Machine Learning

What is Machine Learning?

- **Machine Learning** is the field of study that gives computers the capability to learn without being explicitly programmed.

- ***Machine learning (ML) is defined as a discipline of artificial intelligence (AI) that provides machines the ability to automatically learn from data and past experiences to identify patterns and make predictions with minimal human intervention.***

- Machine learning is focused on building systems that can learn from historical data, identify patterns, and make logical decisions with little to no human intervention

What is Machine Learning?

- Machine Learning(ML) can be explained as automating and improving the learning process of computers based on their experiences without being actually programmed.

- The process starts with feeding good quality data and then training our machines(computers) by building machine learning models using the data and different algorithms. The choice of algorithms depends on what type of data do we have and what kind of task we are trying to automate.
- **Example: Training of students during exams.**

What is learning?

- A computer program is said to *learn* from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks T, as measured by P , improves with experience E.
- Handwriting recognition learning problem
  - Task T : Recognizing and classifying handwritten words within images
  - Performance P : Percent of words correctly classified
  - Training experience E : A dataset of handwritten words with given classifications

- The quality of a machine learning model is dependent on two major aspects:

1. The quality of the input data: "garbage in, garbage out"

- Garbage in, garbage out, or GIGO, refers to the idea that in any system, the quality of output is determined by the quality of the input.
  - For example, if a mathematical equation is improperly stated, the answer is unlikely to be correct. Similarly, if incorrect data is used as input into a computer program, the output is unlikely to be correct or informative.

- **Volume:** the size and amounts of data that companies manage and analyze
- **Value:** the most important "V" from the perspective of the business, the value of data usually comes from insight discovery and pattern recognition that lead to more effective operations.
- **Variety:** the diversity and range of different data types, including unstructured data, semi-structured data and raw data
- **Velocity:** the speed at which companies receive, store and manage data – e.g., the specific number of social media posts or search queries received within a day, hour or other unit of time
- **Veracity:** the "truth" or accuracy of data and information assets, which often determines executive-level confidence

2. The model choice itself.

Why Machine Learning?

- Recent progress in algorithms and theory
- Growing flood of online data
- Computational power is available
- Increasing support from industries(trends in customer behavior and business operational patterns, as well as supports the development of new products)

Three niches for machine learning:

- Data mining : using historical data to improve decisions
  - medical records → medical knowledge
- Software applications we can't program by hand
  - autonomous driving
  - speech recognition
- Self customizing programs
  - Newsreader that learns user interests

How Machine Learning Works

- Machine Learning enables computers to learn from data and make predictions or decisions without explicit programming. The process involves several key steps:

1. **Data Collection:**

   The first step in Machine Learning is gathering relevant data representing the problem or task at hand. This data can be collected from various sources such as databases, sensors, or online platforms.

2. **Data Preprocessing:**

   Once the data is collected, it needs to be pre-processed to ensure its quality and suitability for training the model. This involves cleaning the data, handling missing values, and normalizing or transforming the data to a consistent format.

3. **Feature Extraction and Selection:**

   The collected data may contain many features or attributes in many cases. Feature extraction and selection involve identifying the most informative and relevant features contributing to the learning task.

4. **Model Training:**

   The training phase involves feeding the pre-processed data into a Machine Learning algorithm or model. The model learns from the data by adjusting its internal parameters based on the patterns and relationships it discovers.

