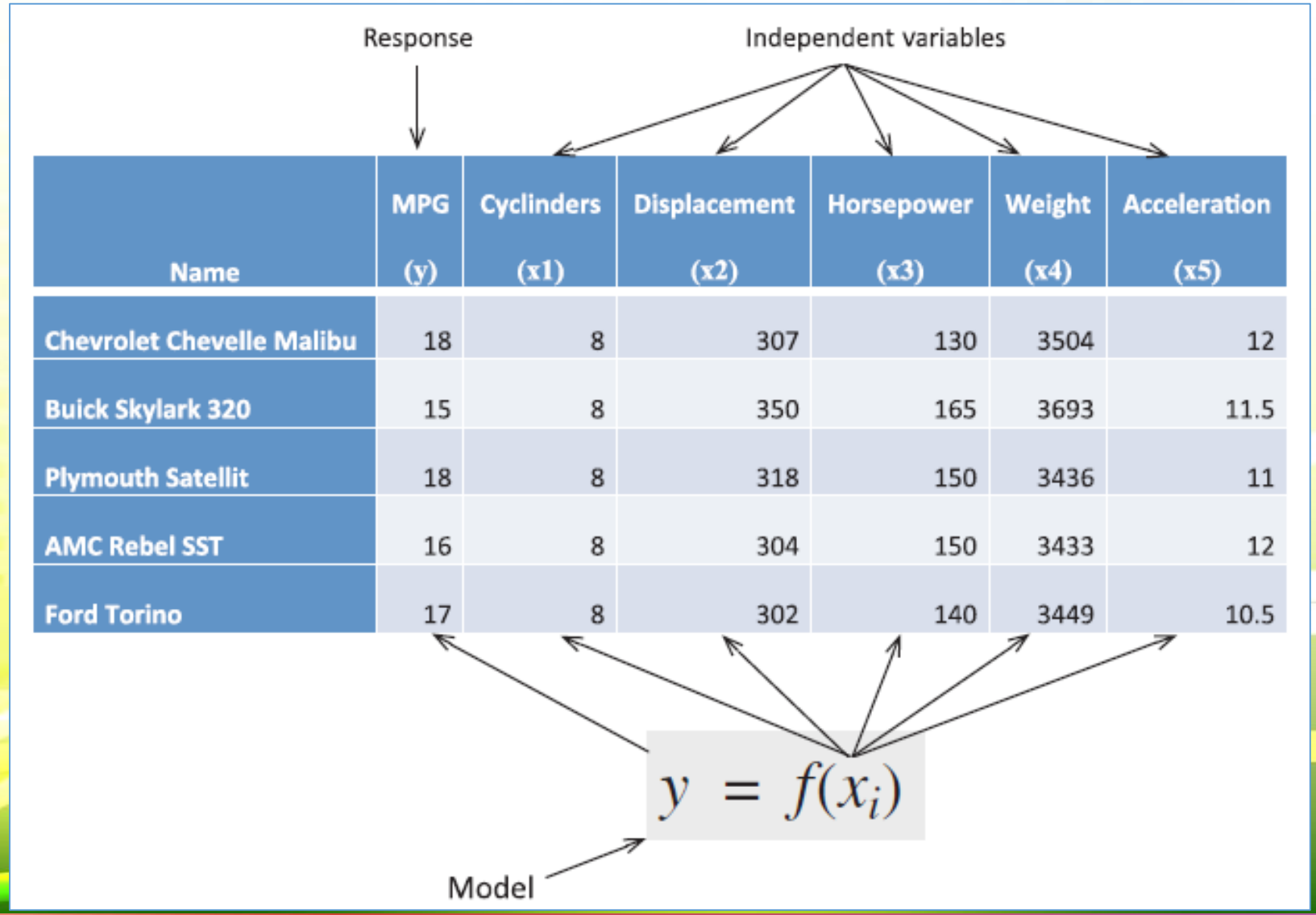
Machine Learning



PREDICTIVE MODELING

- **Predictive modeling** is a commonly used statistical technique to predict future behavior.
- Technology that analyzes historical/current data and generate model to help predict future outcomes.
- Data is collected, statistical model is formulated, predictions are made, and model is validated (or revised) as additional data becomes available.
- Various algorithms in ML used for *prediction problems, classification problems, regression problems,* etc.
- Predictive analytics models are:
  - *Classification model*
  - *Clustering model*
  - *Forecast model*
  - *Outliers model*
  - *Time series model*

PREDICTIVE MODELING

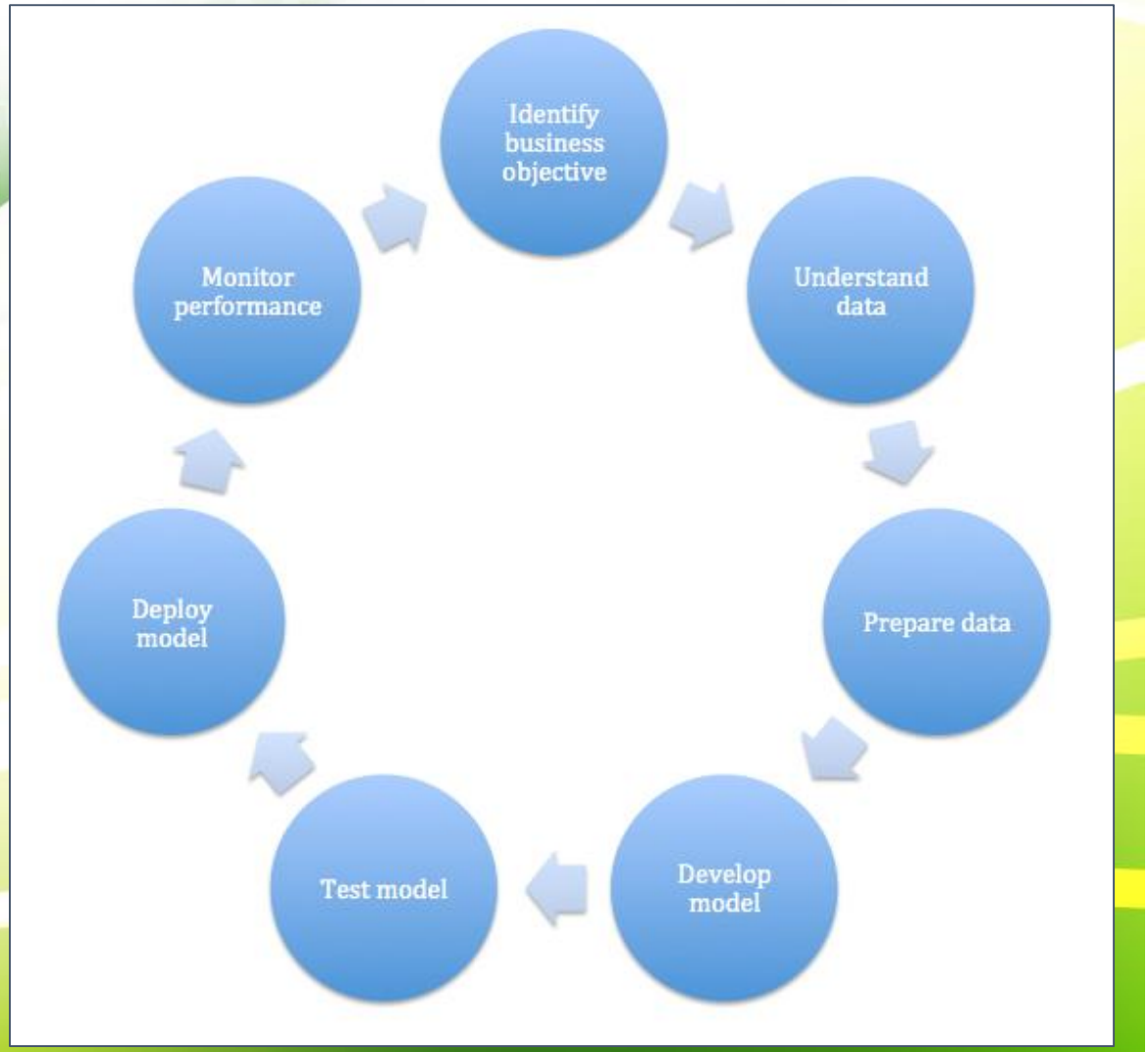


Name	MPG (y)	Cyclinders (x1)	Displacement (x2)	Horsepower (x3)	Weight (x4)	Acceleration (x5)
Chevrolet Chevelle Malibu	18	8	307	130	3504	12
Buick Skylark 320	15	8	350	165	3693	11.5
Plymouth Satellit	18	8	318	150	3436	11
AMC Rebel SST	16	8	304	150	3433	12
Ford Torino	17	8	302	140	3449	10.5

PREDICTIVE MODELING

Phases of Predictive Modeling

- *Understand Business Objective*
- *Define Modelling Goals*
- *Select/Collect Data*
- *Prepare Data*
- *Analyse and Transform Variables,* ***Sampling***
- ***Model*** *Selection, Develop Models (Training)*
- *Validate Models (Testing), Optimize, Profitability*
- *Document Methodology and Models*
- *Implement Models*
- *Monitoring and Performance Tracking*



PREDICTIVE MODELING

Predictive modeling limitations :

- Errors in data labeling.
- Shortage of massive data sets needed to train machine learning.
- Machine's inability to explain what and why it did what it did.
- Generalizability of learning, or rather lack thereof.
- Bias in data and algorithms.

Types of Learning

Supervised learning

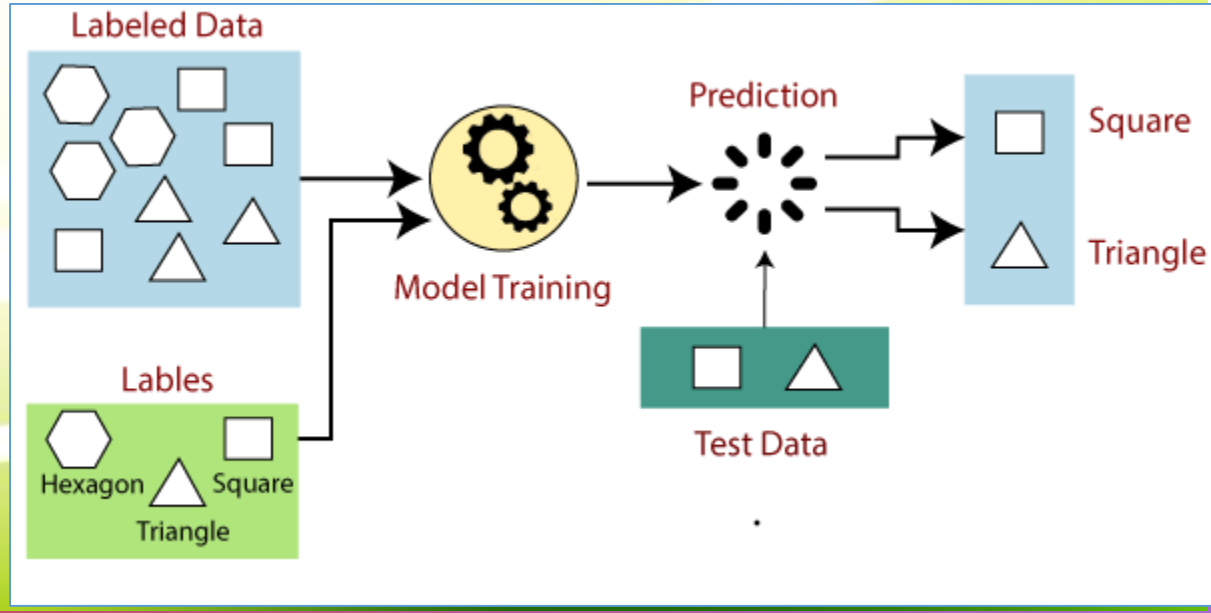
- predict y from x
- Given a labelled set of input-output pairs, Map input x to output y

  Given: Training set $\{(x_i, y_i) \mid i = 1 \ldots n\}$

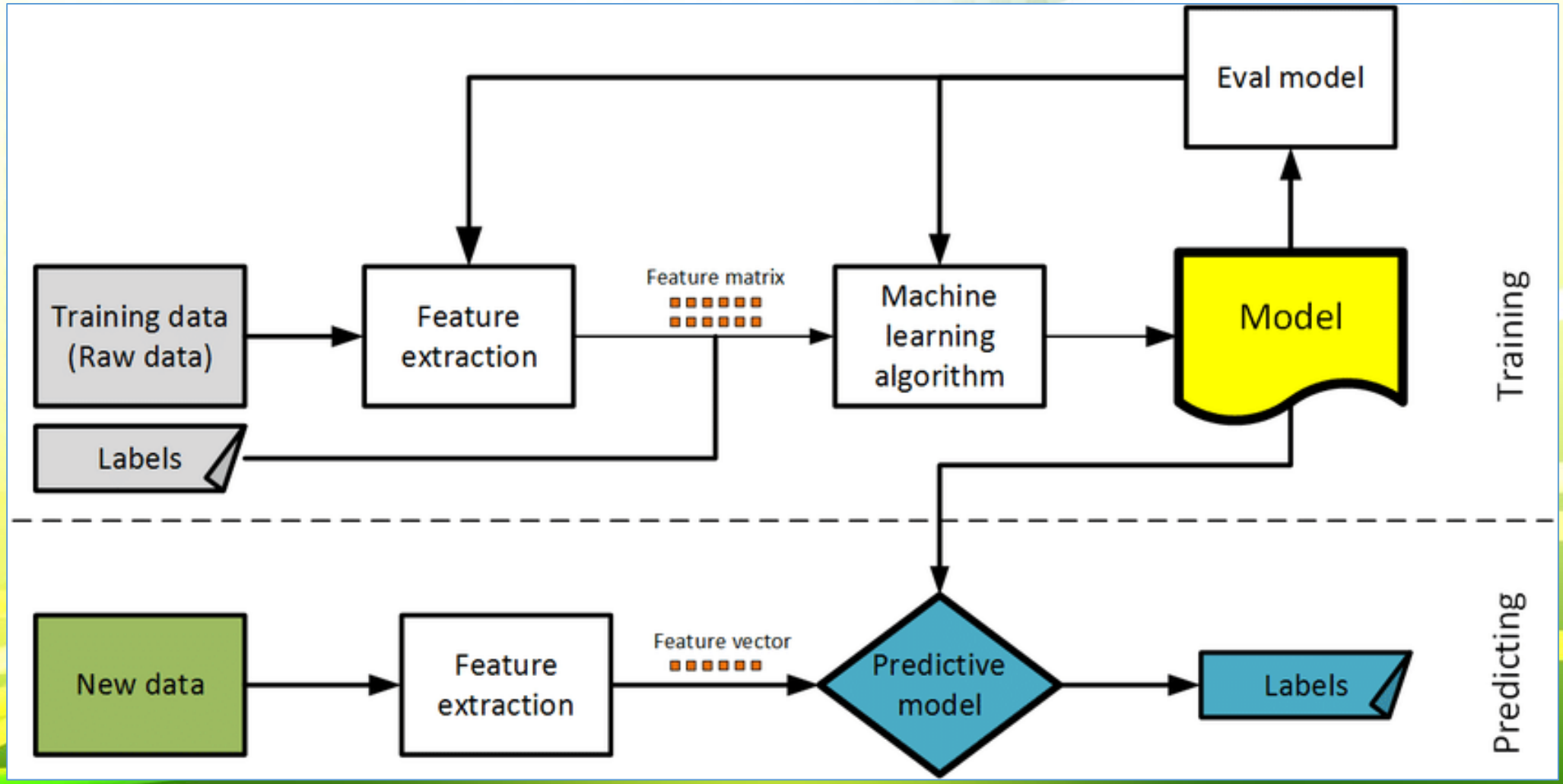
  Find: A good approximation to $f : X \rightarrow Y$
- **Classification** – y is categorical
- **Regression** – y is real values

Examples –

- spam detection, Digit Recognition, stock prices



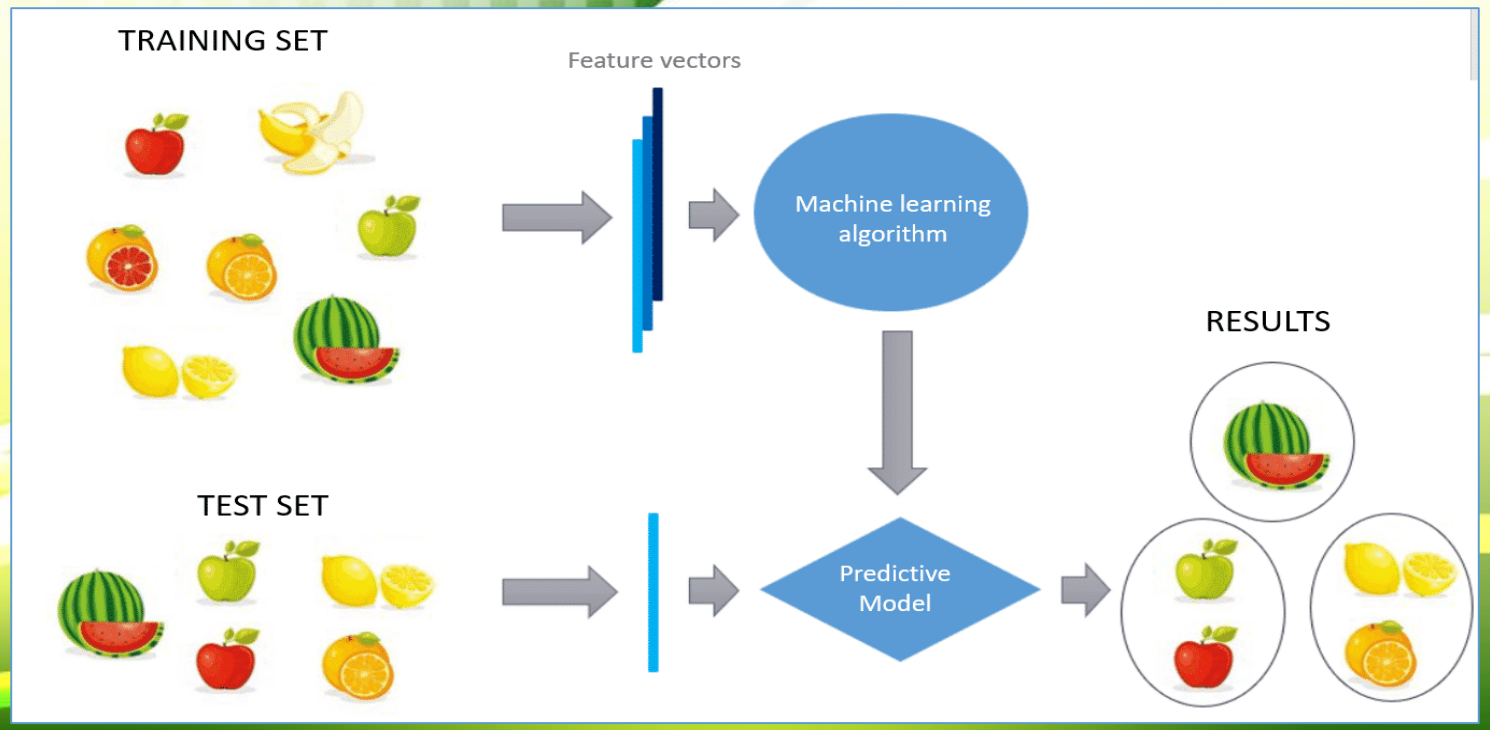
Supervised Learning



Types of Learning

Unsupervised learning

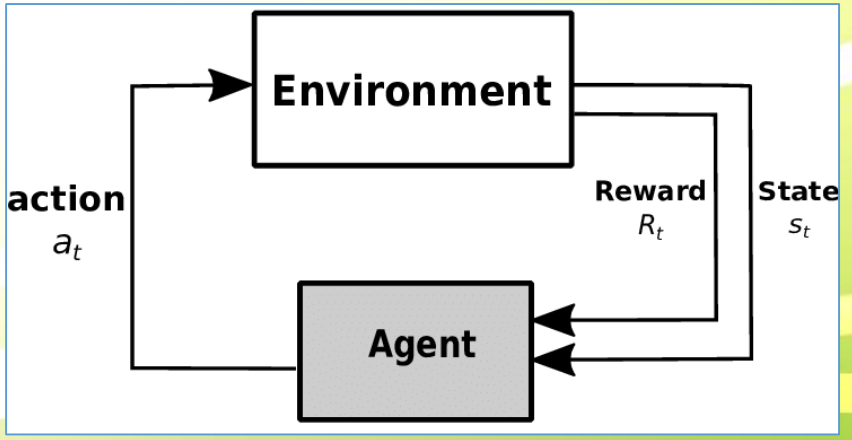
- model p(x), evaluate how likely is to be x
- Find interesting patterns or knowledge discovery
- clustering, association rules/patterns



Types of Learning

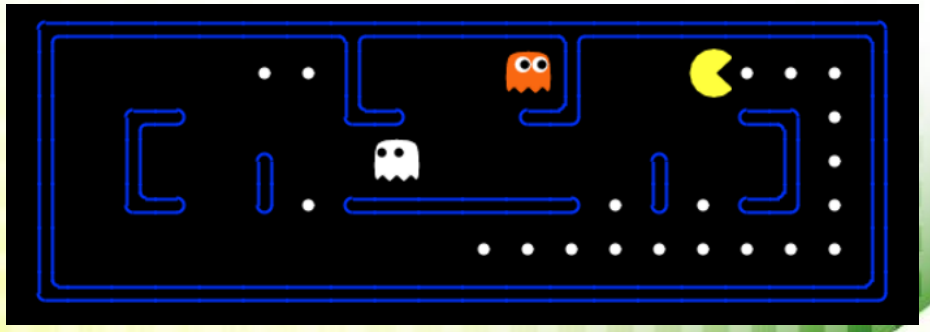
Reinforcement learning

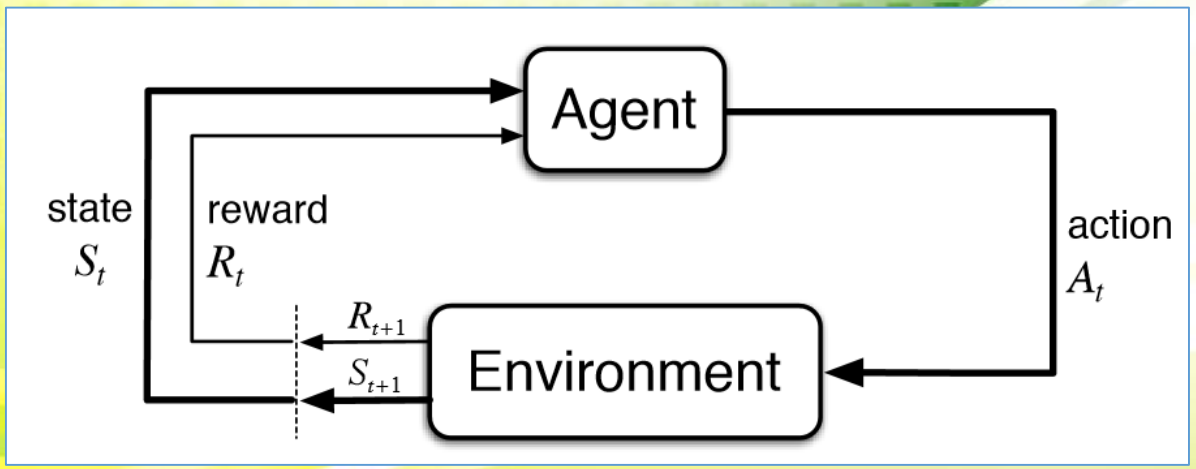
- Agent is acting in an environment.
- Needs to learn action to take at every step
- Action depends on reward or punishment signals that agent gets in each state

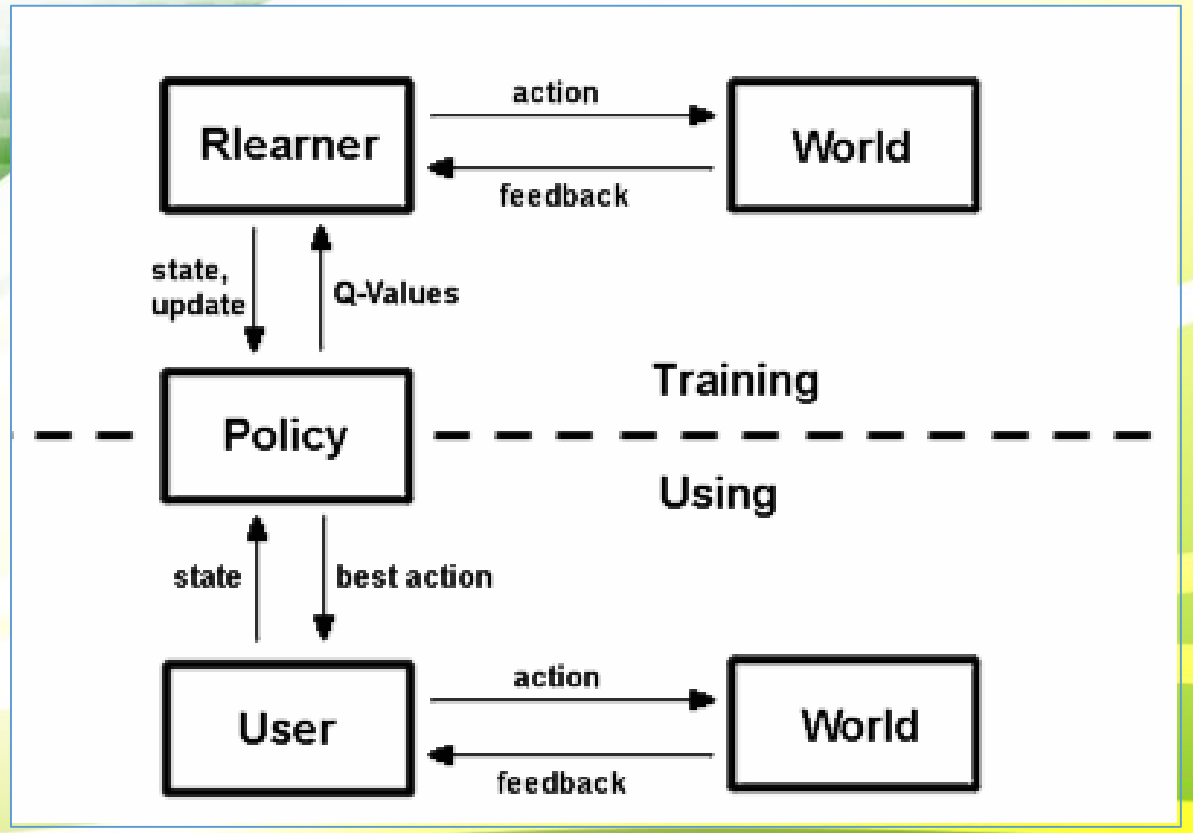


- **Environment —** Physical world in which the agent operates
- **State —** Current situation of the agent
- **Reward —** Feedback from the environment
- **Policy —** Method to map agent's state to actions
- **Value —** Future reward that an agent would receive by taking an action in a particular state

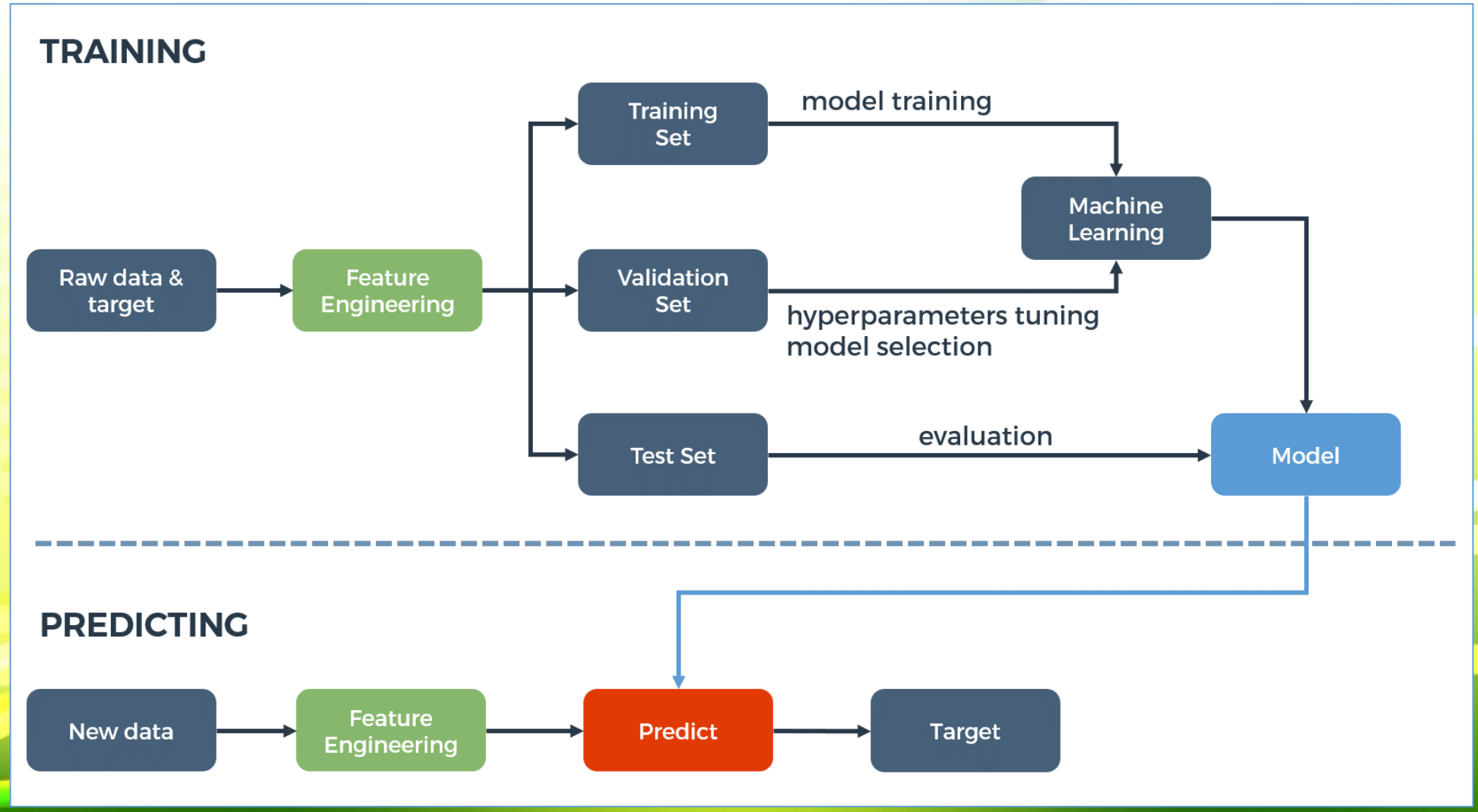
Types of Learning







Machine Learning



Feature Engineering

- Process of **selecting, adding & transforming** variables when creating a predictive model using ML.
  - Good way to enhance predictive models as it involves isolating key information which helps in finding/highlighting patterns.
  - Goal is simplifying & speeding up data transformations while also enhancing model accuracy.

- Construction of additional variables, or features, to dataset to improve machine learning model performance and accuracy.
  - Enables to build more complex models than with only raw data.
  - It also allows to build interpretable models from any amount of data.

- **Feature selection** will help to limit these features to a manageable number.

Feature Engineering

Name	MPG (y)	Cyclinders (x1)	Displacement (x2)	Horsepower (x3)	Weight (x4)	Acceleration (x5)
Chevrolet Chevelle Malibu	18	8	307	130	3504	12
Buick Skylark 320	15	8	350	165	3693	11.5
Plymouth Satellit	18	8	318	150	3436	11
AMC Rebel SST	16	8	304	150	3433	12
Ford Torino	17	8	302	140	3449	10.5

Sq Ft.	Amount
2400	9 Million
3200	15 Million
2500	10 Million
2100	1.5 Million
2500	8.9 Million

Sq Ft.	Amount	Cost Per Sq Ft
2400	9 Million	4150
3200	15 Million	4944
2500	10 Million	3950
2100	1.5 Million	510
2500	8.9 Million	3600

Train, Validation, Test Data

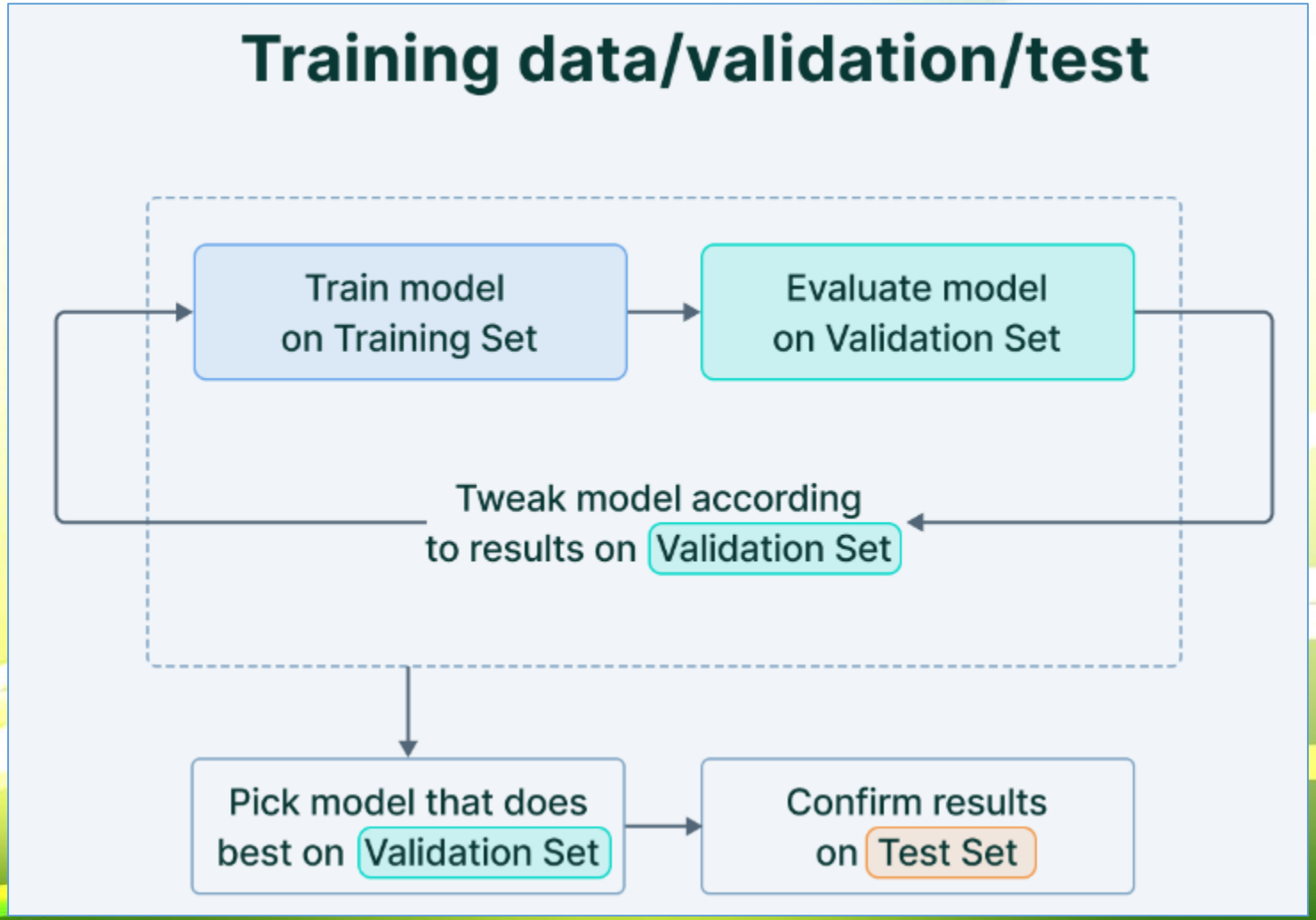
Three split from one large dataset;

- **Training data:** used to train and make the model learn hidden features/patterns in data.

- **Validation data:** provides first test against unseen data, allowing to evaluate how well model makes predictions based on new data.
  - gives information that helps to tune the model's hyperparameters and configurations accordingly.

- **Test data.** test the model after completing the training to evaluate model accuracy in predictions.
  - Training & validation data include labels to monitor performance metrics of model.
  - Testing data should be unlabeled.
  - Provides a final, real-world check of unseen dataset to confirm that model was trained effectively.

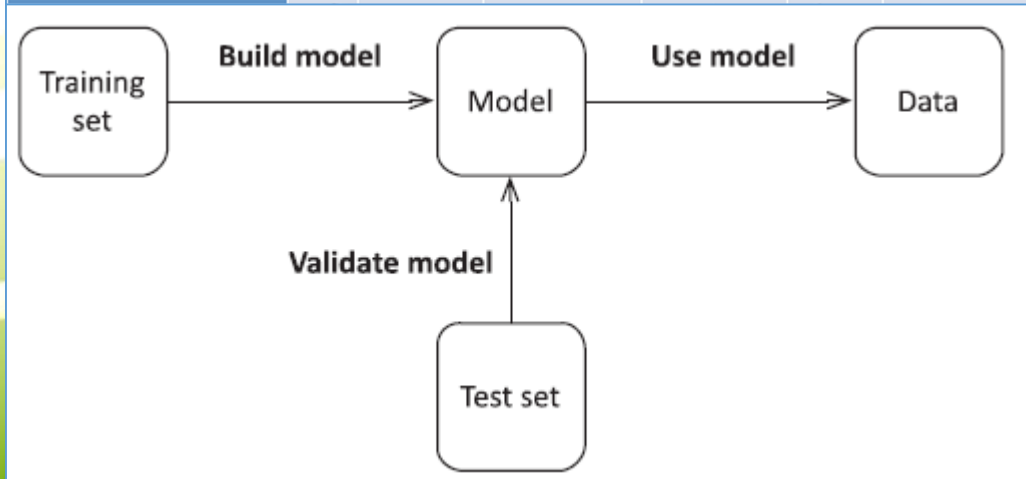
Train, Validation, Test Data



Train, Validation, Test Data

- ***Training** set:* data set used to build a model.
- ***Test set**:* data set used to test the model
- **Cross-validation:** validating model efficiency by training it on subset of input data and testing on previously unseen subset of input data.
- Three steps involved in cross-validation:
  - Reserve some portion of sample data-set.
  - Using the rest data-set train the model.
  - Test the model using the reserve portion of the data-set.
- Methods used for Cross-Validation:
  - Validation Set Approach (50%)
  - Leave-P-out cross-validation (p, n-p)
  - Leave one out cross-validation (1, n-1)
  - K-fold cross-validation
  - Stratified k-fold cross-validation

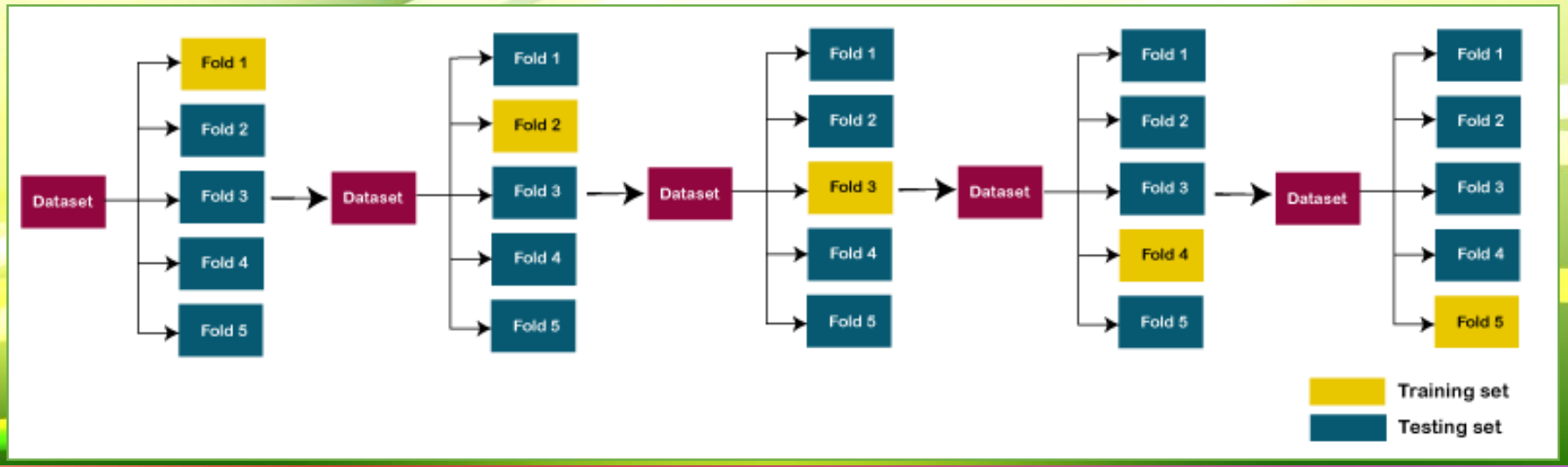
Name	MPG (y)	Cyclinders (x1)	Displacement (x2)	Horsepower (x3)	Weight (x4)	Acceleration (x5)
Chevrolet Chevelle Malibu	18	8	307	130	3504	12
Buick Skylark 320	15	8	350	165	3693	11.5
Plymouth Satellit	18	8	318	150	3436	11
AMC Rebel SST	16	8	304	150	3433	12
Ford Torino	17	8	302	140	3449	10.5



Train, Validation, Test Data

K-fold Cross-validation:

1. Split input dataset into K groups/folds (equal size)
2. Repeat this step by each group on rotation:
   - Take one group as reserve or test dataset.
   - Use remaining groups as training dataset.
   - Fit the model on training set and evaluate performance of the model using the test set.
3. Accuracy of the model is based on average of the k scores.

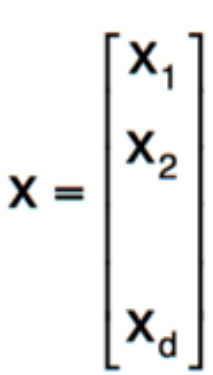


Terminology – Feature Vector

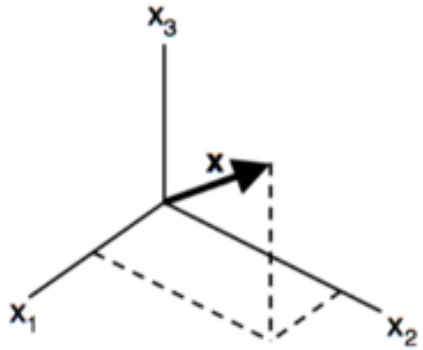
- **Feature**: An individually measurable property of the phenomenon being observed.
  - list of numbers eg: age, name, height, weight etc.
- **Feature Space:** collection of n-features related to some properties of object or event under study.
- **Feature vector:** *n-dimensional vector;* ordered list of numerical properties of observed phenomena.
  - Each row is a feature vector.

ID	First Name	Last Name	Email	Year of Birth
1	Peter	Lee	plee@university.edu	1992
2	Jonathan	Edwards	jedwards@university.edu	1994
3	Marilyn	Johnson	mjohnson@university.edu	1993
6	Joe	Kim	jkim@university.edu	1992
12	Haley	Martinez	hmartinez@university.edu	1993
14	John	Mfume	jmfume@university.edu	1991
15	David	Letty	dletty@university.edu	1995

Feature Vector



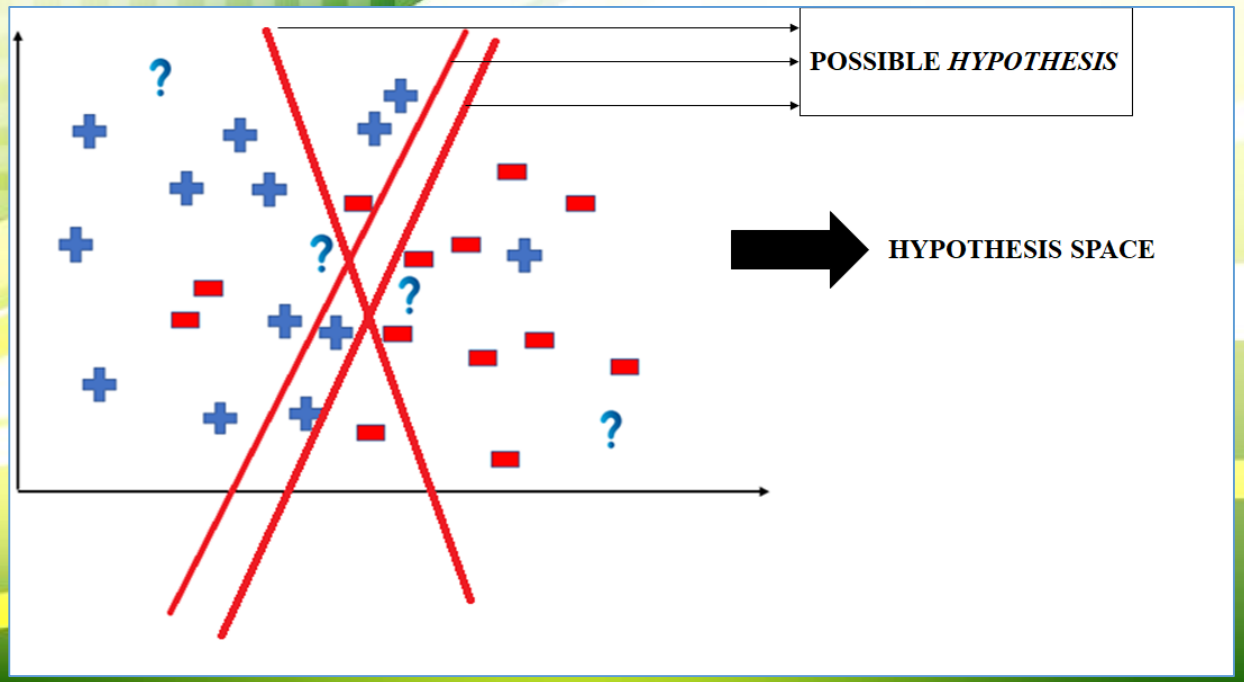
Feature vector



Feature space (3D)

Terminology – Hypothesis Space

- **Hypothesis**: A function that best describes the target in the data (separates classes in clustering).
- **Hypothesis Space:** set of all legal possible functions/hypothesis that are solutions to the task.
  - Comprises of the features chosen and the language or the class of functions
  - ML model determines best possible (only one from this set) which would best describe the target function or the outputs.



Terminology - Inductive Bias

- Set of assumptions that learner uses to predict outputs of given inputs that it has not encountered before.
  - k-Nearest Neighbors (k-NN) algorithm assumes that similar data points are clustered near each other away from the dissimilar ones.
  - Given (X, Y) data points, linear regression assumes that variable (Y) is linearly dependent on explanatory variables (X).
  - Logistic regression assumes that there's a hyperplane that separates two classes from each other.
- Relational inductive biases represents the relationship between entities in the network, while non-relational inductive biases is a set of techniques that further constrain the learning algorithm.

Terminology - Inductive Learning

Inductive/discovery learning/reasoning

- generalization of specific facts
- involves creation of a generalized rule by observing data given to algorithm.
- Can be very complex depending on data; but Effective method in crucial situations.
- uses a bottom-up approach.

a. Apple is a fruit.
b. Apple tastes sweet.
Conclusion: All fruits taste sweet.

deductive learning/reasoning

- uses already available facts and information in order to give a valid conclusion.
- results are certain (not a generalized rule)
- Uses top-down approach

a. All carnivores eat meat.
b. Lion is a carnivore.
Conclusion: – Lion eats meat.

TERMINOLOGY – BIAS & VARIANCE

- In statistics a **fit** is, **how close model is to target class/function/value**.
- Key components in ML modelling:
  - **Signal:** true underlying pattern of data that helps ML model to learn from data.
  - **Noise:** unnecessary and irrelevant data that reduces performance of model.
  - **Bias:** measure of model accuracy.
    - difference between predicted values and actual values.
    - prediction error that is introduced in model due to oversimplifying/optimizing ML algorithms.
  - **Variance:** If ML model performs well (low error) with training dataset, but does not perform well (high error) with test dataset.