5. **Model Evaluation:**

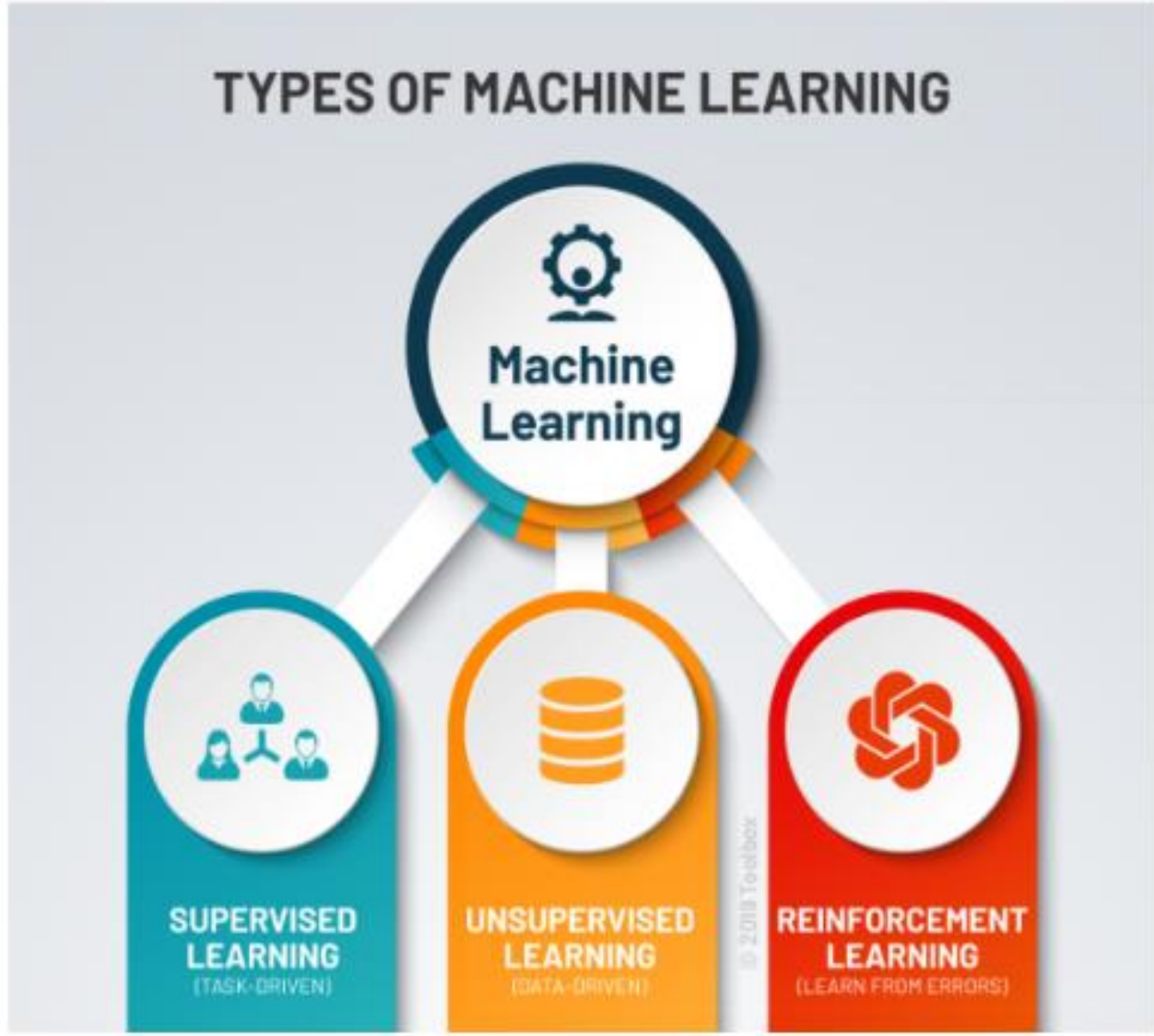
   The model must be evaluated to assess its performance and generalization ability after training it.

6. **Prediction or Decision Making:**

   Once the model is trained and evaluated, it can predict or decide on new, unseen data. The model takes input features and applies the learned patterns to generate the desired output or prediction.