Terminology – Bias & Variance

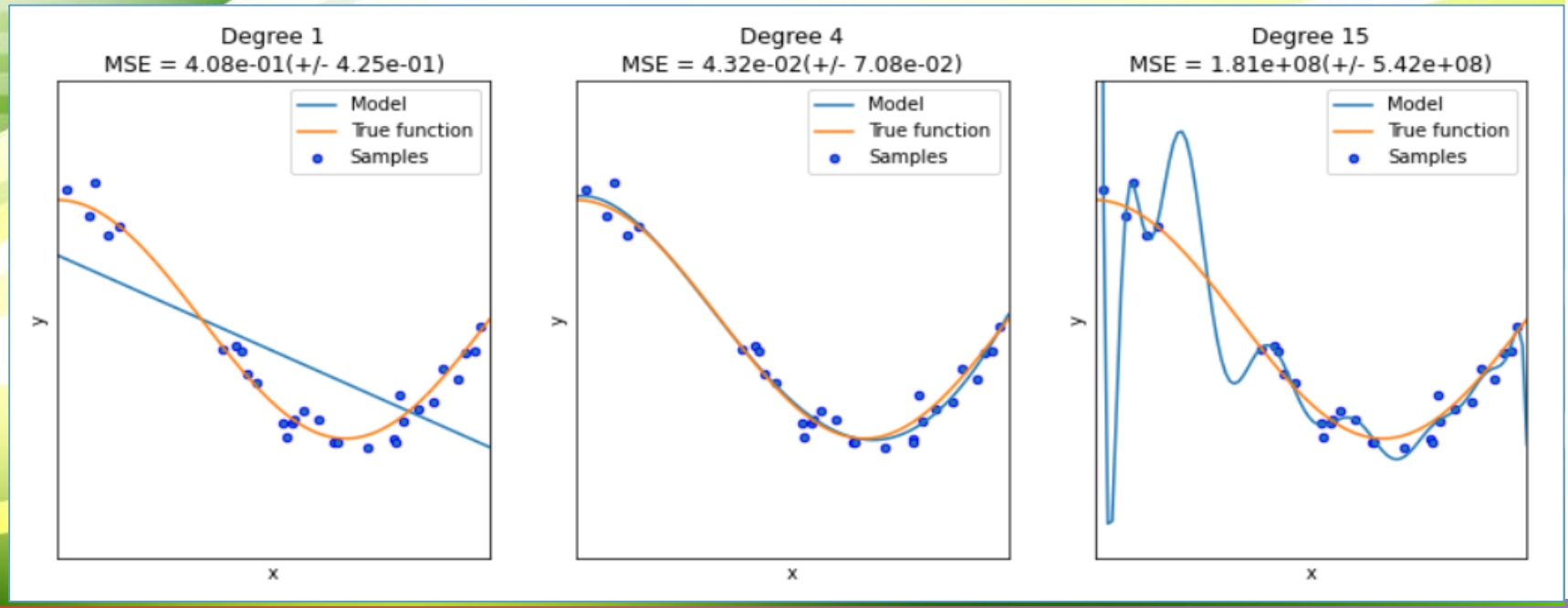
- **Bias** → Average squared difference between predictions and true values.
- Measure of how good model fits the data.

$$Y \approx W_0 + W_1 X_1 + W_2 X_2 + \cdots + W_P X_P$$

- Zero bias → model captures true data generating process perfectly.
- Generally, Both training and validation loss should go to zero.
  - Unrealistic.
  - data is almost always noisy in reality → so some bias is inevitable → **irreducible error.**
- If losses do not decrease as expected, it probably signals that model is not a good fit for data.
- **Variance** → if predictions are sensitive to small changes in (unseen) input.
- High variance often means *overfitting* because model seems to have captured random noise or outliers.
- Overfitting and underfitting are two biggest causes of poor performance of ML models.

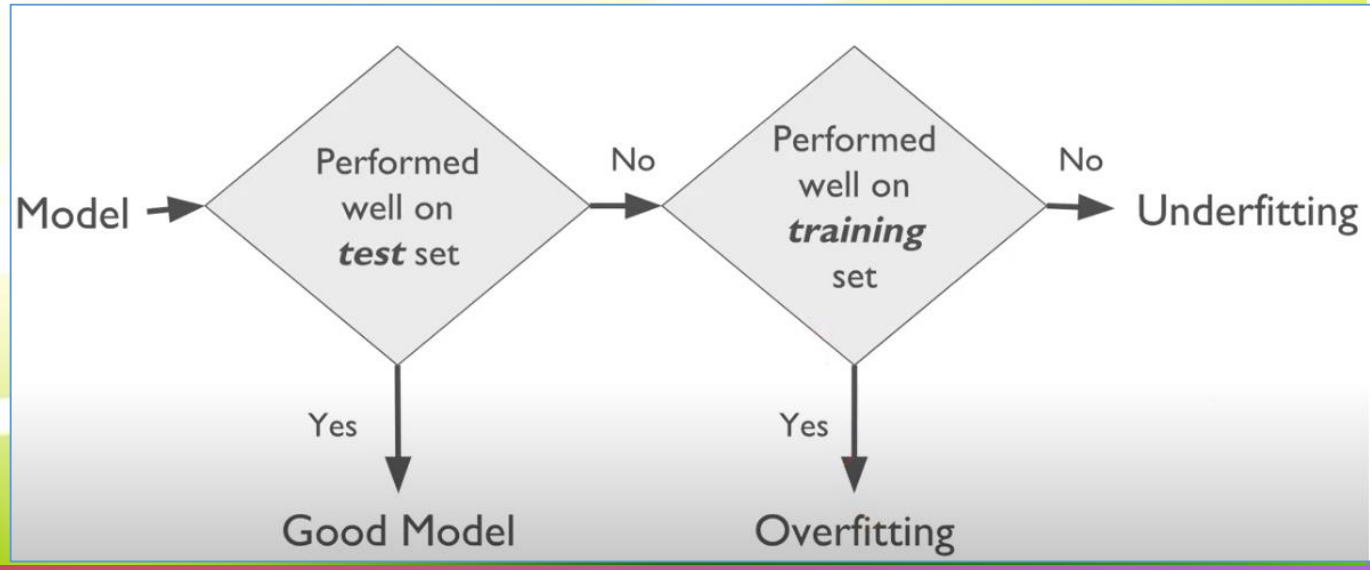
TERMINOLOGY – OVERFITTING

- **Overfitting** model tries to cover all/more than required data points present in given dataset.
  - Model starts catching noise/inaccurate values present in dataset.
  - model can't generalize or fit well on unseen dataset, and reduce model accuracy.
  - Error on testing or validation dataset is much greater than error on training dataset.
  - Overfitted model has low bias and high variance.
- Avoiding Overfitting:
  - Cross-Validation
  - Training with more clean data
  - Removing features
  - Early stopping the training
  - Regularization
  - Ensembling



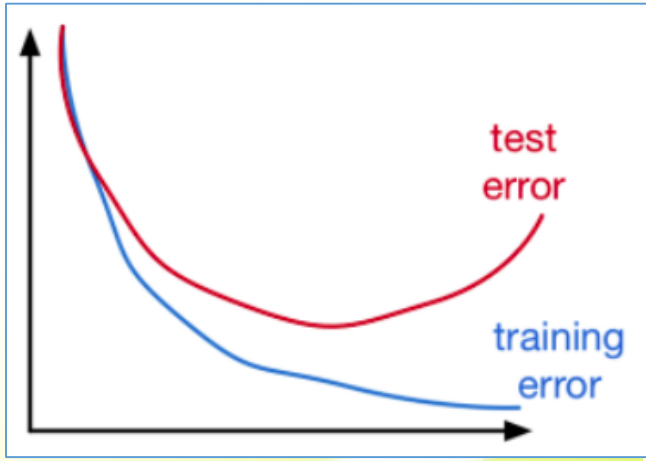
TERMINOLOGY – UNDERFITTING

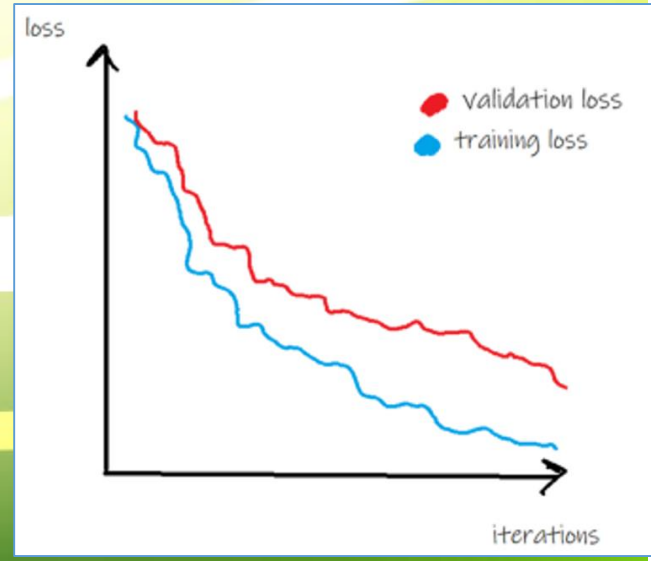
- Opposite of overfitting is underfitting.
- **Underfitting** when model is not able to capture underlying trend of data.
  - To avoid overfitting in model, volume/time of training data is compromised.
  - Hence, model is not able to learn enough from training data, and hence it reduces accuracy and produces unreliable predictions.
  - Underfitted model has high bias and low variance.
- Avoid underfitting:
  - increasing training time/volume of model.
  - increasing number of features.



Terminology - Generalization

- ultimate goal of machine learning is to find statistical patterns in a training set that generalize to data outside the training set.
- **Generalization**: model's ability to adapt and react appropriately to previously unseen, fresh data chosen from the same distribution as the model's initial input.
- Generalization assesses a model's ability to process new data and generate accurate predictions after being trained on a training set.
- Validation/Testing loss also decreases with decrease in Training loss.





Terminology – Parametric & Non Parametric Models

- **Parametric model:** A learning model that summarizes data with a set of parameters of fixed size (independent of the number of training examples).
- No matter how much data thrown at a parametric model, it won't change its mind about how many parameters it needs.
- Some examples of parametric machine learning algorithms:
  - Logistic Regression, Naive Bayes, Simple Neural Networks, Linear Discriminant Analysis, Perceptron

$$Y \approx W_0 + W_1 X_1 + W_2 X_2 + \cdots + W_P X_P$$

Benefits of Parametric Machine Learning Algorithms:

- **Simpler**: easier to understand and interpret results.
- **Speed**: very fast to learn from data.
- **Less Data**: do not require as much training data; can work well even if fit to data is not perfect.

Limitations of Parametric Machine Learning Algorithms:

- **Limited Complexity**: more suited to simpler problems.
- **Poor Fit**: In practice the methods are unlikely to match the underlying mapping function.

Terminology – Parametric & Non Parametric Models

- **Non-Parametric model:** Do not make strong assumptions about the mapping function.
- By not making assumptions, they are free to learn any functional form from the training data.
- Effective when data is huge and no prior knowledge, and uncertainty in choosing right features.
- Some examples of parametric machine learning algorithms:
    - k-Nearest Neighbors, Support Vector Machines, Decision Trees, etc.
    - KNN makes predictions based on k most similar training patterns for a new data instance; does not assume anything about form of mapping function other than similarity patterns.

Benefits of Nonparametric Machine Learning Algorithms:

- **Flexibility**: Capable of fitting a large number of functional forms.
- **Power**: No assumptions (or weak assumptions) about the underlying function.

Limitations of Nonparametric Machine Learning Algorithms:

- **More data**: Require a lot more training data to estimate the mapping function.
- **Slower**: A lot slower to train as they often have far more parameters to train.

Terminology – Dimensionality

- **Dimensionality**: number of input variables or features for a dataset.
- **Dimensionality reduction**: techniques that reduce number of input variables in a dataset.

Benefits of applying Dimensionality Reduction

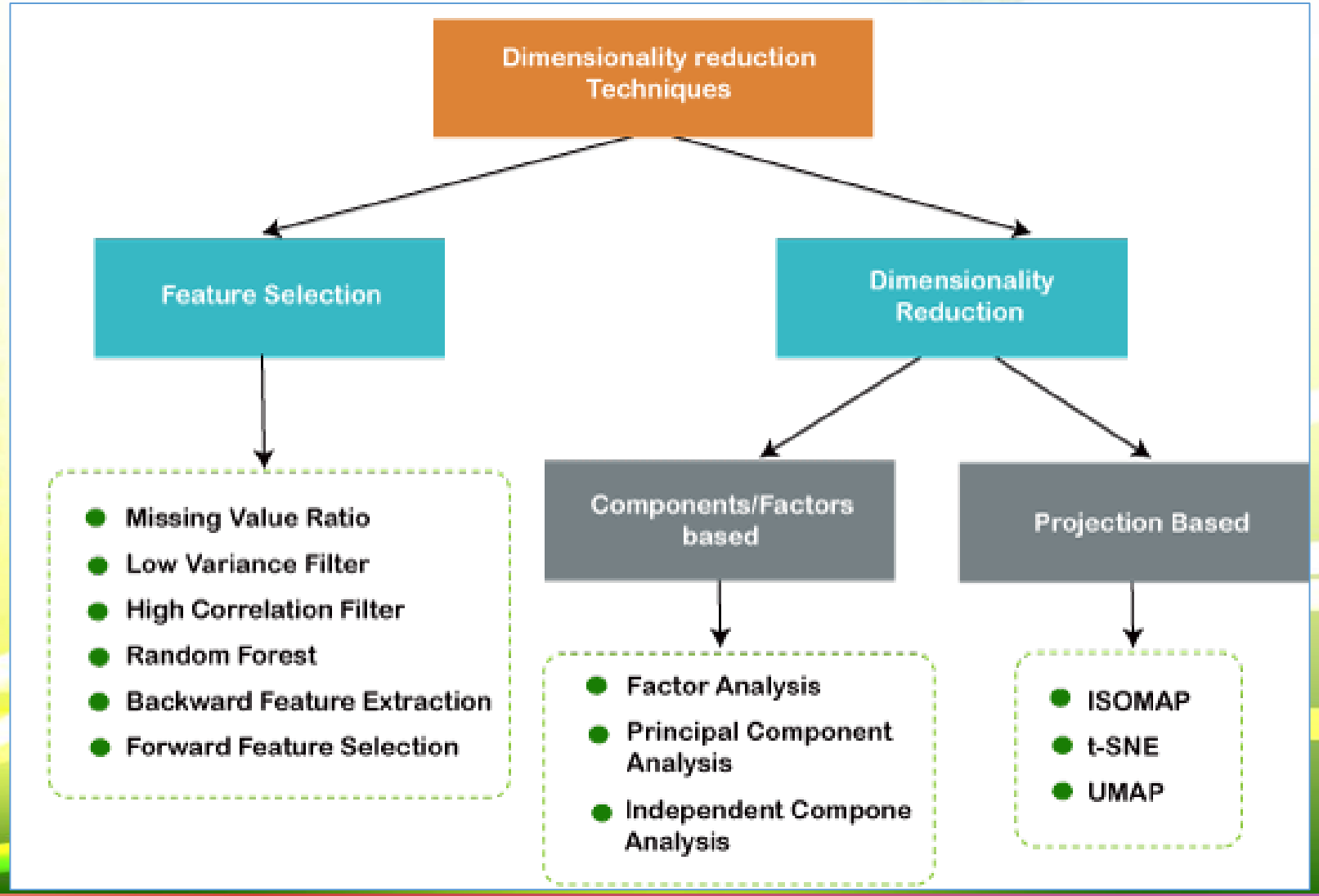
- Space required to store the dataset also gets reduced.
- Less Computation training time is required.
- Help in visualizing the data quickly.
- **R**emoves the redundant features (if present).

Player	Run Scored	Ball Faced

Disadvantages of dimensionality Reduction

- Some data may be lost due to dimensionality reduction.
- Sometimes the principal components required to consider are unknown.

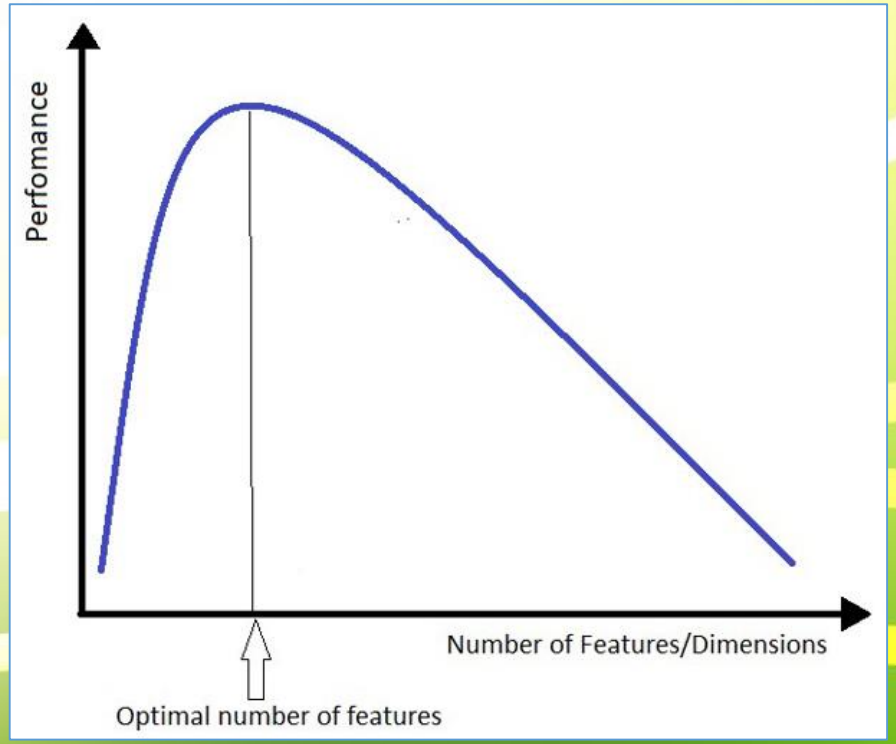
Terminology – Dimensionality Reduction



Terminology – Curse of dimensionality (CoD)

- Handling the high-dimensional data is very difficult in practice.
- **CoD** - difficulties related to training machine learning models due to high dimensional data
- If dimensionality of input dataset increases;
- number of samples also gets increased proportionally,
- chance of overfitting also increases.
- any machine learning model becomes more complex.
- If model is trained on high-dimensional data, it becomes overfitted and results in poor performance.

- Dimensionality reduction is very important.



Terminology – Model Selection

- Process of selecting one final machine learning model from among a collection of candidate machine learning models for a training dataset.
  - e.g., linear or logistic regression models with different degree polynomials,
  - KNN classifiers with different values of K, etc.

  - across different types of models (e.g. logistic regression, SVM, KNN, etc.)
  - across models of same type configured with different model hyperparameters (e.g. different kernels in an SVM).

- Model selection techniques:
  - **Probabilistic Measures**: Choose a model via in-sample error and complexity.
    - Analytical scoring of candidate model using both its performance on training dataset and the complexity of model.
  - **Resampling Methods**: Choose a model via estimated out-of-sample error.
    - estimate performance of model (development process) on out-of-sample data.
    - K-fold cross-validation: performance evaluation over different samples

Terminology - No free lunch theories

- NFL theory implies that no single machine learning algorithm is universally the best-performing algorithm for all problems.
- No one optimum optimization algorithm exists.
- The theory asserts that when performance of all optimization methods is averaged across all conceivable problems, they all perform equally well.
- In supervised learning, cross-validation is frequently used to compare prediction accuracy of many models of various complexity in order to select optimal model.
- A good model may also be trained using several methods.
- If no good conditions for the objective function are known, and one is just working with a black box, no guarantee can be made that this or that method outperforms.

MODEL EVALUATION - CONFUSION MATRIX

Model Evaluation: Process of using different evaluation metrics to understand a machine learning model's performance (also its strengths and weaknesses).

Confusion matrix: N X N matrix, where N is number of classes being predicted.

ACTUAL	Course-1	Course-2	Course-3
Pass	90	70	80
Fail	30	50	40

PREDICTED	Course-1	Course-2	Course-3
Pass	80	50	70
Fail	40	70	50

Course-1

	Predicted Pass	Predicted Fail	Total
Actual Pass	70	20	90
Actual Fail	10	20	30
Total	80	40	

MODEL EVALUATION - CONFUSION MATRIX

Confusion matrix: N X N matrix, where N is number of classes being predicted.

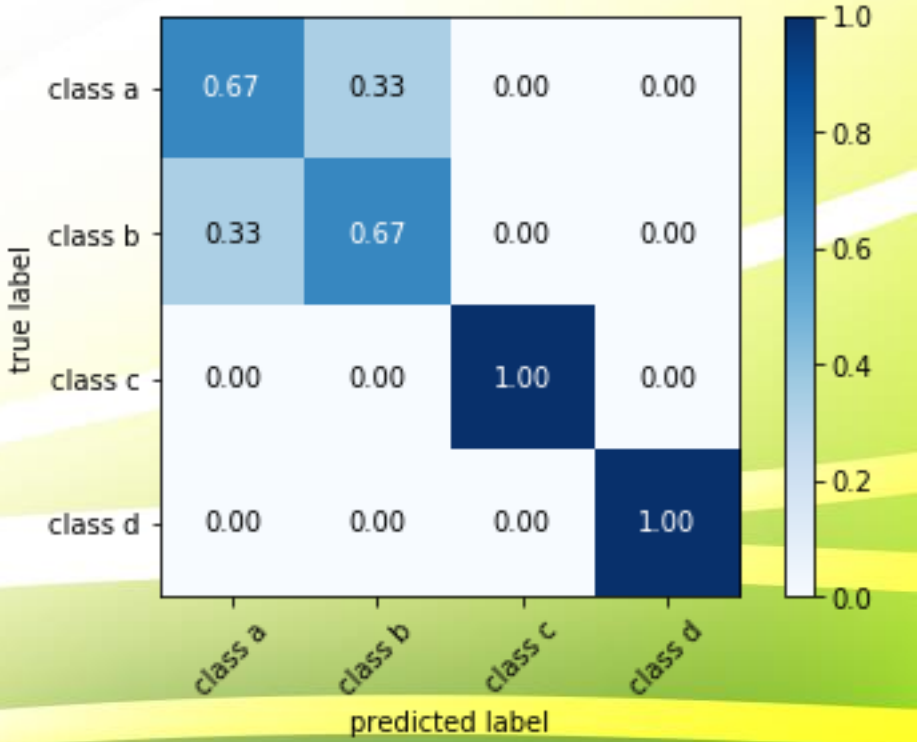
- **Error / residual:** difference between predicted value and actual value.
- **Accuracy / C*oncordance*:** proportion of total number of predictions that were correct.
- **Positive Predictive Value/Precision**: proportion of positive cases that were correctly identified.
- **Negative Predictive Value** : proportion of negative cases that were correctly identified.
- **Sensitivity or Recall** : proportion of actual positive cases which are correctly identified.
- **Specificity** : proportion of actual negative cases which are correctly identified.

Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

CONFUSION MATRIX

Predicted Class

Actual Class	**Positive**	**Negative**	
Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$



CONFUSION MATRIX

- **Classification accuracy** → total number of correct predictions divided by total number of predictions made for a dataset.
  - Accuracy is not always appropriate for all problems.

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

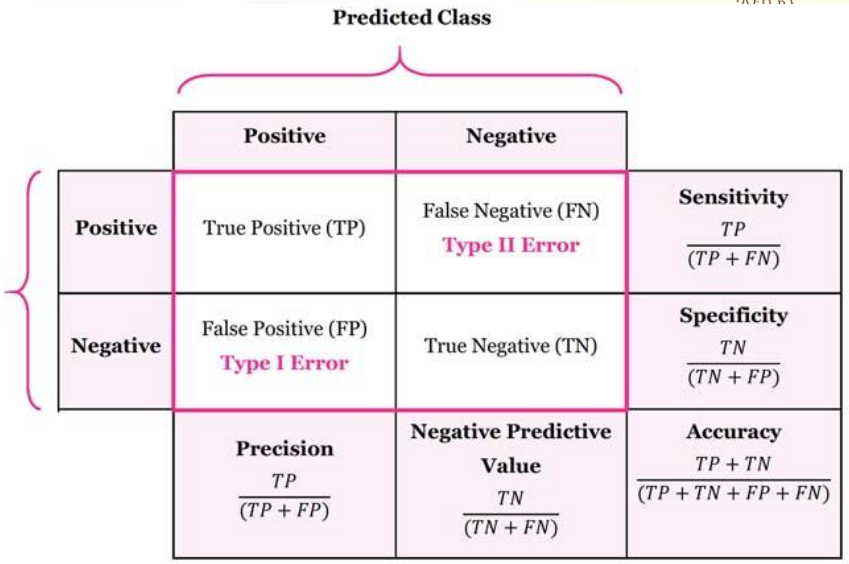
- Alternative to using classification accuracy → precision and recall metrics.
  - **Precision** → number of positive class predictions that actually belong to the positive class.
  - **Recall (sensitivity)** → number of positive class predictions made out of all positive examples in the dataset.
  - **F-Measure** → A single score that balances both the concerns of precision and recall.

CONFUSION MATRIX

- **Precision** → number of correct positive predictions made.

  Precision = TruePositives / (TruePositives + FalsePositives)

- 0.0 for no precision and 1.0 for full or perfect precision.



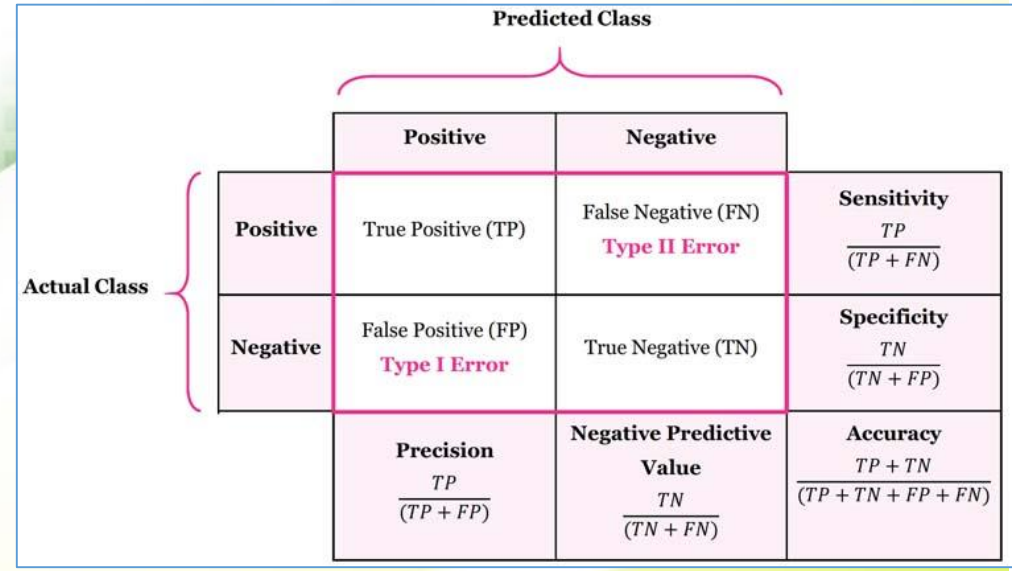
- A model makes predictions and predicts 120 examples as belonging to the positive class, 90 of which are correct, and 30 of which are incorrect.

  Precision = TruePositives / (TruePositives + FalsePositives) = 90 / (90 + 30) = 90 / 120 = 0.75

- Same dataset, another model predicts 50 examples belonging to the positive class, 45 of which are true positives and five of which are false positives.

  Precision = TruePositives / (TruePositives + FalsePositives) = 45 / (45 + 5) = 45 / 50 = 0.90

CONFUSION MATRIX



- **Recall** → number of correct positive predictions made out of all positive predictions that could have been made.

  Recall = TruePositives / (TruePositives + FalseNegatives)

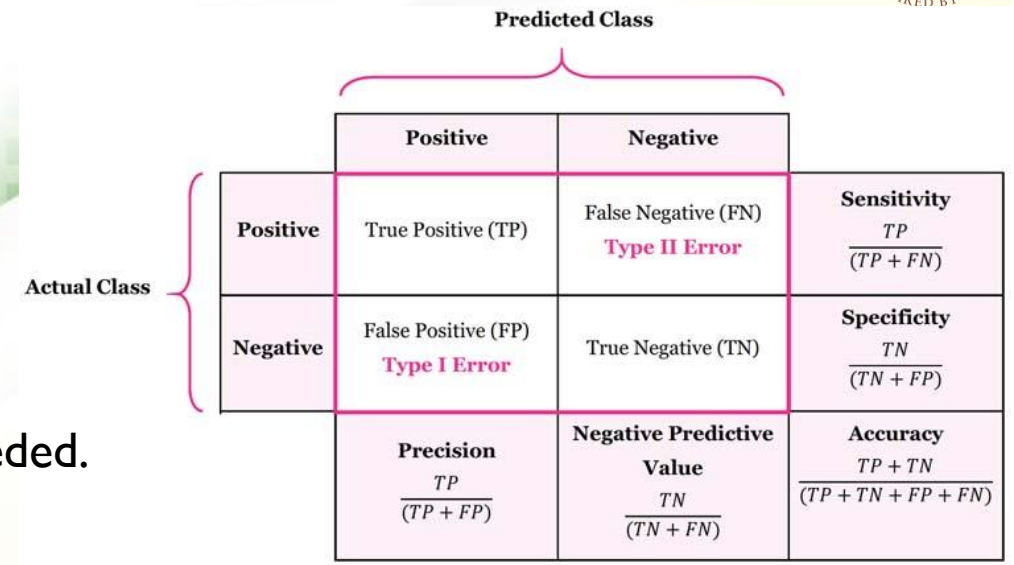
- 0.0 for no recall and 1.0 for full or perfect recall.

- A model makes predictions and predicts 90 of the positive class predictions correctly and 10 incorrectly.

  Recall = TruePositives / (TruePositives + FalseNegatives)

  Recall = 90 / (90 + 10) = 90 / 100 = 0.9

CONFUSION MATRIX

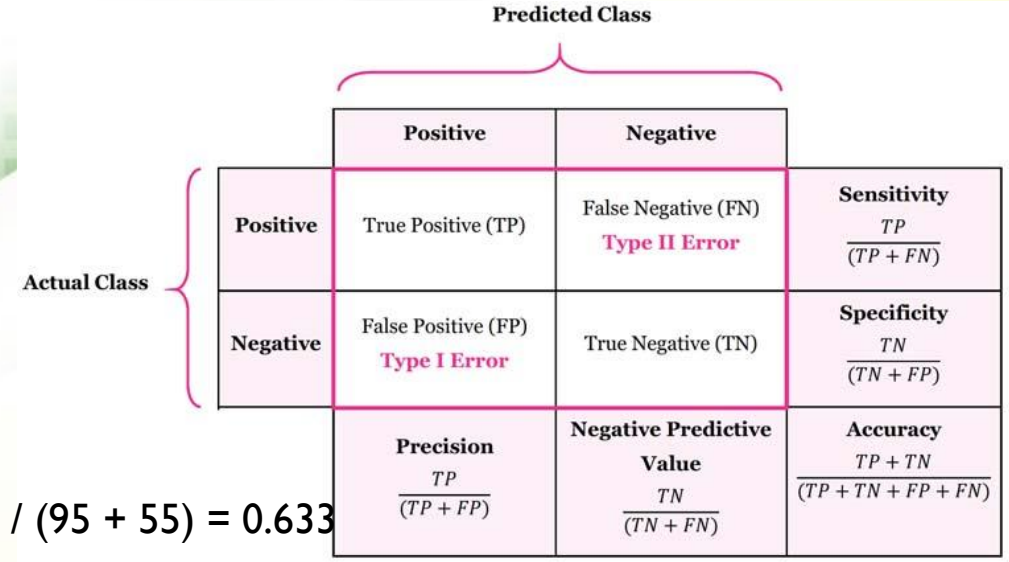


- Maximizing precision will minimize the number false positives.
- Maximizing the recall will minimize the number of false negatives.
  - For excellent predictions both high precision and high recall needed.
  - Neither precision or recall tells whole story.
    - *Excellent precision with terrible recall, or terrible precision with excellent recall!!!*
  - Increases in recall often come at the expense of decreases in precision, and vice-versa.
- Instead of picking any one measure, a new metric can be used that combines both precision and recall into one score → F/F1-score (harmonic mean of both)

  F-Measure = (2 * Precision * Recall) / (Precision + Recall)
- 0.0 is poor F-Measure score and 1.0 is best or perfect F-Measure score

CONFUSION MATRIX

- Consider a model that predicts 150 examples for the positive class, 95 are correct (true positives), meaning five were missed (false negatives) and 55 are incorrect (false positives).



Precision = TruePositives / (TruePositives + FalsePositives) = 95 / (95 + 55) = 0.633

Recall = TruePositives / (TruePositives + FalseNegatives) = 95 / (95 + 5) = 0.95

Model has poor precision, but excellent recall.

F-Measure = (2 * Precision * Recall) / (Precision + Recall)

= (2 * 0.633 * 0.95) / (0.633 + 0.95) = (2 * 0.601) / 1.583 = 0.759

- Good recall levels-out poor precision, giving an okay or reasonable F-measure score.

CONFUSION MATRIX

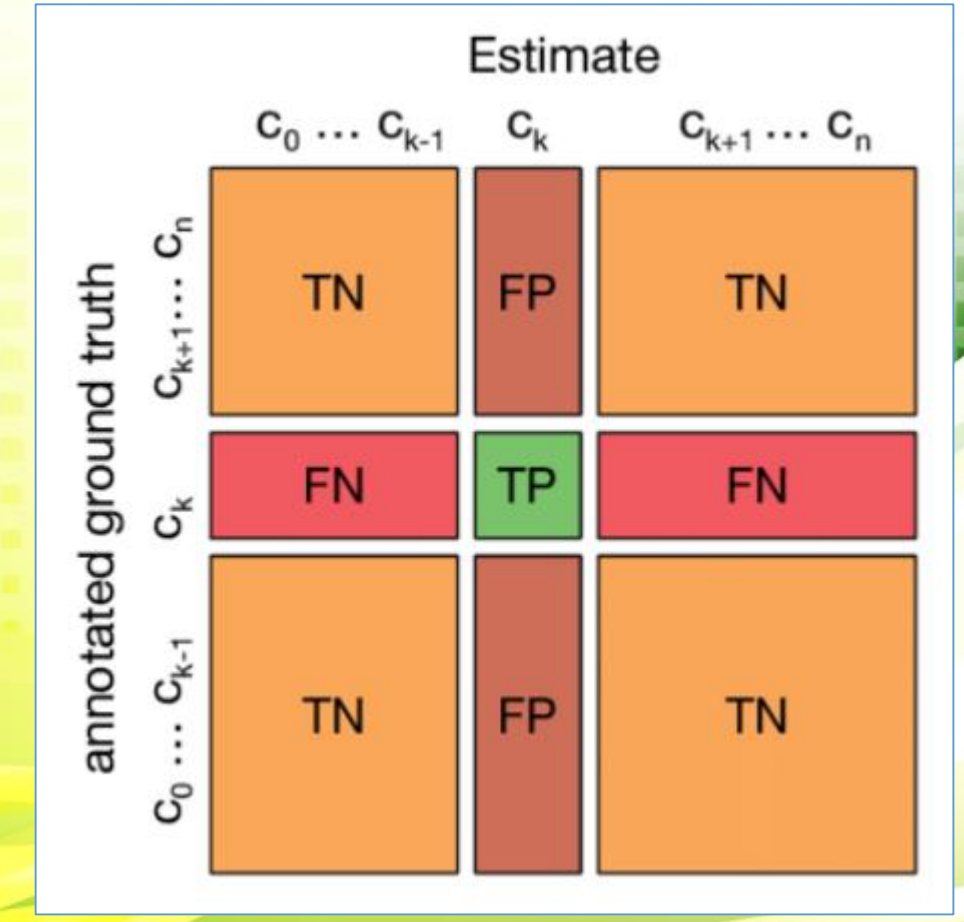
Example: Suppose we are trying to create a model that can predict the result for the disease that is either a person has that disease or not. The table is given for the two-class classifier, which has two predictions "Yes" and "NO." **Yes** defines that patient has the disease, and **No** defines that patient does not has that disease.

- The classifier has made a total of 100 predictions.
- Out of 100 predictions, 89 are true predictions.
- The model has given prediction "yes" for 32 times. Whereas the actual "Yes" was 27 times.
- In 3 instances, an actual patient was wrongly diagnosed.

Prepare the Confusion matrix and calculate all the major performance parameters.

n = 100	Actual: No	Actual: Yes	
Predicted: No	TN: 65	FP: 3	68
Predicted: Yes	FN: 8	TP: 24	32
	73	27	

CONFUSION MATRIX - MULTICLASS



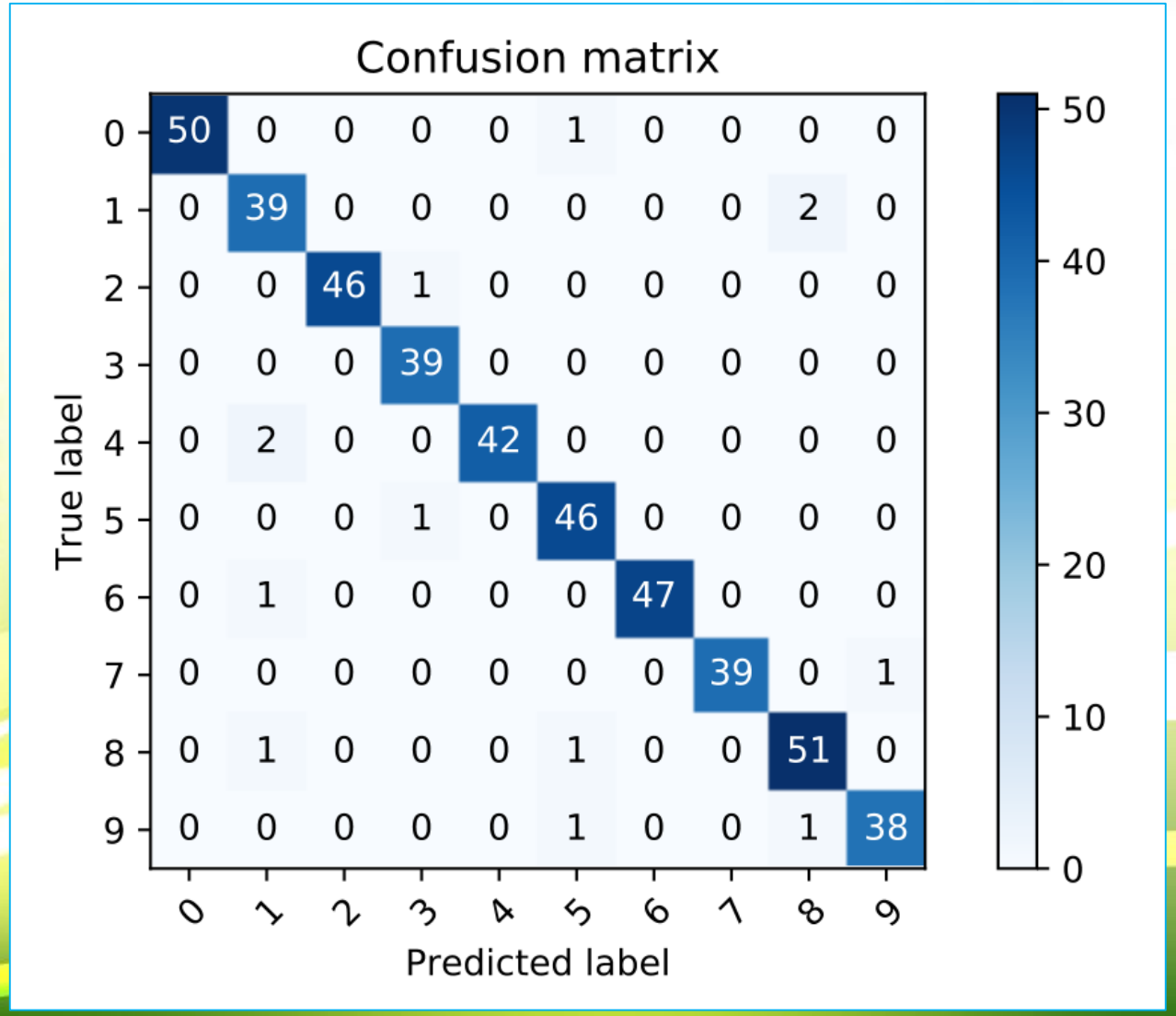
Predicted Class

Actual Class	Positive	Negative	
Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
	Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

gold labels

system output	urgent	normal	spam	
urgent	8	10	1	**precision**$_u$ = $\frac{8}{8+10+1}$
normal	5	60	50	**precision**$_n$ = $\frac{60}{5+60+50}$
spam	3	30	200	**precision**$_s$ = $\frac{200}{3+30+200}$
	recall$_u$ = $\frac{8}{8+5+3}$	**recall**$_n$ = $\frac{60}{10+60+30}$	**recall**$_s$ = $\frac{200}{1+50+200}$	

CONFUSION MATRIX - MULTICLASS



CONFUSION MATRIX- MULTICLASS

Example: For the given confusion matrix, calculate all relevant performance measures.

		Predicted Values		
		Setosa	Versicolor	Virginica
Actual Values	Setosa	16 (cell 1)	0 (cell 2)	0 (cell 3)
	Versicolor	0 (cell 4)	17 (cell 5)	1 (cell 6)
	Virginica	0 (cell 7)	0 (cell 8)	11 (cell 9)

Regression Model Evaluation

- ***Residual***: error term representing difference between observed value (*y*) and predicted value.

$$\hat{e} = y - \hat{y}$$

- Residual analysis helps to better understand how well model is performing.
- ***Sum of squares total* (*SST*)** : measure of variation of *y*-values about their mean.
- ***Sum of squares due to regression* (*SSR*):** differences between predicted/regression values and average y-value.
- ***Sum of squares* of *error* (*SSE*):** differences between actual *y*-values and predicted *y*-values.

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SST = SSR + SSE$$

Regression Model Evaluation

- ***Coefficient of determination*** **(R^2)**: proportion of variation.

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad R^2 = \frac{SSR}{SST}$$

- R^2 values vary between 0 and 1.
- R^2 closer to 1 → more accurate model predictions (models have a *closer fit).*
- In multiple linear regression, *adjusted* R^2 value (R^2 adj) is usually considered to better account for the multiple independent variables used in analysis as well as sample size.

$$R^2_{\text{adj}} = 1 - \left[\frac{(1 - R^2)(n - 1)}{n - k - 1}\right]$$

n : number of observations
k : number of independent variables.

Regression Model Evaluation

- *standard error of the estimate* ($S_{y,x}$) : measure of variation of *y*-values about regression line.

$$s_{y.x} = \sqrt{\frac{\text{SSE}}{n-2}}$$

  - interpreted in a similar manner to standard deviation.
  - indicates model's accuracy: larger the value for standard error of estimate, lower the precision.
- t-Test, F-Test performed to assess variable dependencies and model performance.
- **Mean Absolute Error (MAE):** simple metric; Not preferred where outliers are prominent.
- **Mean Squared Error (MSE):** most common metric for regression models.
- **Root Mean Squared Error (RMSE):** square root MSE.
  - RMSE penalizes large errors..

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

Regression Model Evaluation

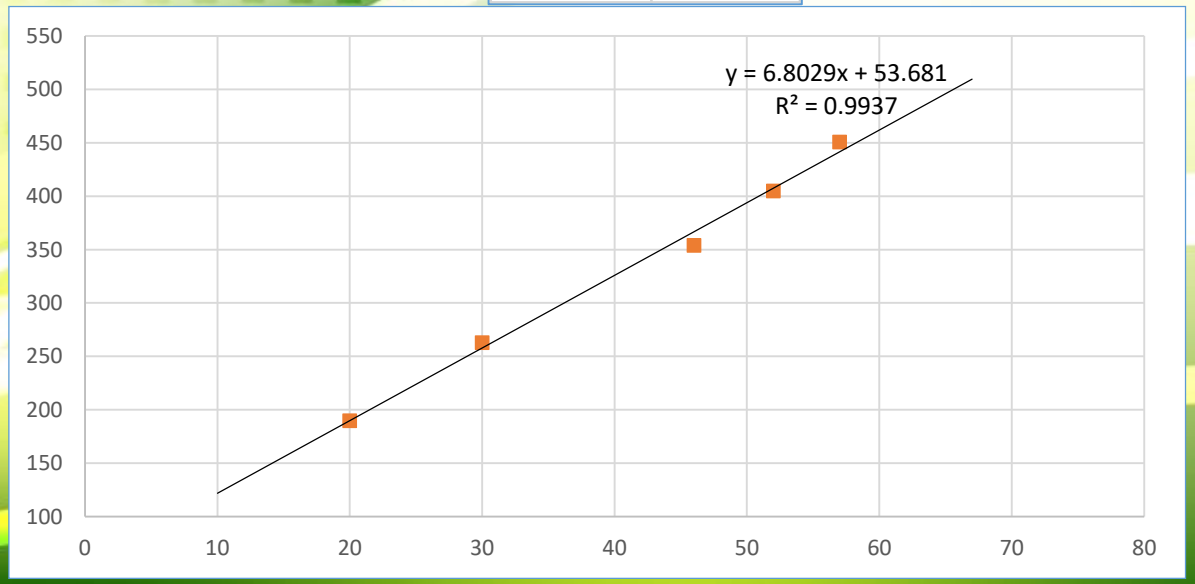
Example: For the given data, calcucate the MAE, MSE, RMSE, and R^2 for simple linear regression model.

Age (x)	BP (Y)
46	354
20	190
52	405
30	263
57	451

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$ **56.8**

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$ **6.0**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$ **7.5**



Probability Theory

Probability

- Event A
- P(A)
- P(not A)

Probability → likelihood/chances of occurrence of an activity

$$P(A) = \frac{\text{Number of favorable outcomes to A}}{\text{Total number of possible outcomes}}$$

P(A') = 1 - P(A). P(A')

Probability Theory

Probability

What is the probability of drawing a queen from a deck of cards?

$$P(A) = \frac{\text{Number of favorable outcomes to A}}{\text{Total number of possible outcomes}}$$

A deck of cards has 4 suits.

Each suit consists of 13 cards.

total number of possible outcomes = (4)(13) = 52

There can be 4 queens, one belonging to each suit.

number of favorable outcomes = 4.

card probability = 4 / 52 = 1 / 13

Probability Theory

Probability

When two dice are rolled what is the probability of getting a sum of 8?

When two dice are rolled there are 36 possible outcomes.

To get the sum as 8 there are 5 favorable outcomes.

[(2, 6), (6, 2), (3, 5), (5, 3), (4, 4)]

$$P(A) = \frac{\text{Number of favorable outcomes to A}}{\text{Total number of possible outcomes}}$$

= 5 / 36

Probability Theory

PROBABILITY: likelihood for one variable.

Probability of Event A or Event B.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- **Marginal probability P(A):** probability of an event irrespective of outcome of another variable
- **Joint probability P(A,B)**: probability that two events will both occur (likelihood of two events occurring together).

$$\text{Joint Probability} = P(A \cap B) = P(A) \times P(B)$$

- **Conditional probability:** probability of one event occurring given that second event has happened.

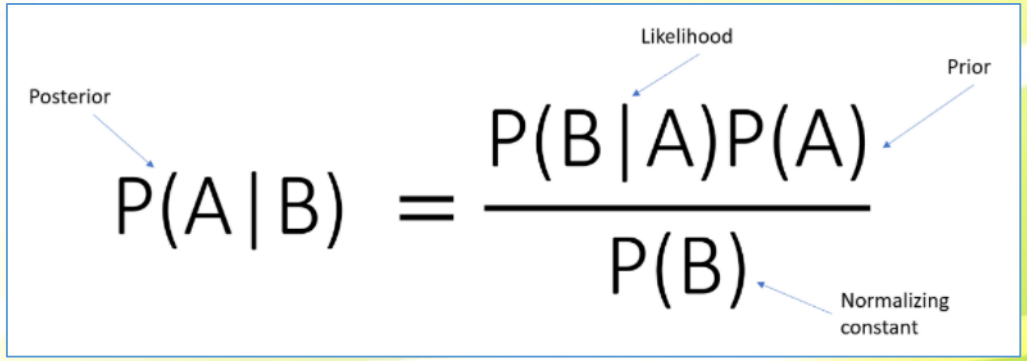
$$P(X/Y) = \frac{P(X \cap Y)}{P(Y)}, \text{ if } P(Y) \neq 0$$

$$P(Y/X) = \frac{P(Y \cap X)}{P(X)}, \text{ if } P(X) \neq 0$$

Probability Theory

Bayes' Theorem:

- Used to determine probability of a hypothesis with prior knowledge.
- It depends on conditional probability.
- Using Bayes theorem helps finding the probability of A, given that B occurred.