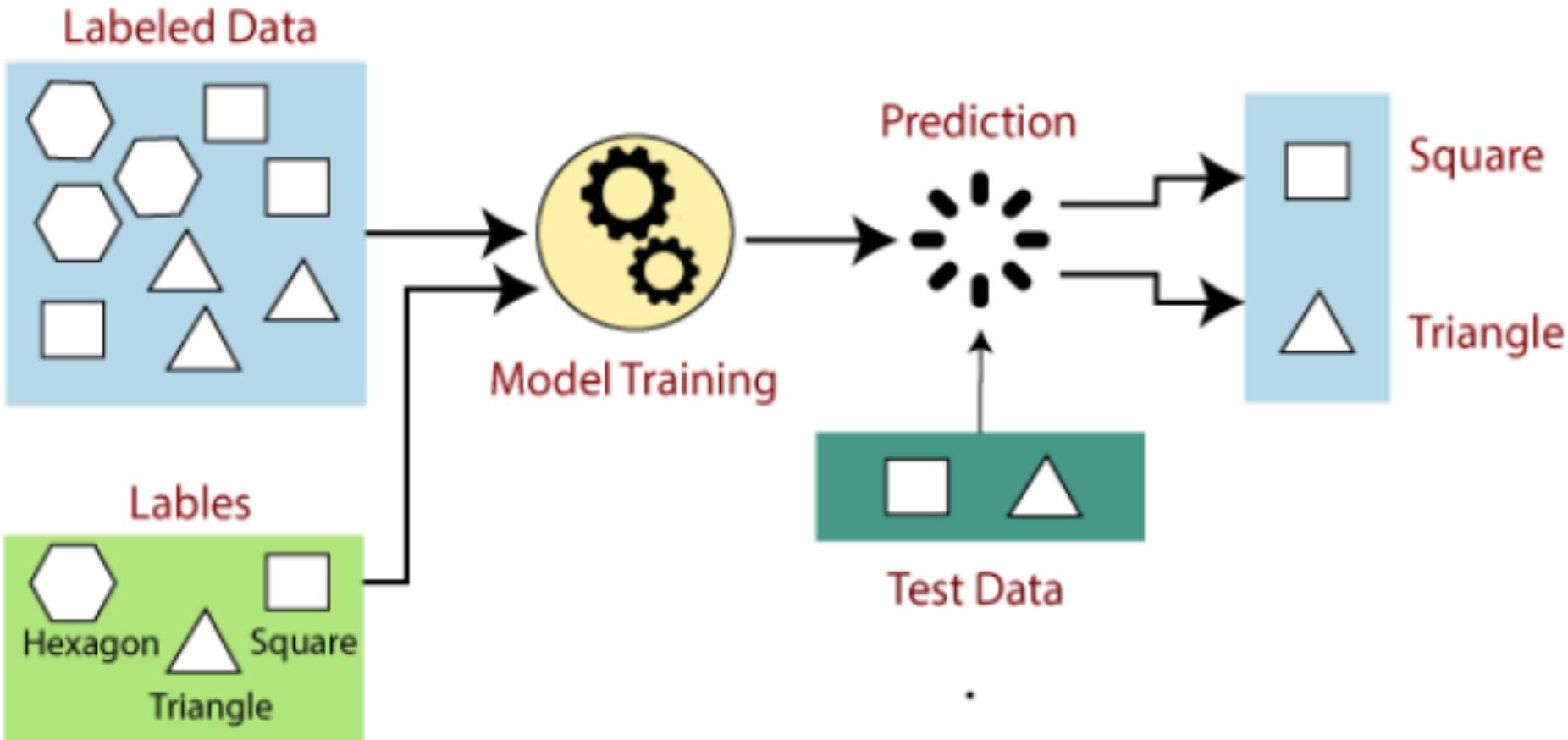
7. **Model Refinement and Iteration:**

   ML is an iterative process that involves refining the model based on their feedback and new dataset. If the model's performance is unsatisfactory and not accurate, then we can make adjustments by retraining the model with additional data, changing the algorithm, or tuning the model's parameters.


TYPES OF MACHINE LEARNING
Machine Learning
SUPERVISED LEARNING
(TASK-DRIVEN)
UNSUPERVISED LEARNING
(DATA-DRIVEN)
REINFORCEMENT LEARNING
(LEARN FROM ERRORS)

Supervised machine learning

- The model or algorithm is presented with example inputs and their desired outputs and then finding patterns and connections between the input and the output.
- The goal is to learn a general rule that maps inputs to outputs. The training process continues until the model achieves the desired level of accuracy on the training data.


Labeled Data
Lables
Hexagon
Square
Triangle
Model Training
Prediction
Test Data
Square
Triangle

Types of Supervised Machine Algorithm

- **Regression**: It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc.

- Algorithms:
  - Linear Regression
  - Regression Trees
  - Non-Linear Regression
  - Bayesian Linear Regression
  - Polynomial Regression