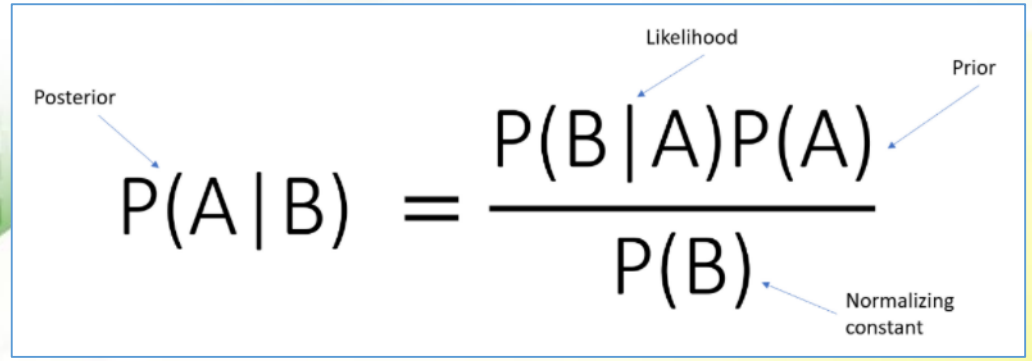


where,

- A and B are the events and P (B) ≠ 0
- **Conditional/Posterior probability: P(A/B), P(B/A)**
- **Marginal/prior probability: P(A), P(B)**
- P(A/B) is a **conditional probability** that describes the occurrence of event **A** is given that **B** is true.
- P(B/A) is a **conditional probability** that describes the occurrence of event **B** is given that **A** is true.
- P(X) and P(Y) are the probabilities of observing X and Y independently of each other.

Probability Theory

Bayes' Theorem:



- A is the hypothesis and B is the evidence/condition.
- P(A) and P(B) is the independent probabilities of A and B.
- **P(A) is Prior Probability**: Probability of hypothesis before observing the evidence.
- **P(B) is Marginal Probability**: Probability of Evidence.
- **P(A|B) is Posterior probability**: Probability of hypothesis A on the observed event B.
- **P(B|A) is Likelihood probability**: Probability of the evidence given that the probability of a hypothesis is true.

Probability Theory

Probability

Out of 10 people, 3 bought pencils, 5 bought notebooks and 2 got both pencils and notebooks. If a customer bought a notebook what is the probability that she also bought a pencil.

Using the concept of conditional probability in probability theory,

$$P(A \mid B) = P(A \cap B) / P(B)$$

A : event of people buying pencils

B : event people of buying notebooks.

$P(A) = 3 / 10 = 0.3$

$P(B) = 5 / 10 = 0.5$

$P(A \cap B) = 2 / 10 = 0.2$

$P(A \mid B) = 0.2 / 0.5 = 0.4$

Probability Theory

- **Chain rule of conditional probability:**

  $P(A,B) = p(A|B)\, p(B)$

  $P(A,B|C) = P(A|B,C)\, P(B|C)$

$$P(X/Y) = \frac{P(X \cap Y)}{P(Y)}, \text{ if } P(Y) \neq 0$$

$$P(Y/X) = \frac{P(Y \cap X)}{P(X)}, \text{ if } P(X) \neq 0$$

- Extend this for three variables:
  - *joint probability distribution of conditional probabilities.*

  $P(A,B,C) = P(A|\, B,C)\, P(B,C) = P(A|B,C)\, P(B|C)\, P(C)$

$$P(Y \cap X) = P(X \cap Y)$$

$$\text{or, } P(X \cap Y) = P(X/Y)P(Y) = P(Y/X)P(X)$$

$$\text{or, } P(X/Y) = \frac{P(Y/X)P(X)}{P(Y)}, \text{ if } P(Y) \neq 0$$

- General to n variables:

  $P(A1, A2, ..., An) = P(A1|\, A2, ..., An)\, P(A2|\, A3, ..., An)\, P(An\text{-}1|An)\, P(An)$

Probability Theory

- Bayes theorem can be derived from the conditional probability:

$$P(X/Y) = \frac{P(X \cap Y)}{P(Y)}, \text{ if } P(Y) \neq 0$$

$$P(Y/X) = \frac{P(Y \cap X)}{P(X)}, \text{ if } P(X) \neq 0$$

- P (X⋂Y) is **joint probability** of both X and Y being true.
- P(Y) = P(Y | X) * P(X) + P(Y | not X) * P(not X)
- **complement of X:** P(not X) = 1 – P(X)

$$P(Y \cap X) = P(X \cap Y)$$

$$\text{or, } P(X \cap Y) = P(X/Y)P(Y) = P(Y/X)P(X)$$

$$\text{or, } P(X/Y) = \frac{P(Y/X)P(X)}{P(Y)}, \text{ if } P(Y) \neq 0$$

Probability Theory

Bag1 contains 4 white and 8 black balls and Bag2 contains 5 white and 3 black balls. From one of the bag one ball is drawn at random and the ball which is drawn comes out as black. Find the probability that the ball is drawn from Bag1.

Let E1, E2 and A be three events,
E1 = Event of selecting Bag1
E2 = Event of selecting Bag2
A = Event of drawing black ball

P(E1) = P(E2) = 1/2
P(drawing a black ball from Bag1) = P(A|E1) = 8/12 = 2/3
P(drawing a black ball from Bag2) = P(A|E2) = 3/8

P(A) = P(drawing a black ball) = P(A|E1) * P(E1) + P(A|E2) * P(E2)
= 2/3 * 1/2 + 3/8 * 1/2 = 25/48

According to Bayes' Theorem,
Probability(drawing a black ball from Bag1)
P(E1|A) = P(A|E1) * P(E1) / P(A)

According to Bayes' Theorem,
Probability(drawing a black ball from Bag1)
P(E1|A) = P(A|E1) * P(E1) / P(A)
= (2/3 * 1/2) / (25/48) = 16/25

Probability that ball is drawn from Bag1 is 16/25

Probability Theory

Imagine you are a financial analyst at an investment bank. According to your research of publicly-traded companies, 60% of the companies that increased their share price by more than 5% in the last three years replaced their CEOs during the period.

At the same time, only 35% of the companies that did not increase their share price by more than 5% in the same period replaced their CEOs. Knowing that the probability that the stock prices grow by more than 5% is 4%, find the probability that the shares of a company that fires its CEO will increase by more than 5%.

Define the notation of probabilities.
P(I) –probability that stock price increases by 5% = (4%) = 0.04
P(R) –probability that CEO is replaced = (60%) = 0.6
P(R | I) –probability of CEO replacement given stock price has increased by 5% = (60%) = 0.6
P(I | R) –probability of stock price increases by 5% given that CEO has been replaced = ???

P(I`) = 1 – 0.4 = 0.6
P(R | I`) = (35%) = 0.35

According to Bayes' Theorem,
P(I | R) = P(R | I) * P(I) / P(R)

P(R) = P(R | I) * P(I) + P(R | I`) * P(I`)

$$P(A|B) = \frac{0.60 \times 0.04}{0.60 \times 0.04 + 0.35 \times (1 - 0.04)} = 0.067 \text{ or } 6.67\%$$

Probability Theory

XYZ works as a mail filter to identify spam in which users train the system. In emails, it considers patterns in words which are marked as spam by users. For Example, it may have learned that the word "release" is marked as spam in 30% of the emails. Concluding 0.8% of non-spam mails which includes the word "release" and 40% of all emails which are received by user is spam. Find probability that a mail is a spam if the word "release" seems in it.

P(R | S) = 0.30
P(R | N) = 0.008
P(S) = 0.40
=> P(N) = 0.60

P(S | R) = ?

Now, using Bayes' Theorem:
P(S | R) = P(R | S) * P(S) / P(R)

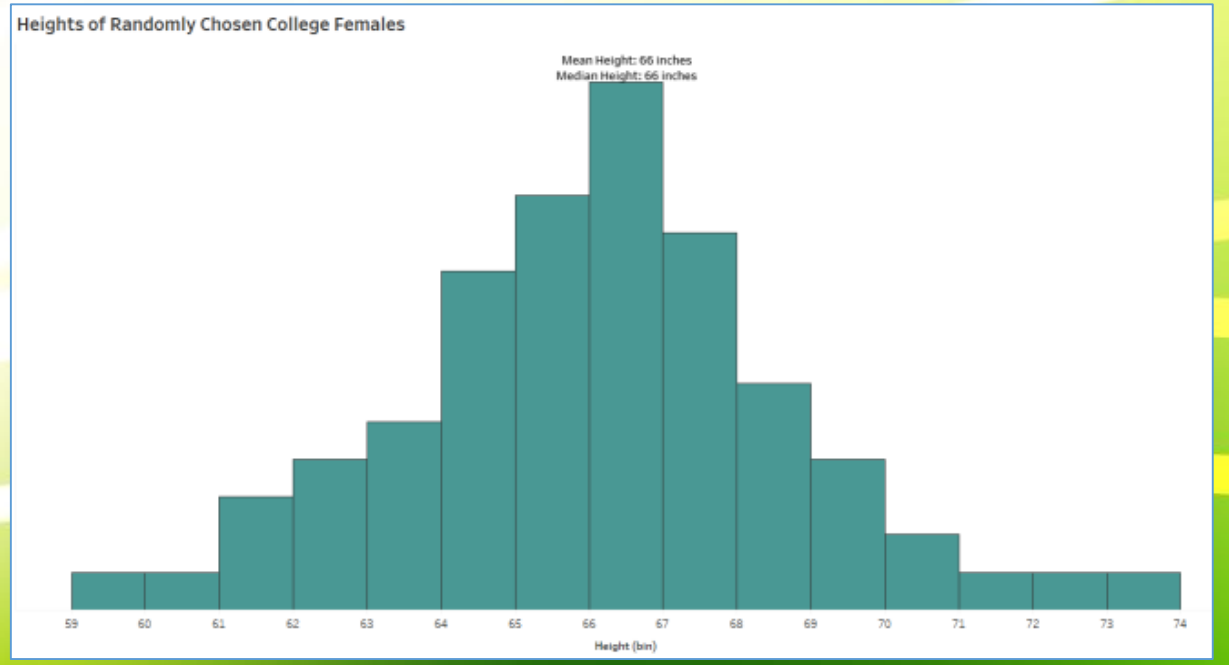
P(R) = P(R | S) * P(S) + P(R | N) * P(N)
= 0.40 * 0.30 + 0.30 *
0.008 = 0.1224

Using Bayes' Theorem:
P(S | R) = P(R | S) * P(S) / P(R)
= 0.30 * 0.40 / 0.1224 = 0.980

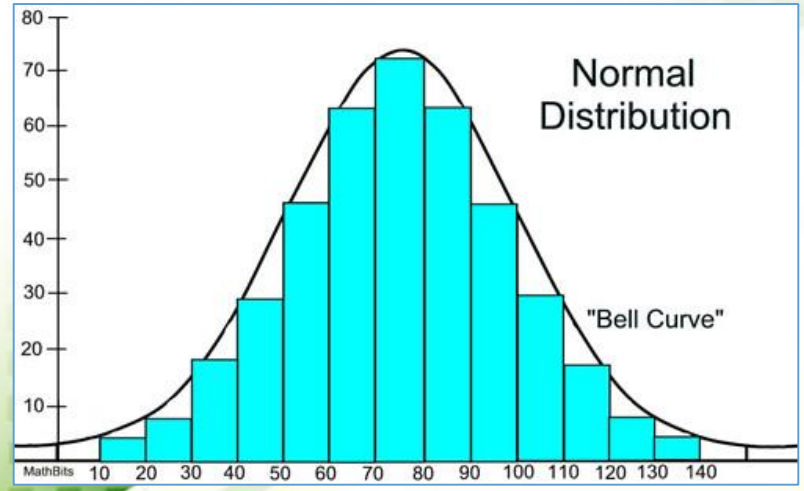
Probability Distribution

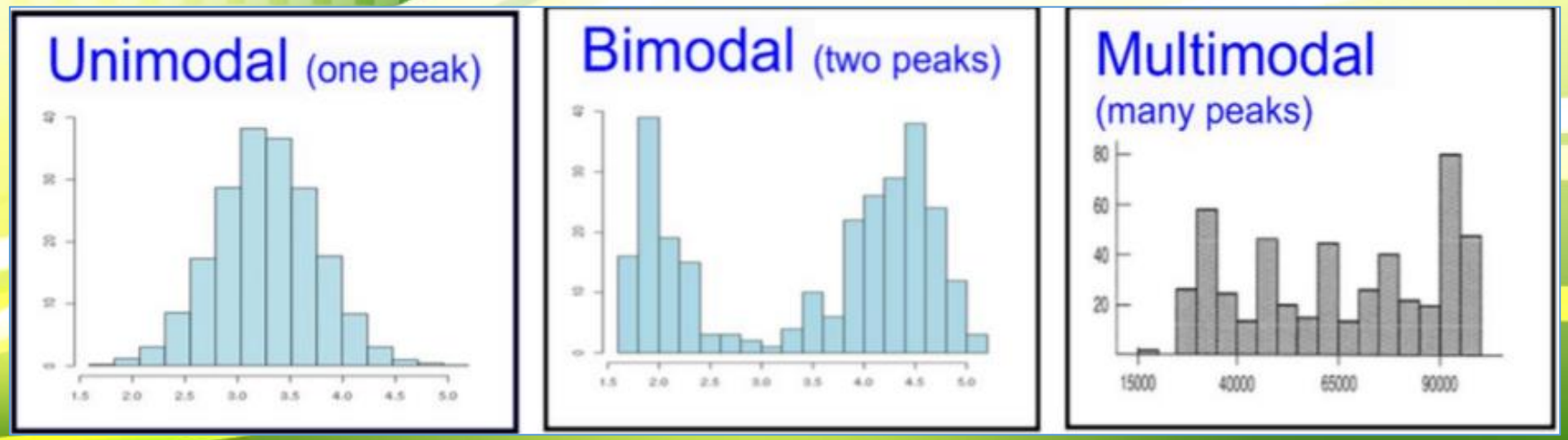
- **Distribution:** a function that shows possible values for a variable and how often they occur.
- **Frequency distribution** gives an idea about how frequently given data point occurs & how probable it is to occur.
- **Probability distribution** gives probability of occurrence of given data point.
- For large test cases frequency distribution and probability distributions are similar in shape.

- When shape of the **distribution is symmetric and unimodal**, the mean, median, and mode are equal.



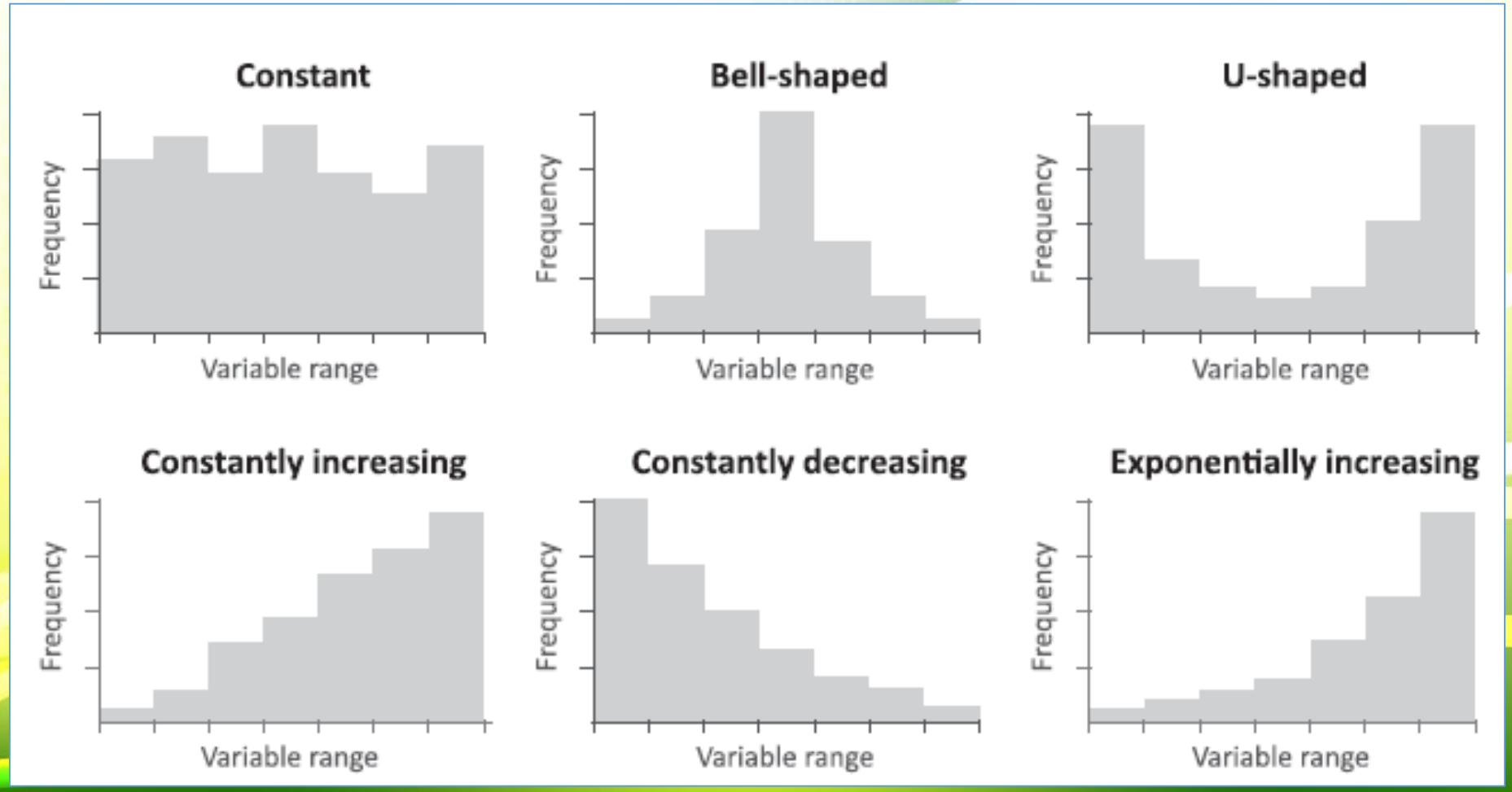
Probability Distribution





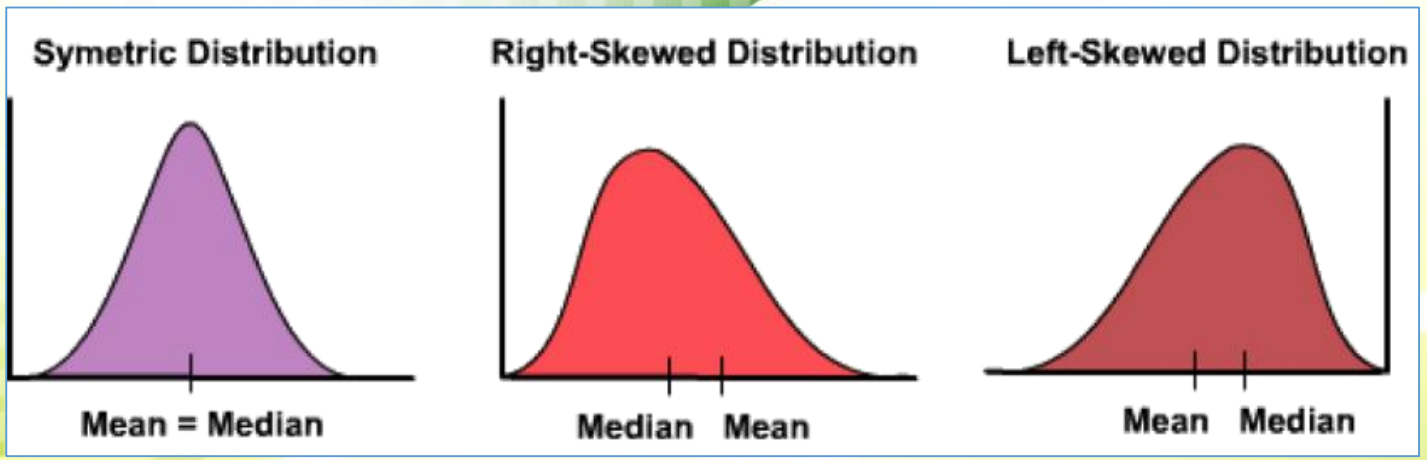
Probability Distribution

Types of distribution.



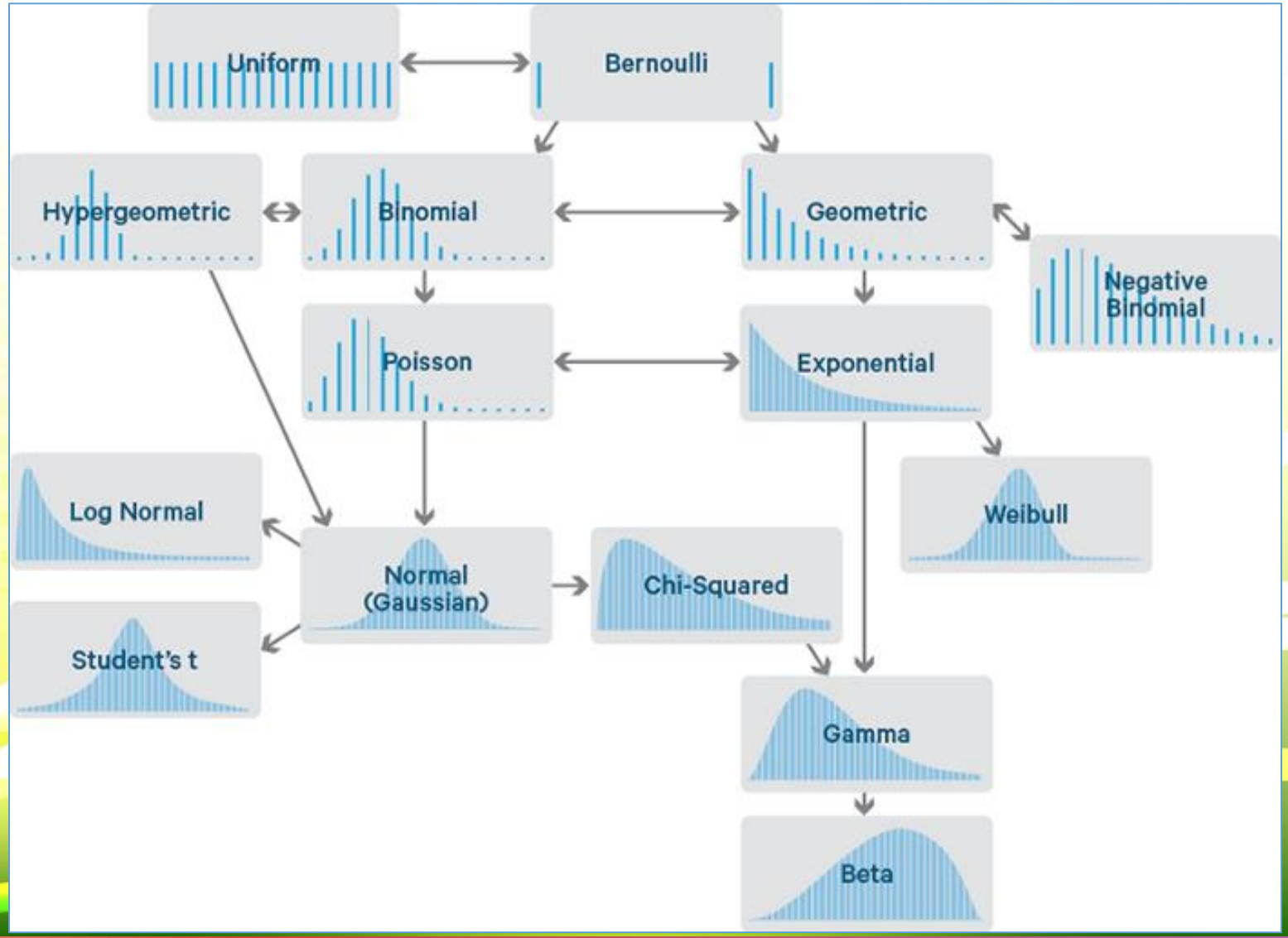
Probability Distribution

- **Skewed distribution** is neither symmetric nor normal → data values trail off more sharply on one side than on the other.



- Skewed data Degrades model's ability.
- Right skewed data will predict better on data points with lower value as compared to those with higher values.
- Skewed data also does not work well with many statistical methods.
- To ensure that model capabilities is not affected, skewed data has to be transformed to approximate to a normal distribution.

Probability Distribution



Probability Density Function (PDF)

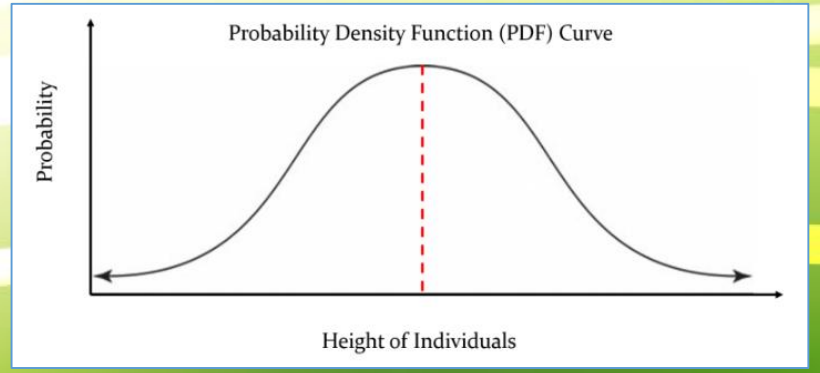
- **Probability density** is the relationship between observations and their probability.
- Random variable x has a probability distribution p(x).
- If random variable is continuous, then probability can be calculated via **probability density function**.
- **Probability distribution**: Shape of PDF across the domain for a random variable.
  - *describes all possible values and likelihoods that a random variable can take within a given range.*
- Common probability distributions are uniform, normal, exponential, etc.

$$P(a \leq X \leq b) = \int_a^b f(x) d_x$$

$[a, b]$ = Interval in which x lies.

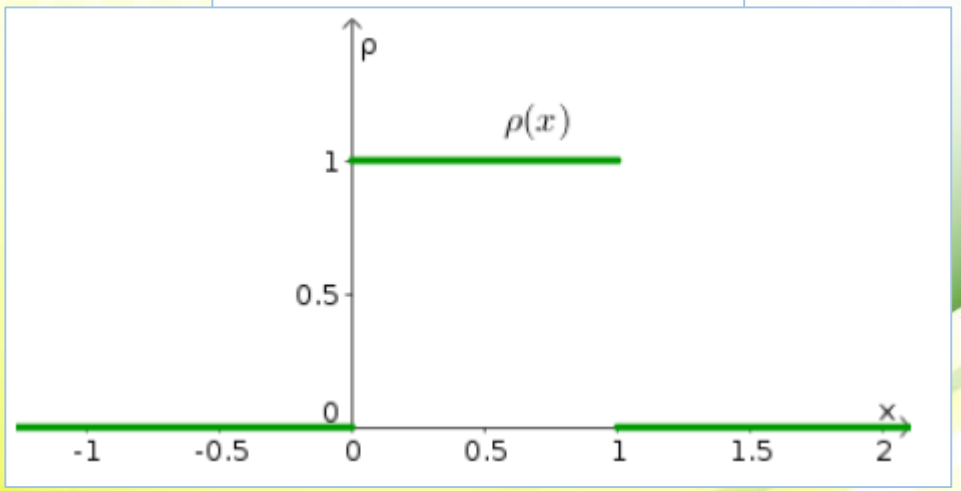
$P(a \leq X \leq b)$ = probability that some value x lies within this interval.

d_x = b-a

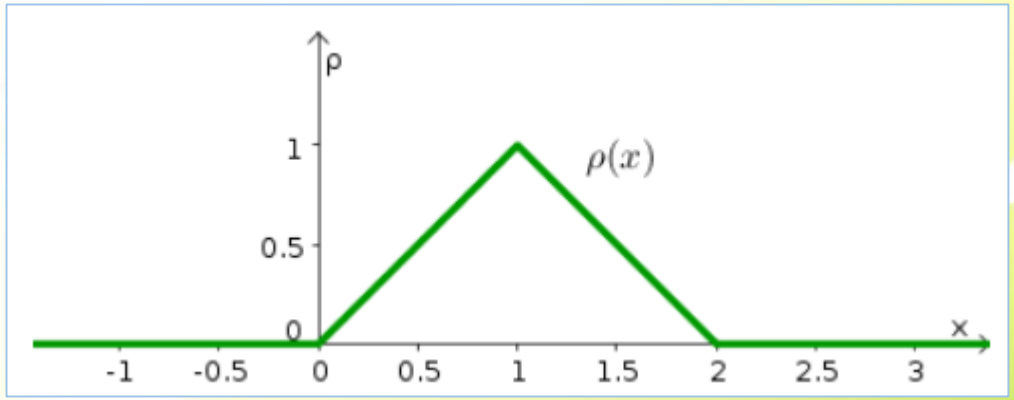


Probability Density Function (PDF)

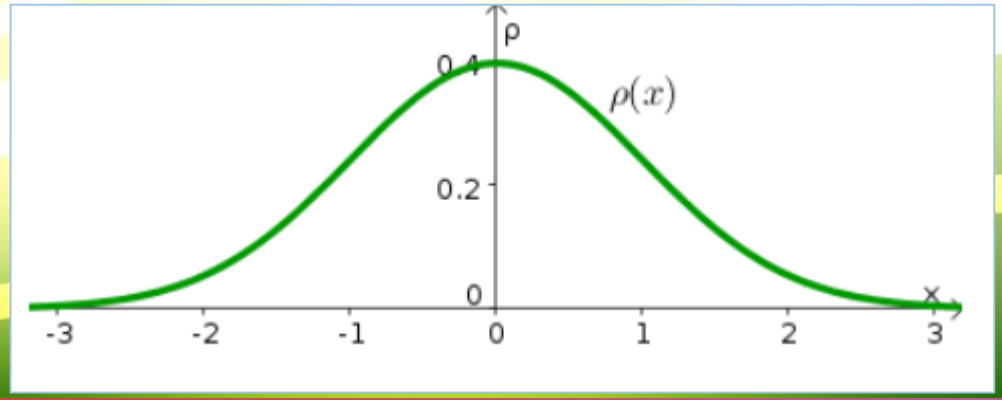
$$\rho(x) = \begin{cases} 1 & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$



$$\rho(x) = \begin{cases} x & \text{if } 0 < x < 1 \\ 2 - x & \text{if } 1 < x < 2 \\ 0 & \text{otherwise,} \end{cases}$$



$$\rho(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$



y or $f(x)$	$\frac{dy}{dx}$ or $f'(x)$
K, K is a constant	$\frac{d}{dx}(K) = 0$
x^n	$\frac{d}{dx}(x^n) = n x^{n-1}$
$\sqrt{x}$	$\frac{d}{dx}(\sqrt{x}) = \frac{1}{2\sqrt{x}}$
e^x	$\frac{d}{dx}(e^x) = e^x$
a^x, $a > 0$	$\frac{d}{dx}(a^x) = a^x \log a$

$$\int cf(x)dx = c\int f(x)dx$$

$$\int [f(x)+g(x)]dx = \int f(x)dx + \int g(x)dx$$

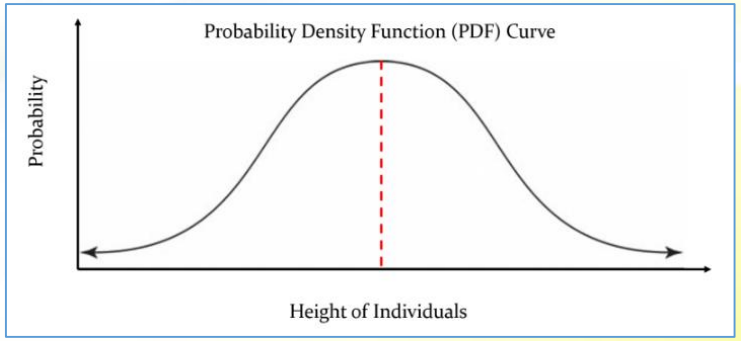
$$\int k\,dx = kx + C$$

$$\int x^n dx = \frac{x^{n+1}}{n+1} + C \quad (n \neq 1)$$

$$\int e^u du = e^u + C$$

$$\int a^u du = \frac{a^u}{\ln a} + C \qquad (a > 0, a \neq 1)$$

Probability Density

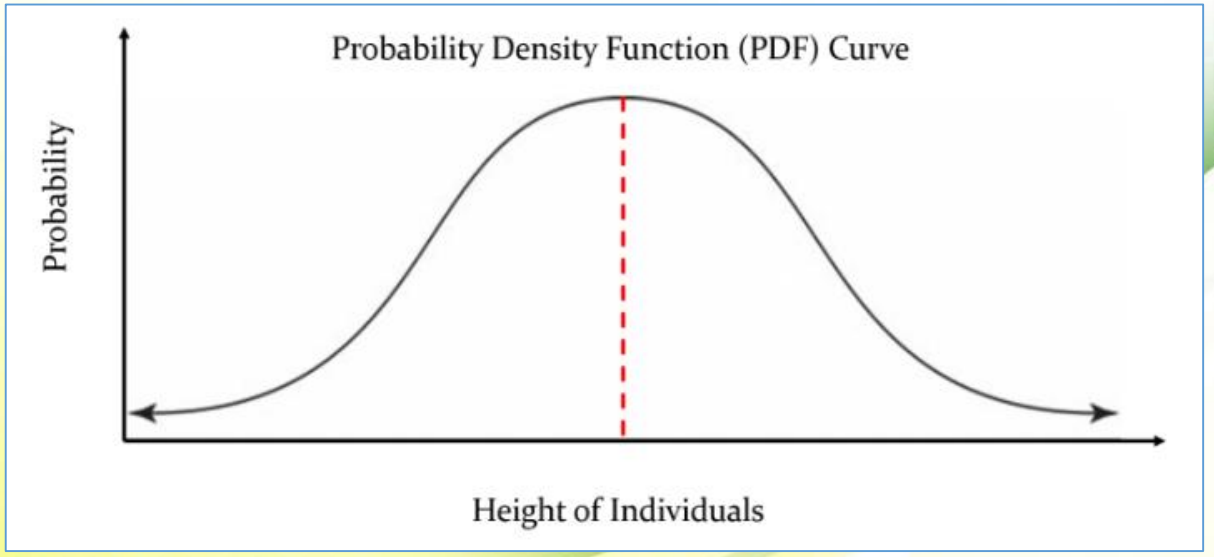


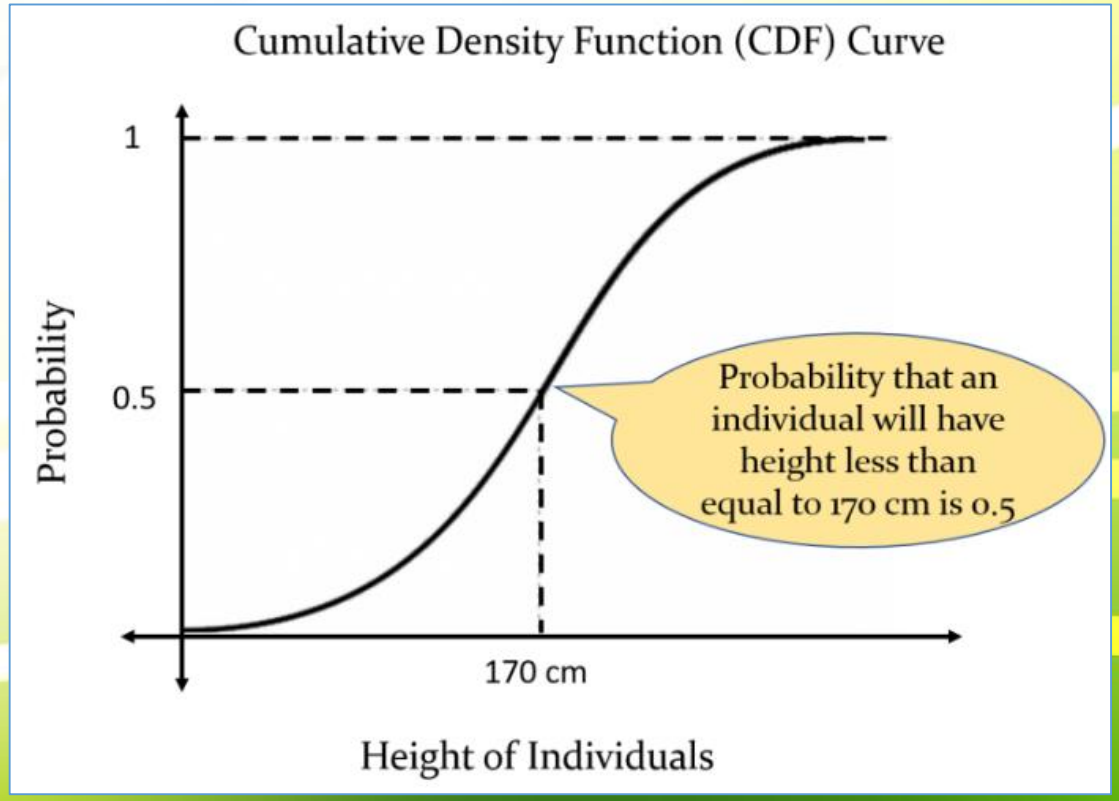
- Probability density generating functions;
  - Probability Mass function (PMF) for discrete distributions
  - Probability Density function (PDF) for continuous distributions.
- The total value of PMF and PDF over the entire domain is always equal to one.

Cumulative Distribution Function (CDF)

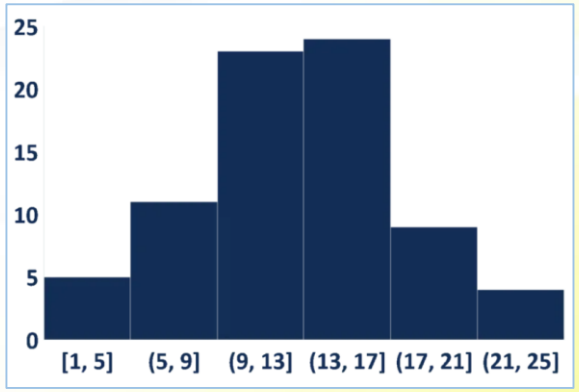
- PDF gives probability of a particular outcome.
- CDF gives probability of seeing an outcome less than or equal to particular value of random variable.
- CDFs are used to check how the probability has added up to a certain point.
- ***Example**, if P(X = 5) is probability of heads on flipping a coin as 5 then, P(X <= 5) denotes the cumulative probability of obtaining 1 to 5 heads.*

Probability Density





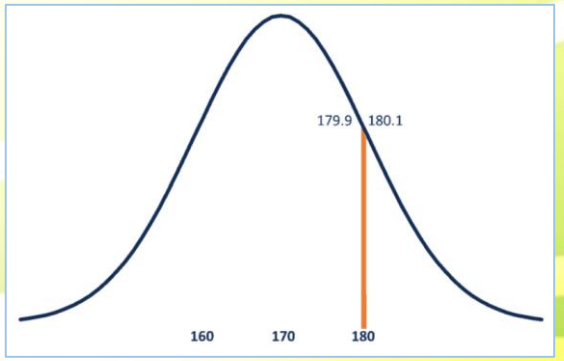
Probability Distribution



Two Types of distributions based on type of data generated by experiments.

Discrete probability distributions

- Probabilities of random variables can have discrete values as outcomes.
- ***Example**, possible values for random variable X for head occurrence in coin tossed twice {0, 1, 2}; but not any value from 0 to 2 like 0.1 or 1.6 etc.*
- Bernoulli, Binomial, Negative Binomial, Hypergeometric, etc..



Continuous probability distributions

- Probabilities of random variables can have any possible outcome.
- ***Example**, possible values for random variable X for weights of citizens can have any value like 34.5, 47.7, etc..*
- Normal, Student's T, Chi-square, Exponential, etc..

Discrete Distributions - Binomial

- **Binomial distribution** is a discrete probability distribution that gives only two possible results in an experiment, either **Success or Failure**.
- If we toss a coin, there could be only two possible outcomes: heads or tails.
- A single success/failure test is also called a **Bernoulli trial or Bernoulli experiment**.
- A series of outcomes is called a **Bernoulli process**.
- Probability through the trials remains constant and each trial is independent of the other.

Probability Mass function (PMF)

$$P(x:n,p) = {}^{n}C_{x}\, p^{x}\, (1-p)^{n-x}$$

$$_nC_r = \frac{n!}{r!(n-r)!}$$

Mean, μ = np
Variance, σ^2 = npq
Standard Deviation σ= √(npq)

n = the number of experiments
x = number of times for a specific outcome within n trials
p = Probability of Success in a single experiment
q = probability of failure, where q = 1-p

Discrete Distributions - Bernoulli

- For n = 1, i.e. a single experiment, the binomial distribution is a **Bernoulli distribution**.

Probability Mass function (PMF)

$$\text{PMF} = \begin{cases} p, & \text{Success} \\ 1 - p, & \text{Failure} \end{cases}$$

Example: flipping a coin once.
p is the probability of getting a head
q = 1 – p is probability of getting a tail.

Discrete Distributions – Negative Binomial

- Sometimes we want to check how many Bernoulli (single) trials we need to make in order to get a particular outcome.
- Number of *successes* is denoted by 'r'.
  - ***Example**, if we throw a dice and determine occurrence of 1 as a failure and all non-1's as successes.*
  - *Now, if we throw a dice frequently until 1 appears the third time, i.e., r = three success.*

Probability Mass function (PMF)

$$f(x; r, P) =^{x-1} C_{r-1} \times P^r \times (1-P)^{x-r}$$

x = Total number of trials.
r = Number of occurrences of success.
P = Probability of success on each occurrence.
q = Probability of failure on each occurrence.

Mean, μ = r/p
Variance, σ^2 = rq/p^2

Discrete Distributions – Geometric

- Geometric distribution is a special case of negative binomial distribution.
- Number of trials required for a single success.
- number of successes (r) = 1.

Probability Mass function (PMF)

$$P(X = x) = p \times q^{x-1}$$

x = Total number of trials.
r = 1
P = Probability of success on each occurrence.
q = Probability of failure on each occurrence.

Mean, μ = 1/p
Variance, $\sigma^2 = q/p^2$

Discrete Distributions – Geometric

Example: In an amusement fair, a competitor is entitled for a prize if he throws a ring on a peg from a certain distance. It is observed that only 30% of the competitors are able to do this. If someone is given 5 chances, what is the probability of his winning the prize when he has already missed 4 chances?

$$P(X = x) = p \times q^{x-1}$$

$$_nC_r = \frac{n!}{r!(n-r)!}$$

If someone has already missed four chances and has to win in fifth chance, then it is a probability experiment of getting the first success in 5 trials → geometric probability distribution.

p=30%=0.3
number of trials = x = 5

$$P(X = 5) = 0.3 \times (1 - 0.3)^{5-1},$$
$$= 0.3 \times (0.7)^4,$$
$$\approx 0.072$$
$$\approx 7.2\%$$

x = Total number of trials.
r = 1
P = Probability of success on each occurrence.
q = Probability of failure on each occurrence.

Discrete Distributions – Negative Binomial

Example: Robert is a football player. His success rate of goal hitting is 70%.

What is the probability that Robert hits his third goal on his fifth attempt?

$$f(x;r,P) = {}^{x-1}C_{r-1} \times P^r \times (1-P)^{x-r}$$

$${}_nC_r = \frac{n!}{r!(n-r)!}$$

Here probability of success, P = 0.70.
Number of trials, x = 5
number of successes r = 3.

$$f(x;r,P) = {}^{x-1}C_{r-1} \times P^r \times (1-P)^{x-r}$$

$$\implies f(5;3,0.7) = {}^{4}C_{2} \times 0.7^3 \times 0.3^2$$

$$= 6 \times 0.343 \times 0.09$$

$$= 0.18522$$

x = Total number of trials.
r = Number of occurrences of success.
P = Probability of success on each occurrence.
q = Probability of failure on each occurrence.

Discrete Distributions - Binomial

Example: If a coin is tossed 5 times, find the probability of:

(a) Exactly 2 heads

(b) At least 4 heads.

(c) At most 2 heads.

$$P(x:n,p) = {}^{n}C_{x}\, p^{x}\, (1-p)^{n-x}$$

$$_{n}C_{r} = \frac{n!}{r!(n-r)!}$$

(a) Repeated tossing of coin is an example of a Bernoulli trial.
Number of trials: n=5
Probability of head: p= 1/2
probability of tail, q =1/2
For exactly two heads: x=2

$P(x=2) = {}^{5}C2\ p^{2}\ q^{5-2} = 5!\ /\ 2!\ 3! \times (½)^{2} \times (½)^{3}$

$P(x=2) = 5/16$

(b) For at least four heads,

$x \geq 4,\ P(x \geq 4) = P(x = 4) + P(x=5)$

Hence,

$P(x = 4) = {}^{5}C4\ p^{4}\ q^{5-4} = 5!/4!\ 1! \times (½)^{4} \times (½)^{1} = 5/32$

$P(x = 5) = {}^{5}C5\ p^{5}\ q^{5-5} = (½)^{5} = 1/32$

Therefore,

$P(x \geq 4) = 5/32 + 1/32 = 6/32 = 3/16$

Solution: P (at most 2 heads) = $P(X \leq 2) = P\ (X = 0) + P\ (X = 1)$

$P(X = 0) = (½)^{5} = 1/32$

$P(X=1) = {}^{5}C_{1}\ (½)^{5} = 5/32$

Therefore,

$P(X \leq 2) = 1/32 + 5/32 = 3/16$

Discrete Distributions - Binomial

Example: Suppose, according to the latest police reports, 80% of all petty crimes are unresolved, and in your town, at least three of such petty crimes are committed. The three crimes are all independent of each other. From the given data, what is the probability that one of the three crimes will be resolved?

Number of fixed trials (n): 3 (Number of petty crimes)
Number of mutually exclusive outcomes: 2 (solved and unsolved)
The probability of success (p): 0.2 (20% of cases are solved)
Independent trials: Yes

$$P(x:n,p) = {}^{n}C_x \, p^x \, (1-p)^{n-x}$$

$${}_nC_r = \frac{n!}{r!(n-r)!}$$

Trial 1 = Solved 1st, unsolved 2nd, and unsolved 3rd

= 0.2 x 0. 8 x 0.8

= 0.128

Trial 2 = Unsolved 1st, solved 2nd, and unsolved 3rd

= 0.8 x 0.2 x 0.8

= 0.128

Trial 3 = Unsolved 1st, unsolved 2nd, and solved 3rd

= 0.8 x 0.8 x 0.2

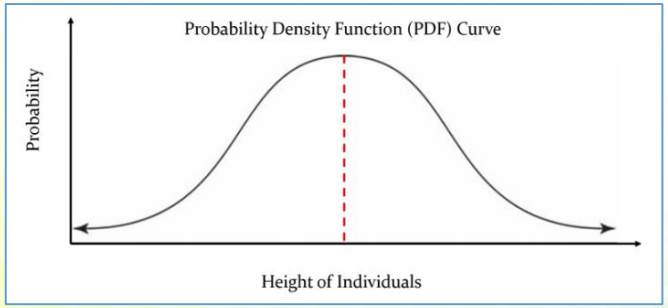
= 0.128

Total (for the three trials):

= 0.128 + 0.128 + 0.128

= 0.384

Discrete Distributions – Mean & Variance



$$\int_{-\infty}^{\infty} f(x)\,dx = 1.$$

$$P(x{:}n,p) = {}^{n}C_{x}\, p^{x}\, (1\text{-}p)^{n\text{-}x}$$

$$f(x;r,P) = {}^{x-1}C_{r-1} \times P^{r} \times (1-P)^{x-r}$$

$$P(X = x) = p \times q^{x-1}$$

	Discrete Random Variable	Continuous Random Variable
Mean (Expected Value)	$\mu = E(X) = \sum_{i=1}^{n} x f(x)$	$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$
Variance	$\sigma^2 = V(X) = \sum_{i=1}^{n} (x-\mu)^2 f(x)$	$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$
Standard Deviation (Standard Error)	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$

$$\mu = \mu_X = \mathrm{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)\,dx.$$

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - \mu^2 = \left(\int_{-\infty}^{\infty} x^2 \cdot f(x)\,dx\right) - \mu^2$$

Discrete Distributions – Mean & Variance

$$P(X = x) = p \times q^{x-1}$$

Mean, μ = 1/p
Variance, σ² = q/p²

	Discrete Random Variable	Continuous Random Variable
Mean (Expected Value)	$\mu = E(X) = \sum_{i=1}^{n} x f(x)$	$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$
Variance	$\sigma^2 = V(X) = \sum_{i=1}^{n} (x-\mu)^2 f(x)$	$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$
Standard Deviation (Standard Error)	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$

$$\begin{aligned}
\mathrm{E}(Y) &= \sum_{k=0}^{\infty} (1-p)^k p \cdot k \\
&= p \sum_{k=0}^{\infty} (1-p)^k k \\
&= p(1-p) \sum_{k=0}^{\infty} (1-p)^{k-1} \cdot k \\
&= p(1-p) \left[\frac{d}{dp} \left(-\sum_{k=0}^{\infty} (1-p)^k \right) \right] \\
&= p(1-p) \frac{d}{dp} \left(-\frac{1}{p} \right) = \frac{1-p}{p}.
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}[X^2] - \mathrm{E}[X]^2 \\
&= \frac{2-3p+p^2}{p^2} - \left(\frac{1-p}{p} \right)^2 \\
&= \frac{2-3p+p^2-(1-2p+p^2)}{p^2} \\
&= \frac{2-3p+p^2-1+2p-p^2}{p^2} \\
&= \frac{1-p}{p^2}
\end{aligned}$$

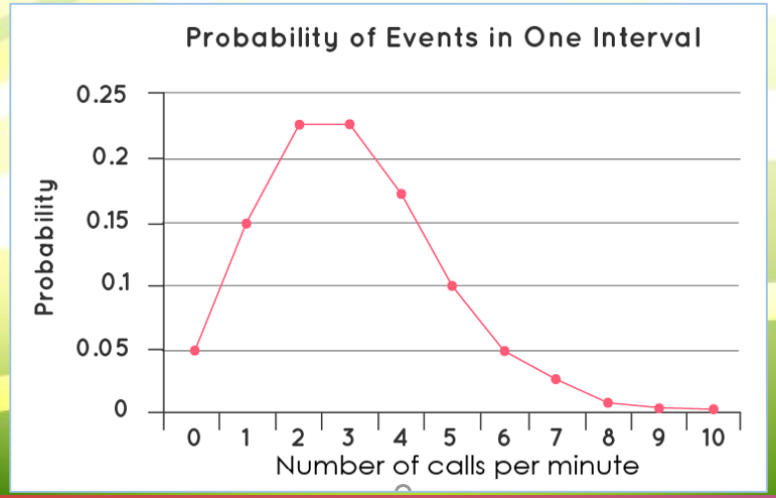
$$\mu = \mu_X = \mathrm{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx.$$

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - \mu^2 = \left(\int_{-\infty}^{\infty} x^2 \cdot f(x)\, dx \right) - \mu^2$$

Discrete Distributions - Poisson

- **Poisson distribution** measures how many times an event is likely to occur within “x” period of time.
- Poisson experiment is a statistical experiment that classifies the experiment into two categories (success/failure).
- Poisson distribution is a limiting process of the binomial distribution.
- λ : Poisson rate parameter that indicates expected value of average number of events in fixed time interval.