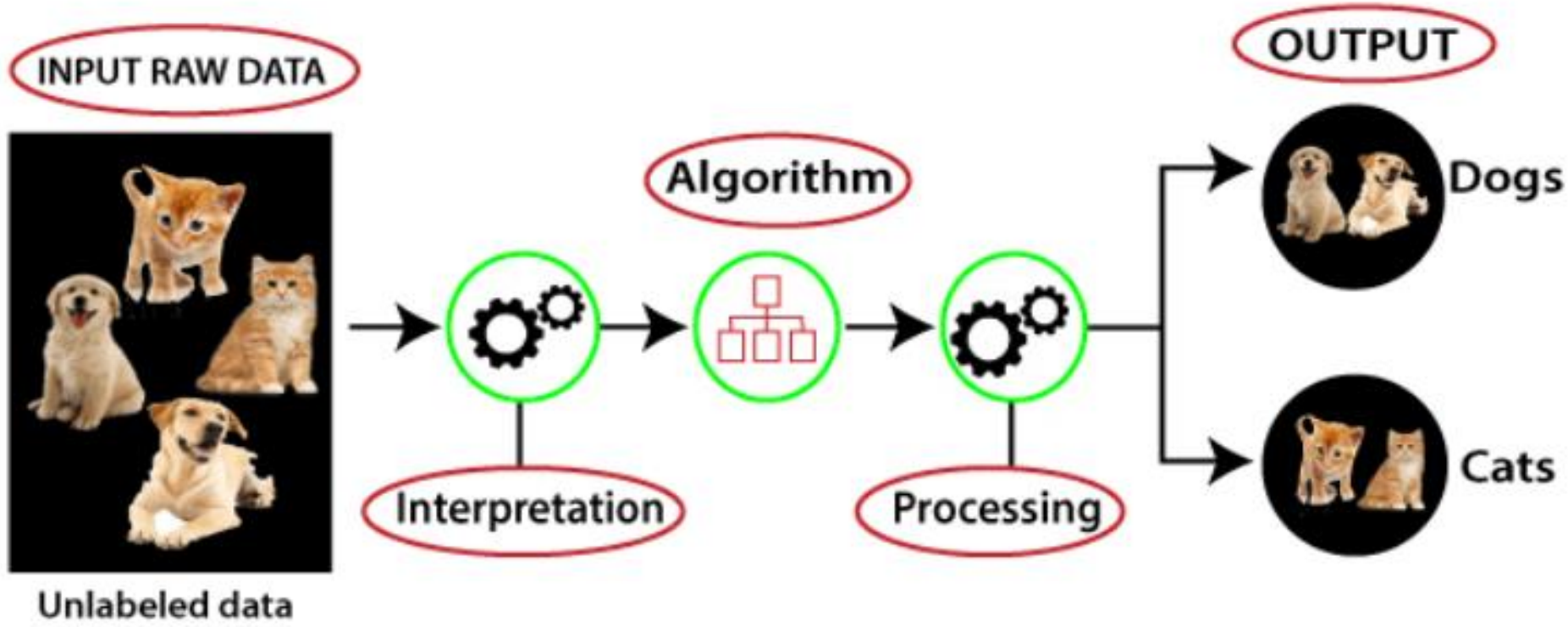
Types of Supervised Machine Algorithm

- **Classification** algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.

- Algorithms:
  - Random Forest
  - Decision Trees
  - Logistic Regression
  - Support vector Machines

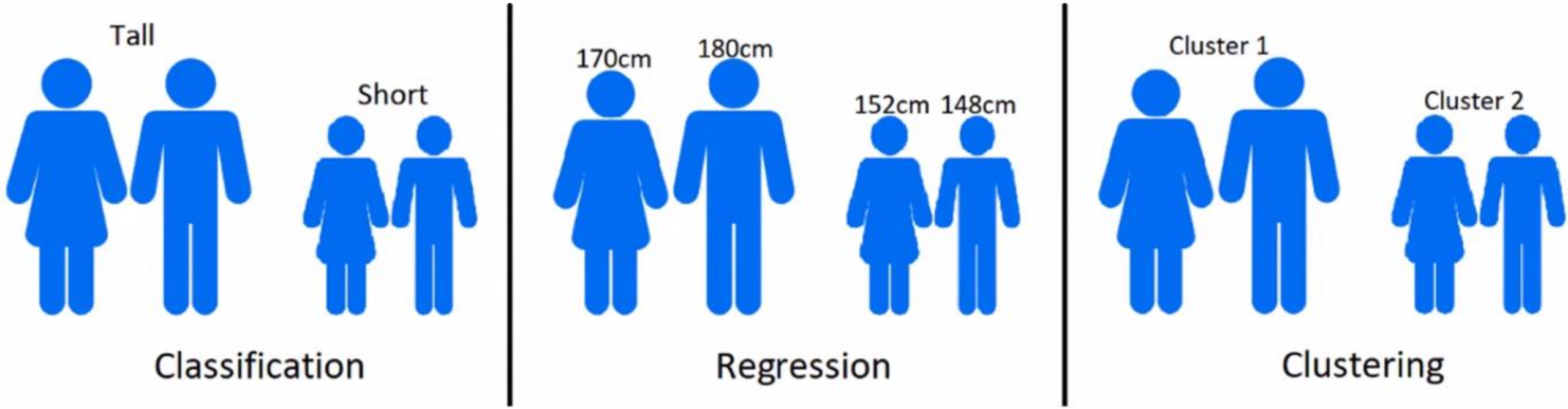
Unsupervised machine learning

- No labels are given to the learning algorithm, leaving it on its own to find structure in its input.
- Models itself find the hidden patterns and insights from the given data. It can be compared to learning which takes place in the human brain while learning new things.
- **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
- **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.