- **Example**, customer care center receives 100 calls per hour, 8 hours a day. Calls are independent of each other. There can be any number of calls per minute irrespective of the number of calls received in the previous minute.
- Probability of number of calls per minute has a Poisson probability distribution.



Discrete Distributions - Poisson

Poisson circulation is utilized as part of circumstances where happening's likelihood of an occasion is little, i.e., occasion once in a while happens.

- likelihood of faulty things in an assembling organization is little,
- likelihood of happening tremor in a year is little,
- mischance's likelihood on a street is little, etc..

Poisson distribution is used under certain conditions.

number of trials “n” tends to infinity
Probability of success “p” tends to zero
np = 1 is finite

Probability Mass function (PMF)

$$P(x, \lambda) = (e^{-\lambda} \lambda^{x})/x!$$

X : number of successes in experiment (Poisson random variable)
e : base of the logarithm
λ : Probability of success (average number of successes within a given range)

mean E(X) = λ.
mean and variance are equal.
E(X) = V(X)

Discrete Distributions - Poisson

Example: A producer of pins realized that on a normal 5% of his item is faulty. He offers pins in a parcel of 100 and insurances that not more than 4 pins will be flawed. What is the likelihood that a bundle will meet the ensured quality? [e^{-m}=0.0067]

$$n = 100, p = \frac{5}{100},$$

$$\Rightarrow np = 100 \times \frac{5}{100} = 5$$

$$P(x, \lambda) = (e^{-\lambda} \lambda^{x})/x!$$

X : number of successes in experiment
λ : Probability of success

Required probability = P [packet will meet the guarantee]
= P [packet contains up to 4 defectives]
= P (0) +P (1) +P (2) +P (3) +P (4)

$$= e^{-5} \cdot \frac{5^0}{0!} + e^{-5} \cdot \frac{5^1}{1!} + e^{-5} \cdot \frac{5^2}{2!} + e^{-5} \cdot \frac{5^3}{3!} + e^{-5} \cdot \frac{5^4}{4!},$$

$$= e^{-5}[1 + \frac{5}{1} + \frac{25}{2} + \frac{125}{6} + \frac{625}{24}],$$

$$= 0.0067 \times 65.374 = 0.438$$

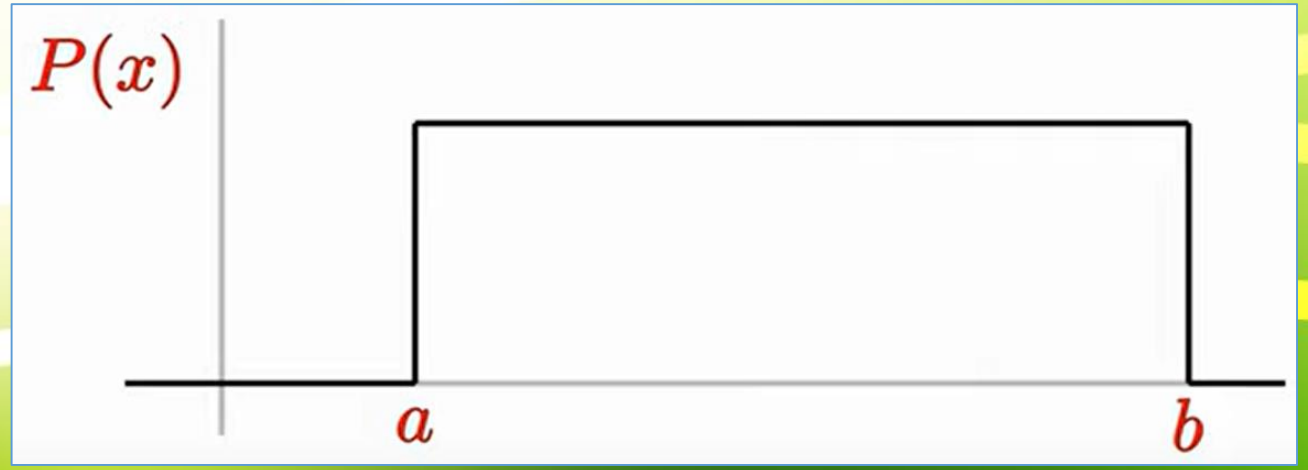
Continuous Distributions - Uniform

- Uniform distribution has both continuous and discrete forms.
- Continuous uniform distribution plots random variables whose values have equal probabilities of occurring.
- Example: flipping a fair die → all 6 outcomes are equally likely to happen → probability is constant.
- Uniform distribution (rectangular distribution) has constant probability.

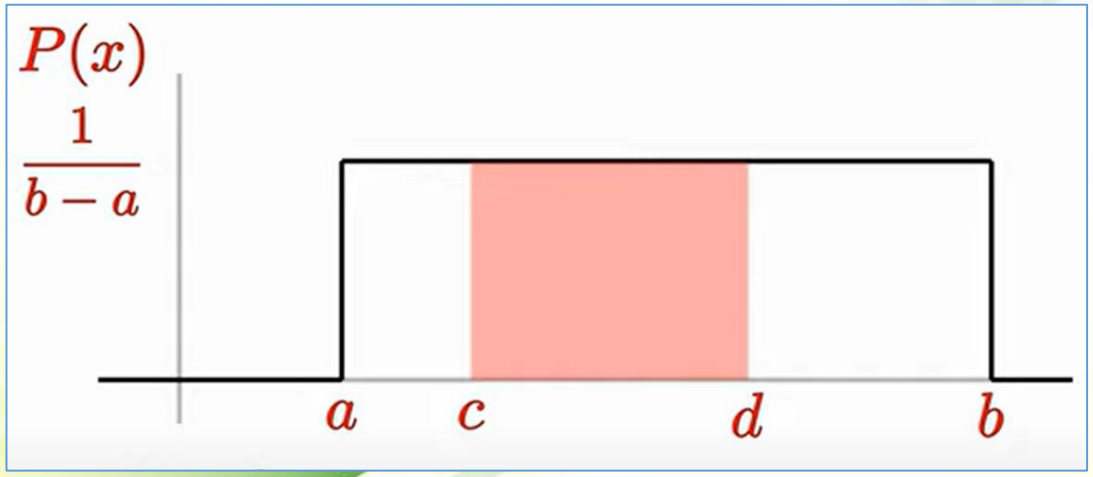
Probability Density function (PDF)

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b] \\ 0, & otherwise \end{cases}$$

where, a: minimum value
b: maximum value.



Continuous Distributions - Uniform




```
P(c ≤ x ≤ d) = (d - c) / (b - a)
```

$$\mu = \frac{a+b}{2}$$

$$\sigma = \sqrt{\frac{(b-a)^2}{12}}$$

where, a: minimum value
b: maximum value.

Uniform Distributions – Mean & Variance

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & x \in [a,b] \\ 0, & otherwise \end{cases}$$

$$\begin{aligned} \mathrm{E}(X) &= \int_{-\infty}^{a} 0x\,\mathrm{d}x + \int_{a}^{b} \frac{x}{b-a}\,\mathrm{d}x + \int_{b}^{\infty} 0x\,\mathrm{d}x \\ &= \left[\frac{x^2}{2(b-a)}\right]_a^b \\ &= \frac{b^2-a^2}{2(b-a)} \\ &= \frac{(b-a)(b+a)}{2(b-a)} \\ &= \frac{a+b}{2} \end{aligned}$$

	Discrete Random Variable	Continuous Random Variable
Mean (Expected Value)	$\mu = E(X) = \sum_{i=1}^{n} xf(x)$	$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$
Variance	$\sigma^2 = V(X) = \sum_{i=1}^{n} (x-\mu)^2 f(x)$	$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx$
Standard Deviation (Standard Error)	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1.$$

$$\mu = \mu_X = \mathrm{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)\,dx.$$

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - \mu^2 = \left(\int_{-\infty}^{\infty} x^2 \cdot f(x)\,dx\right) - \mu^2$$

Uniform Distributions – Mean & Variance

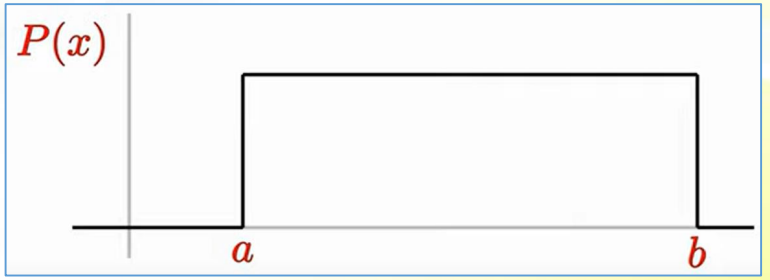
$$f(x) = \begin{cases} \dfrac{1}{b-a}, & x \in [a,b] \\ 0, & otherwise \end{cases}$$

$$\mu = \mu_X = \mathrm{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)\,dx.$$

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - \mu^2 = \left(\int_{-\infty}^{\infty} x^2 \cdot f(x)\,dx \right) - \mu^2$$

$$\begin{aligned}
\mathrm{var}\,(X) &= \int_{-\infty}^{a} 0x^2\,\mathrm{d}x + \int_{a}^{b} \frac{x^2}{b-a}\,\mathrm{d}x + \int_{b}^{\infty} 0x^2\,\mathrm{d}x - \frac{(a+b)^2}{4} \\
&= \left[\frac{x^3}{3\,(b-a)} \right]_a^b - \frac{(a+b)^2}{4} \\
&= \frac{b^3 - a^3}{3\,(b-a)} - \frac{(a+b)^2}{4} \\
&= \frac{4\,(b-a)\left(a^2+ab+b^2\right)}{12\,(b-a)} - \frac{3(a+b)^2}{12} \\
&= \frac{4a^2+4ab+4b^2-3a^2-6ab-3b^2}{12} \\
&= \frac{b^2-2ab+a^2}{12} \\
&= \frac{(b-a)^2}{12}
\end{aligned}$$

Continuous Distributions - Uniform



Bus is uniformly late between 2 and 10 minutes.

(a) How long can you expect to wait? With what S.D.?

(b) If its >7 mins late, you will be late for work. Whats the probability that you will be late?

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & x \in [a,b] \\ 0, & otherwise \end{cases}$$

$$\mu = \frac{a+b}{2}$$

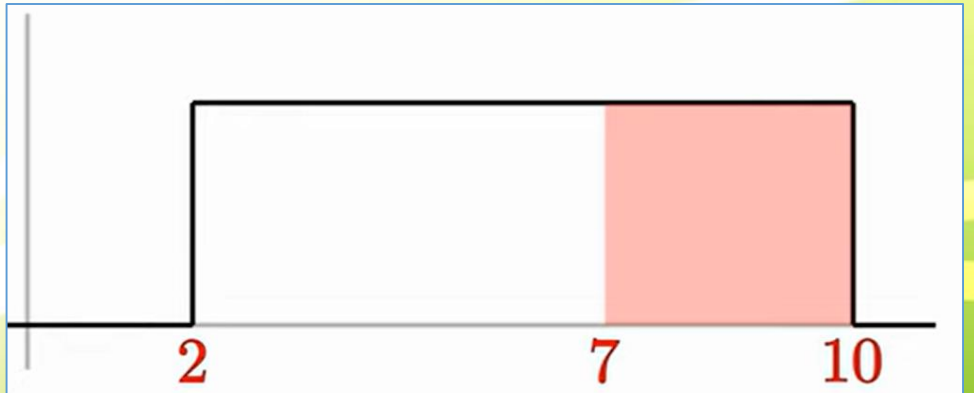
$$\sigma = \sqrt{\frac{(b-a)^2}{12}}$$

```
P(c ≤ x ≤ d) = (d - c) / (b - a)
```

Given, a = 2, b = 10

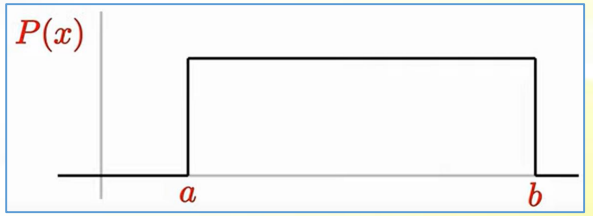
$$\mu = \frac{2+10}{2} = \frac{12}{2} = 6$$

$$\sigma = \sqrt{\frac{(10-2)^2}{12}} = \sqrt{5.33} = 2.31$$



$$P(7 \le X \le 10) = \frac{10-7}{10-2} = 0.375$$

Continuous Distributions - Uniform



The average weight gained by a person over the winter months is uniformly distributed and ranges from 0 to 30 lbs..

(a) How much weight Rhoit may expect to gain? With what S.D.?

(b) Find the probability of a person that he will gain between 10 and 15lbs in the winter months?

(c) Determine P(X ≤ 10) for the above-given question.

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b] \\ 0, & otherwise \end{cases}$$

`P(c ≤ x ≤ d) = (d - c) / (b - a)`

Given, a = 0, b = 30

$$\text{Probability} = 5 \times \frac{1}{30} = \frac{5}{30} = \frac{1}{6}.$$

$$10 \times \frac{1}{30} = \frac{10}{30} = \frac{1}{3}.$$

$$\mu = \frac{a+b}{2}$$

$$\sigma = \sqrt{\frac{(b-a)^2}{12}}$$

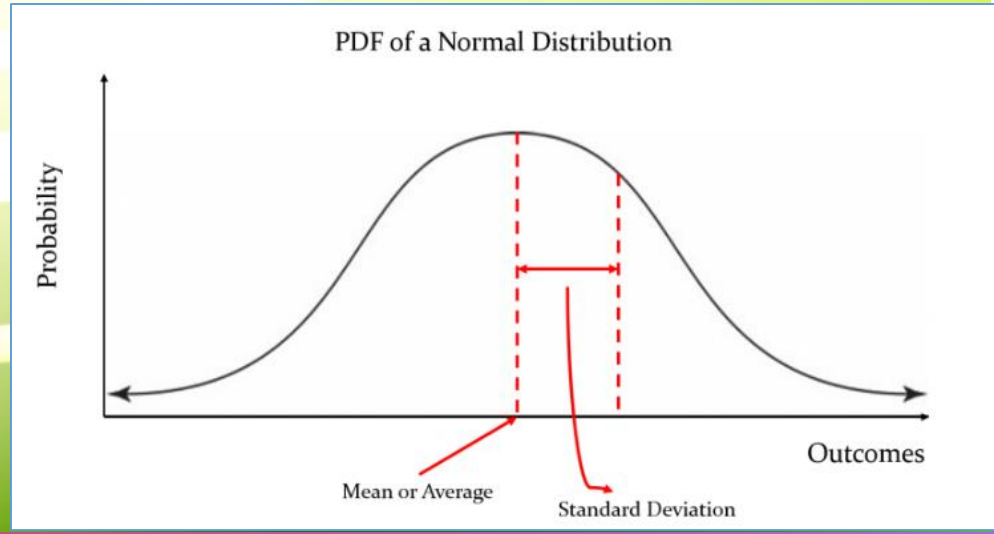
Continuous Distributions - Normal

- Most commonly found distribution.
- Many continuous distributions often reach normal distribution given a large enough sample.
- Mean has highest probability; all other values are distributed equally on either side of mean in a symmetric fashion.
- **Standard normal distribution**: mean is 0 and standard deviation of 1.
- When variable X follows normal distribution, with mean μ and variance σ^2: **$X \sim N(n,\sigma^2)$**

Probability Density function (PDF)

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

where **μ : mean of random variable X**
σ : standard deviation
e: 2.72 **π: 3.14**



Continuous Distributions - Normal

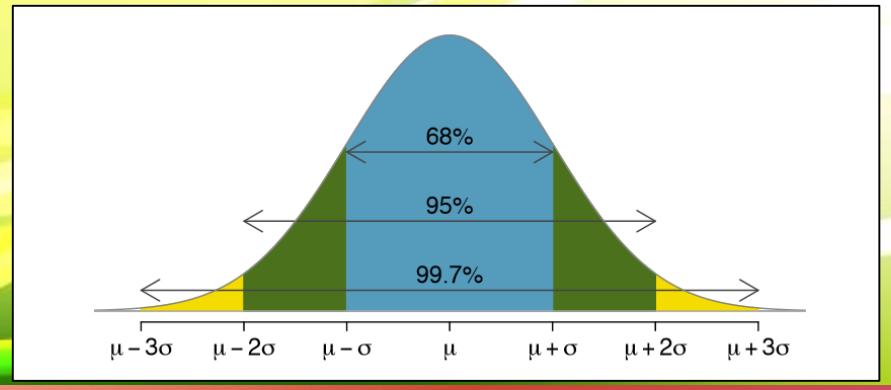
- **Normal/Gaussian/bell-shaped distribution**: symmetrical around its mean.
- Most observations cluster around central peak
- Probabilities for values further away from mean taper off (equally) in both directions.
- Extreme values in both tails of distribution are similarly unlikely.
- It follows that 68% of values are 1 standard deviation away,

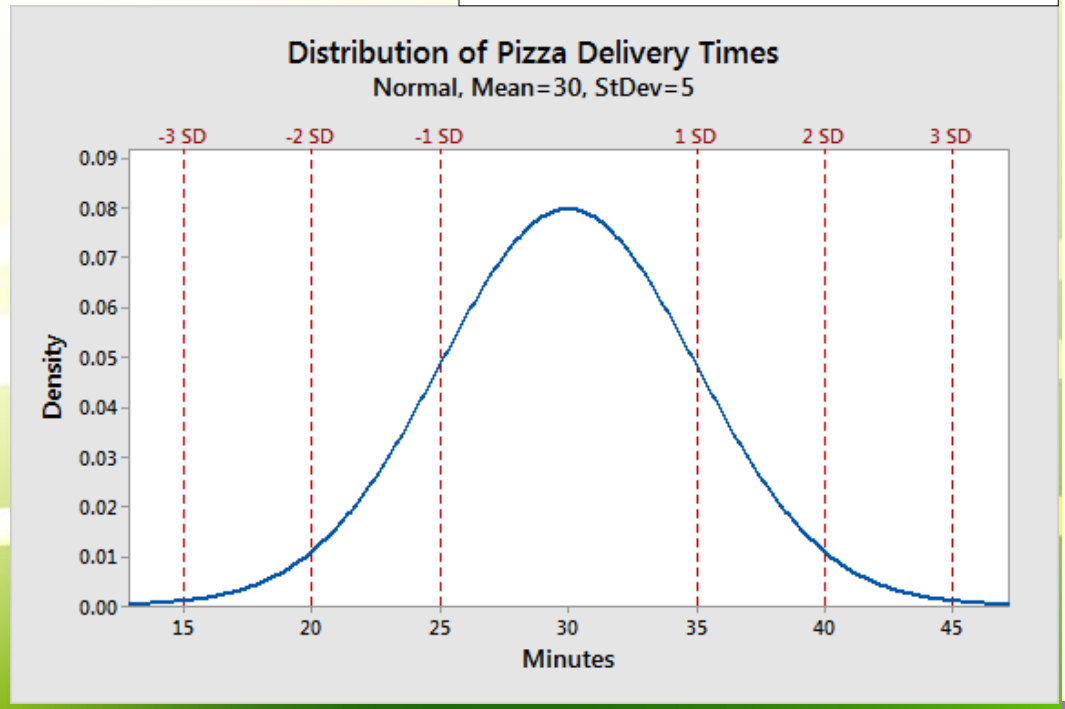
  95% percent of them are 2 standard deviations away, and

  99.7% are 3 standard deviations away from the mean.

  (68-95-99.7 rule)

Mean +/- standard deviations	Percentage of data contained
1	68%
2	95%
3	99.7%





Continuous Distributions – Standard Normal

- Convert normal distributions to **standard normal distribution**.

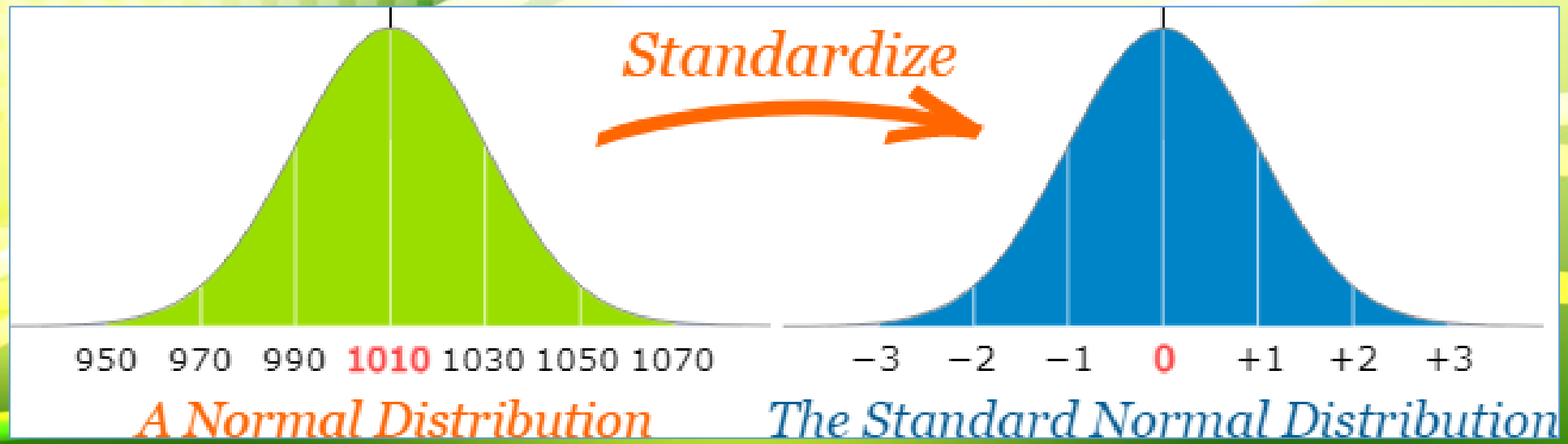
z-score

$$Z = \frac{X - \mu}{\sigma}$$

where **μ : mean of random variable**
σ : standard deviation.

P(x) → p(z)

In standard normal distribution;
μ : 0
σ : 1



Continuous Distributions - Normal

Positive Z-score Table

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

Continuous Distributions - Normal

Negative Z-score Table

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414
-0.1	.46017	.45620	.45224	.44828	.44433	.44034	.43640	.43251	.42858	.42465
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-1	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08692	.08534	.08379	.08226
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-2	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831

Continuous Distributions - Normal

Calculate the probability function value for random variable 3 in a normal distribution with population mean 4, standard deviation 2.

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

x = 3

Mean = μ = 4

Standard Deviation = σ = 2

$$f(x) = \frac{1}{\sqrt{2\pi(2)^2}} e^{\frac{-(3-4)^2}{2(2)^2}}$$

$$f(x) = \frac{1}{2\sqrt{2\pi}} e^{\frac{-1}{8}}$$

$$f(x) = 0.19947 \times e^{-0.125}$$

$$f(x) = 0.19947 \times 0.882496$$

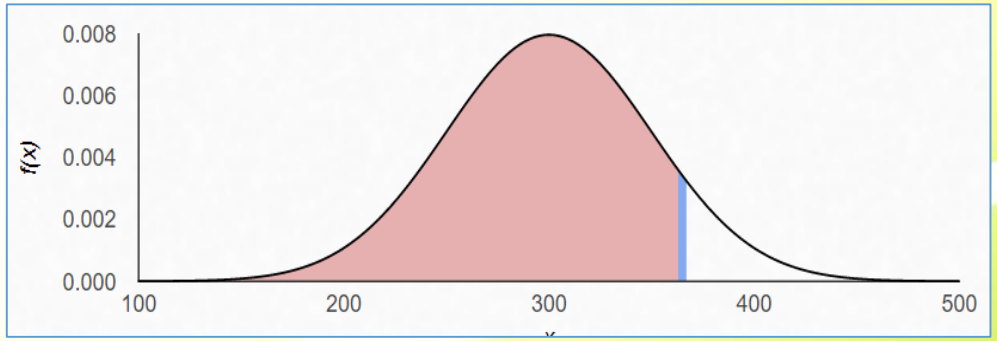
$$f(x) = 0.17603$$

Continuous Distributions - Normal

An average electric bulb lasts for 300 days with a standard deviation of 50 days. Assume that bulb life is normally distributed, what is the probability that the electric bulb will last at most 365 days

$$Z = \frac{X - \mu}{\sigma}$$

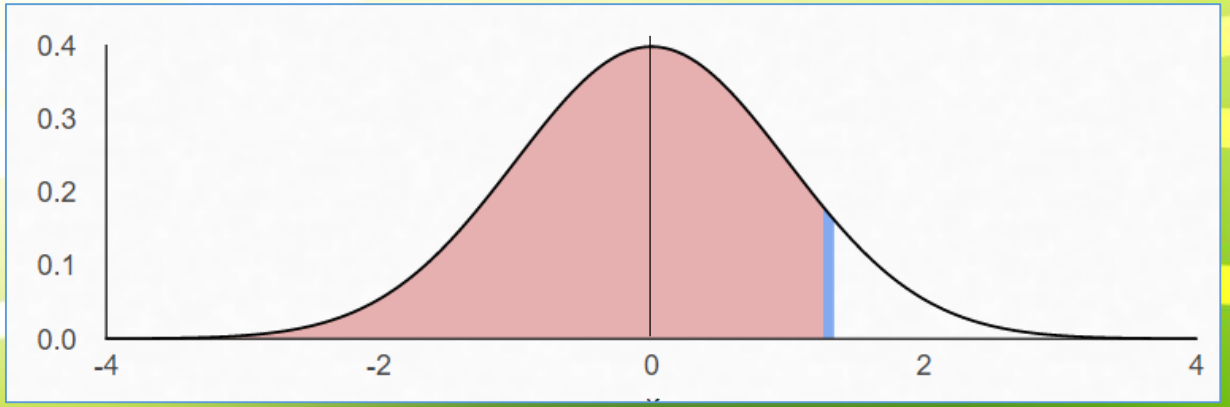
Given, Mean score = 300 days
Standard deviation = 50 days
Probability that bulb life is less than or equals to (random variable value) 365 days.



P (x < 365) = P(z < 1.3) = 0.90 **or**
P (x < 365) = P(z < 1.3) = P(z=0) + P(z=1.3) = 0.5 + 0.40 = 0.90

90% probability that bulb will burn out within 365 days.

$$\text{Z score} = \frac{x - \mu}{\sigma} = \frac{365 - 300}{50} = 1.3$$



Continuous Distributions - Normal

Scores of an exam are represented as N(65,9). Find the percentage of score

(a) Less than 54, (b) at least 80, (c) between 70-86

$$Z = \frac{X - \mu}{\sigma}$$

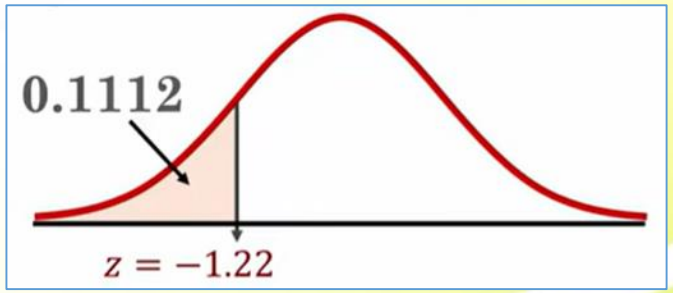
Given, Mean score = 65
Standard deviation =9

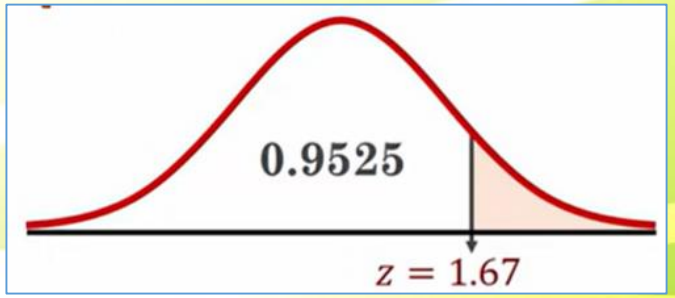
$$x = 54$$
$$z = \frac{x - \mu}{\sigma} = \frac{54 - 65}{9}$$
$$z = -1.2222...$$

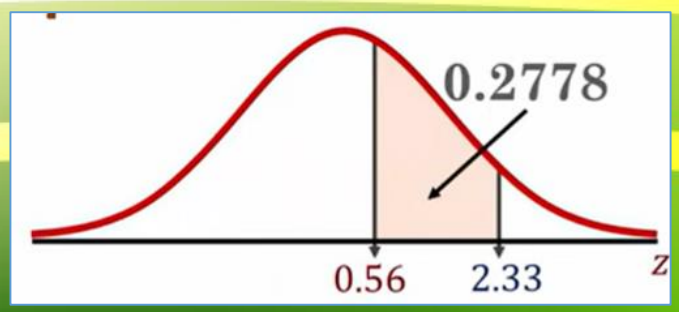
$$x = 80$$
$$z = \frac{x - \mu}{\sigma} = \frac{80 - 65}{9}$$
$$z = 1.67$$

$$x = 70 \quad z = \frac{x - \mu}{\sigma} = \frac{70 - 65}{9} = 0.56$$

$$x = 86 \quad z = \frac{x - \mu}{\sigma} = \frac{86 - 65}{9} = 2.33$$







P (x < 54) = P(z < -1.22) = 0.1112 = 11%

P (x >= 80) = P(x >80) = P(z > 1.67) = 0.9525 = 95%

P (70 < x < 86) = P(0.56 < z < 2.33) = P(z < 2.33) - P(0.56 < z)
= 0.9901 – 0.7123 = 0.2778 = 28%

Continuous Distributions - Normal

For some computers, the time period between charges of the battery is normally distributed with a mean of 50 hours and a standard deviation of 15 hours. Rohan has one of these computers and needs to know the probability that the time period will be between 50 and 70 hours.

$$Z = \frac{X - \mu}{\sigma}$$

Let x be the random variable that represents the time period.
Given Mean, μ= 50
and standard deviation, σ = 15
To find: Probability that x is between 50 and 70 or P(50< x < 70)

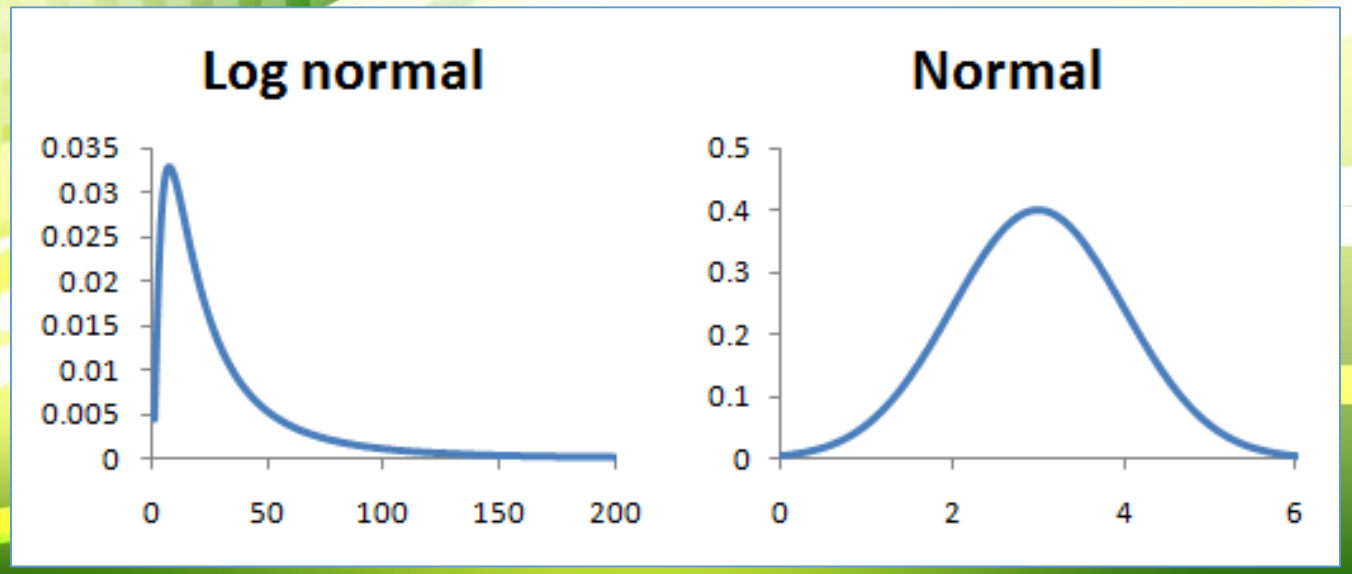
By using the transformation equation,;
z = (X – μ) / σ

For x = 50 , z = (50 – 50) / 15 = 0
For x = 70 , z = (70 – 50) / 15 = 1.33

P(50< x < 70) = P(0< z < 1.33) = [area to the left of z = 1.33] – [area to the left of z = 0]
= P(0< z < 1.33) = 0.9082 – 0.5 = 0.4082

Continuous Distributions - Lognormal

- **Log-normal distribution:** continuous probability distribution of random variable whose logarithm follows the pattern of normal distribution.
- It represents phenomena whose relative growth rate is independent of size.
- Unlike a standard normal distribution, log-normal distribution is not symmetrical.
- Distribution shape is mostly normal with a significant positive skew in one direction.



Continuous Distributions – Student's T

- Similar to normal distribution.
- Difference is that tails of the distribution are thicker.
- Used when sample size is small *(less than 30)* and population variance is not known.
- Defined by the **degrees of freedom**(p) → sample size minus 1 (n – 1).
- As sample size increases, degrees of freedom increases → t-distribution approaches normal distribution and tails become narrower and curve gets closer to mean.

Continuous Distributions – Chi-square

- This distribution is equal to the sum of squares of p normal random variables.

(p: number of degrees of freedom)

- Like t-distribution, as degrees of freedom increase, distribution gradually approaches normal distribution.

Chi-squared distribution is widely used for the following:

- Estimation of Confidence interval for a population standard deviation of a normal distribution using a sample standard deviation.
- To check independence of two criteria of classification of multiple qualitative variables.
- To check the relationships between categorical variables.
- To study the sample variance where the underlying distribution is normal.
- To test deviations of differences between expected and observed frequencies.
- To conduct chi-square test (a goodness of fit test).

Joint Probability Distributions

PROBABILITY: likelihood for one variable.

Probability of Event A or Event B.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- **Marginal probability P(A):** probability of an event irrespective of outcome of another variable
- **Joint probability P(A,B)**: probability that two events will both occur (likelihood of two events occurring together).

$$\text{Joint Probability} = P(A \cap B) = P(A) \times P(B)$$

- **Conditional probability:** probability of one event occurring given that second event has happened.

$$P(X/Y) = \frac{P(X \cap Y)}{P(Y)}, \text{ if } P(Y) \neq 0$$

$$P(Y/X) = \frac{P(Y \cap X)}{P(X)}, \text{ if } P(X) \neq 0$$

Joint Probability Distributions

Joint probability density function of X;

$$P(X_1 \in [a_1, b_1], \ldots, X_K \in [a_K, b_K]) = \int_{a_1}^{b_1} \ldots \int_{a_K}^{b_K} f_X(x_1, \ldots, x_K) dx_K \ldots dx_1$$

- Let X be a continuous random vector formed by random variables x_1, … x_k

  for any choice of the intervals $[a_1, b_1], \ldots, [a_K, b_K]$

$$P(X_1 \in [a_1, b_1], \ldots, X_K \in [a_K, b_K])$$

To solve double integral;

1. in first step, compute the inner integral
2. in second step, compute integral of first step solution.

$$\int_{a_1}^{b_1} \int_{a_2}^{b_2} f_X(x_1, x_2) dx_2 dx_1$$

$$I(x_1) = \int_{a_2}^{b_2} f_X(x_1, x_2) dx_2$$

$$\int_{a_1}^{b_1} \int_{a_2}^{b_2} f_X(x_1, x_2) dx_2 dx_1 = \int_{a_1}^{b_1} I(x_1) dx_1$$

Joint Probability Distributions

Joint pdf is equal to 1 if both entries of the vector belong to the interval [0,1] and it is equal to 0 otherwise. ***(Consider the following joint pdf of two variables.)***

$$f_X(x_1,x_2) = \begin{cases} 1 & \text{if } x_1 \in [0,1] \text{ and } x_2 \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

Compute the probability that both entries will be less than or equal to 1/2.

$$\mathrm{P}(X_1 \le 1/2, X_2 \le 1/2)$$
$$= \mathrm{P}(X_1 \in (-\infty, 1/2], X_2 \in (-\infty, 1/2])$$

$$= \int_{-\infty}^{1/2}\int_{-\infty}^{1/2} f_X(x_1,x_2)\,dx_2\,dx_1$$
$$= \int_{0}^{1/2}\int_{0}^{1/2} dx_2\,dx_1$$
$$= \int_{0}^{1/2} [x_2]_0^{1/2}\,dx_1$$

$$= \int_0^{1/2} \frac{1}{2}\,dx_1$$
$$= \frac{1}{2}[x_1]_0^{1/2}$$
$$= \frac{1}{2}\frac{1}{2} = \frac{1}{4}$$

$$\mathrm{P}(X_1 \in [a_1,b_1],\ldots,X_K \in [a_K,b_K]) = \int_{a_1}^{b_1}\cdots\int_{a_K}^{b_K} f_X(x_1,\ldots,x_K)\,dx_K\ldots dx_1$$

$$\mathrm{P}(X_1 \in [a_1,b_1],\ldots,X_K \in [a_K,b_K])$$

Joint Probability Distributions

Let X and Y be jointly continuous random variables with joint PDF

1. *Are X and Y independent?*
2. *Find P(X>Y)*

$$f_{X,Y}(x,y) = \begin{cases} 6e^{-(2x+3y)} & x,y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathrm{P}(X_1 \in [a_1,b_1],\ldots,X_K \in [a_K,b_K]) = \int_{a_1}^{b_1}\ldots\int_{a_K}^{b_K} f_X(x_1,\ldots,x_K)dx_K\ldots dx_1$$

$$\mathrm{P}(X_1 \in [a_1,b_1],\ldots,X_K \in [a_K,b_K])$$

$$f_X(x) = 2e^{-2x}u(x), \qquad f_Y(y) = 3e^{-3y}u(y).$$

X and Y are independent

X and Y are independent, if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

$$P(X>Y) = \int_0^\infty \int_y^\infty 6e^{-(2x+3y)}\,dxdy$$
$$= \int_0^\infty 3e^{-5y}dy$$
$$= \frac{3}{5}.$$

$$\int e^x dx = e^x \qquad \int e^{-x}dx = -e^{-x}$$
$$\int e^{ax}dx = \frac{1}{a}e^{ax}$$

Joint Probability Distributions

Let X and Y be jointly continuous random variables with joint PDF

$$f(x,y) = \frac{x+y}{3}, \qquad 0<x<2, \quad 0<y<1,$$

zero otherwise.

1. *Find P(X>Y)*
2. *Find the marginal PDFs fX(x) and fY(y)*
3. *Are X and Y independent? Find Cov(x,y).*

$$P(X_1 \in [a_1,b_1],\ldots,X_K \in [a_K,b_K]) = \int_{a_1}^{b_1}\cdots\int_{a_K}^{b_K} f_X(x_1,\ldots,x_K)\,dx_K\ldots dx_1$$

$$P(X>Y) = 1-\int_0^1\left(\int_0^y \frac{x+y}{3}\,dx\right)dy$$

$$= 1-\int_0^1\left(\frac{y^2}{6}+\frac{y^2}{3}\right)dy$$

$$= 1-\int_0^1 \frac{y^2}{2}\,dy = 1-\frac{1}{6} = \mathbf{\frac{5}{6}}.$$

Joint Probability Distributions

Let X and Y be jointly continuous random variables with joint PDF

$$f(x,y) = \frac{x+y}{3}, \quad 0<x<2, \quad 0<y<1,$$
$$\text{zero otherwise.}$$

1. *Find P(X>Y)*
2. *Find the marginal PDFs fX(x) and fY(y)*
3. *Are X and Y independent? Find Cov(x,y).*

$$f_X(x) = \int_0^1 \frac{x+y}{3}\,dy = \left(\frac{xy}{3}+\frac{y^2}{6}\right)\Big|_0^1 = \frac{2x+1}{6}, \qquad 0<x<2.$$

$$f_Y(y) = \int_0^2 \frac{x+y}{3}\,dx = \left(\frac{x^2}{6}+\frac{xy}{3}\right)\Big|_0^2 = \frac{2+2y}{3}, \qquad 0<y<1.$$

Since $f(x,y) \neq f_X(x)\cdot f_Y(y)$, X and Y are **NOT independent**.

X and Y are independent, if

$$f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dx$$

Joint Probability Distributions

Let X and Y be jointly continuous random variables with joint PDF

1. ***Find P(X>Y)***
2. ***Find the marginal PDFs fX(x) and fY(y)***
3. ***Are X and Y independent? Find Cov(x,y).***

$$f(x,y) = \frac{x+y}{3}, \qquad 0<x<2, \quad 0<y<1,$$

	Discrete Random Variable	Continuous Random Variable
Mean (Expected Value)	$\mu = E(X) = \sum_{i=1}^{n} x f(x)$	$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$
Variance	$\sigma^2 = V(X) = \sum_{i=1}^{n} (x-\mu)^2 f(x)$	$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$
Standard Deviation (Standard Error)	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$

$$\mu = \mu_X = \mathrm{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x)\,dx.$$

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - \mu^2 = \left(\int_{-\infty}^{\infty} x^2 \cdot f(x)\,dx\right) - \mu^2$$

$$\mathrm{E}(\mathrm{X}) = \int_{-\infty}^{\infty} x \cdot f_{\mathrm{X}}(x)\,dx = \int_{0}^{2} x \cdot \frac{2x+1}{6}\,dx = \left(\frac{x^3}{9} + \frac{x^2}{12}\right)\Big|_0^2 = \frac{11}{9}.$$

$$\mathrm{E}(\mathrm{Y}) = \int_{-\infty}^{\infty} y \cdot f_{\mathrm{Y}}(y)\,dy = \int_{0}^{1} y \cdot \frac{2+2y}{3}\,dy = \left(\frac{y^2}{3} + \frac{y^3}{9}\right)\Big|_0^1 = \frac{5}{9}.$$

$$\mathrm{E}(\mathrm{XY}) = \int_{0}^{2}\left(\int_{0}^{1} xy \cdot \frac{x+y}{3}\,dy\right)dx = \int_{0}^{2}\left(\frac{x^2}{6} + \frac{x}{9}\right)dx = \left(\frac{x^3}{18} + \frac{x^2}{18}\right)\Big|_0^2 = \frac{2}{3}.$$

$$\mathrm{Cov}(\mathrm{X},\mathrm{Y}) = \mathrm{E}(\mathrm{XY}) - \mathrm{E}(\mathrm{X}) \times \mathrm{E}(\mathrm{Y}) = \frac{2}{3} - \frac{11}{9}\cdot\frac{5}{9} = -\frac{\mathbf{1}}{\mathbf{81}} \approx -0.012345679.$$

Transformation of random variables

Let the probability density function of X be given by:

$$f(x) = \begin{cases} 6x(1-x), & 0 < x < 1 \\ 0 \ otherwise \end{cases}$$

Find the probability density of $Y = X^3$

$$\begin{aligned} G(y) &= P(Y \leq y) \\ &= P(X^3 \leq y) \\ &= P\left(X \leq y^{1/3}\right) \\ &= \int_0^{y^{1/3}} 6x(1-x)\,dx \\ &= \int_0^{y^{1/3}} (6x - 6x^2)\,dx \\ &= (3x^2 - 2x^3)\,|_0^{y^{1/3}} \\ &= 3y^{2/3} - 2y \end{aligned}$$

$$\begin{aligned} g(y) &= \frac{dG(y)}{dy} \\ &= \frac{d}{dy}\left(3y^{2/3} - 2y\right) \\ &= 2y^{-1/3} - 2 \\ &= 2(y^{-1/3} - 1),\ 0 < y < 1 \end{aligned}$$

Transformation of random variables

Let the probability density function of X_1 and X_2 be given by:

Find the probability density of $Y = X_1 + X_2$

$$f(x_1, x_2) = \begin{cases} 2e^{-x_1 - 2x_2}, & x_1 > 0, x_2 > 0 \\ 0 \; otherwise \end{cases}$$

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= \int_0^y \int_0^{y-x_2} 2e^{-x_1 - 2x_2} \, dx_1 \, dx_2 \\ &= \int_0^y -2e^{-x_1 - 2x_2} \Big|_0^{y-x_2} \, dx_2 \\ &= \int_0^y \left[\left(-2e^{-y + x_2 - 2x_2} \right) - \left(-2e^{-2x2} \right) \right] dx_2 \\ &= \int_0^y -2e^{-y - x_2} + 2e^{-2x2} \, dx_2 \\ &= \int_0^y 2e^{-2x2} - 2e^{-y - x_2} \, dx_2 \end{aligned}$$

$$\begin{aligned} F(y) &= P(Y \leq y) \\ &= \int_0^y 2e^{-2x2} - 2e^{-y - x_2} \, dx_2 \\ &= -e^{-2x2} + 2e^{-y - x_2} \Big|_0^y \\ &= -e^{-2y} + 2e^{-y - y} - \left[-e^0 + 2e^{-y} \right] \\ &= e^{-2y} - 2e^{-y} + 1 \end{aligned}$$

$$\begin{aligned} F_Y(y) &= \frac{dF(y)}{dy} \\ &= \frac{d}{dy} \left(e^{-2y} - 2e^{-y} + 1 \right) \\ &= -2e^{-2y} + 2e^{-y} \\ &= 2e^{-2y} \left(-1 + e^y \right) \end{aligned}$$

Monte Carlo Approximation

- Monte Carlo approximation is used for randomly sampling a probability distribution.
- Three main reasons to use Monte Carlo:
  - **Estimate density:** gather samples to approximate the distribution of a target function.
  - **Approximate a quantity:** such as the mean or variance of a distribution.
  - **Optimize a function:** locate a sample that maximizes or minimizes the target function.
- For large data, more the random trials performed, more accurate the approximated quantity will become.

$$I = \int_a^b h(x)g(x)dx$$

$Approximation(Average(X))$	$E(X) \approx \frac{1}{N}\sum_{n=1}^{N} x_n$

END !!