INPUT RAW DATA
Unlabeled data
Interpretation
Algorithm
Processing
OUTPUT
Dogs
Cats

Unsupervised Learning Algorithms

- **K-means clustering**
- **KNN (k-nearest neighbors)**
- **Hierarchal clustering**
- **Anomaly detection**
- **Neural Networks**
- **Principle Component Analysis**
- **Independent Component Analysis**
- **Apriori algorithm**


Tall
Short
Classification
170cm
180cm
152cm
148cm
Regression
Cluster 1
Cluster 2
Clustering

Reinforcement learning

- Reinforcement learning works by programming an algorithm with a distinct goal and a prescribed set of rules for accomplishing that goal.

- Data scientists also program the algorithm to seek positive rewards -- which it receives when it performs an action that is beneficial toward the ultimate goal -- and avoid punishments -- which it receives when it performs an action that gets it farther away from its ultimate goal.

Reinforcement learning

- **Robotics:** Robots can learn to perform tasks the physical world using this technique.
- **Video gameplay:** Reinforcement learning has been used to teach bots to play a number of video games.

Supervised Learning

Takes labeled inputs and maps it to the known outputs

Unsupervised Learning

Understands patterns and trends in the data and discovers the output

Reinforcement Learning

Follows trial and error method to arrive at the desired solution

Issues in Machine learning

- Lack Of Quality Data: noisy, incorrect, incomplete data
- Getting bad recommendations : proposal engines with complex algorithms may tend to provide wrong results
- Talent deficit: rare experts are available
- Implementation: slow deployment, data security, lack of data
- Making the wrong assumptions
- Having algorithms become obsolete when data grows
- Absence of skilled resources
- Complexity

Application of Machine Learning

Application of Machine Learning

- **Automation**
- **Banking and Finance**
- **Transportation and Traffic Prediction**
- **Image Recognition**
- **Speech Recognition**
- **Product Recommendation**
- **Virtual Personal Assistance**
- **Email Spam and Malware detection & Filtering**
- **Language Translation**
- **Self-driving cars**

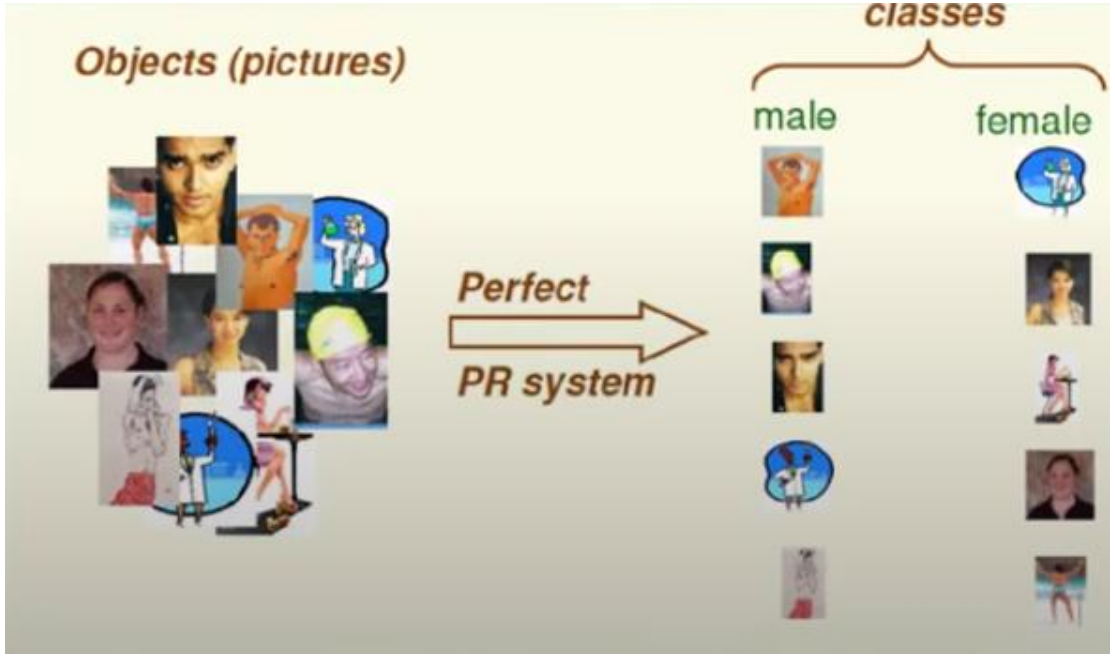
Pattern Recognition(PR)

- **Pattern** is everything around in this digital world.
- A pattern can either be seen physically or it can be observed mathematically by applying algorithms.
- **Example:** The colors on the clothes, speech pattern, etc.
- In computer science, a pattern is represented using vector feature values.

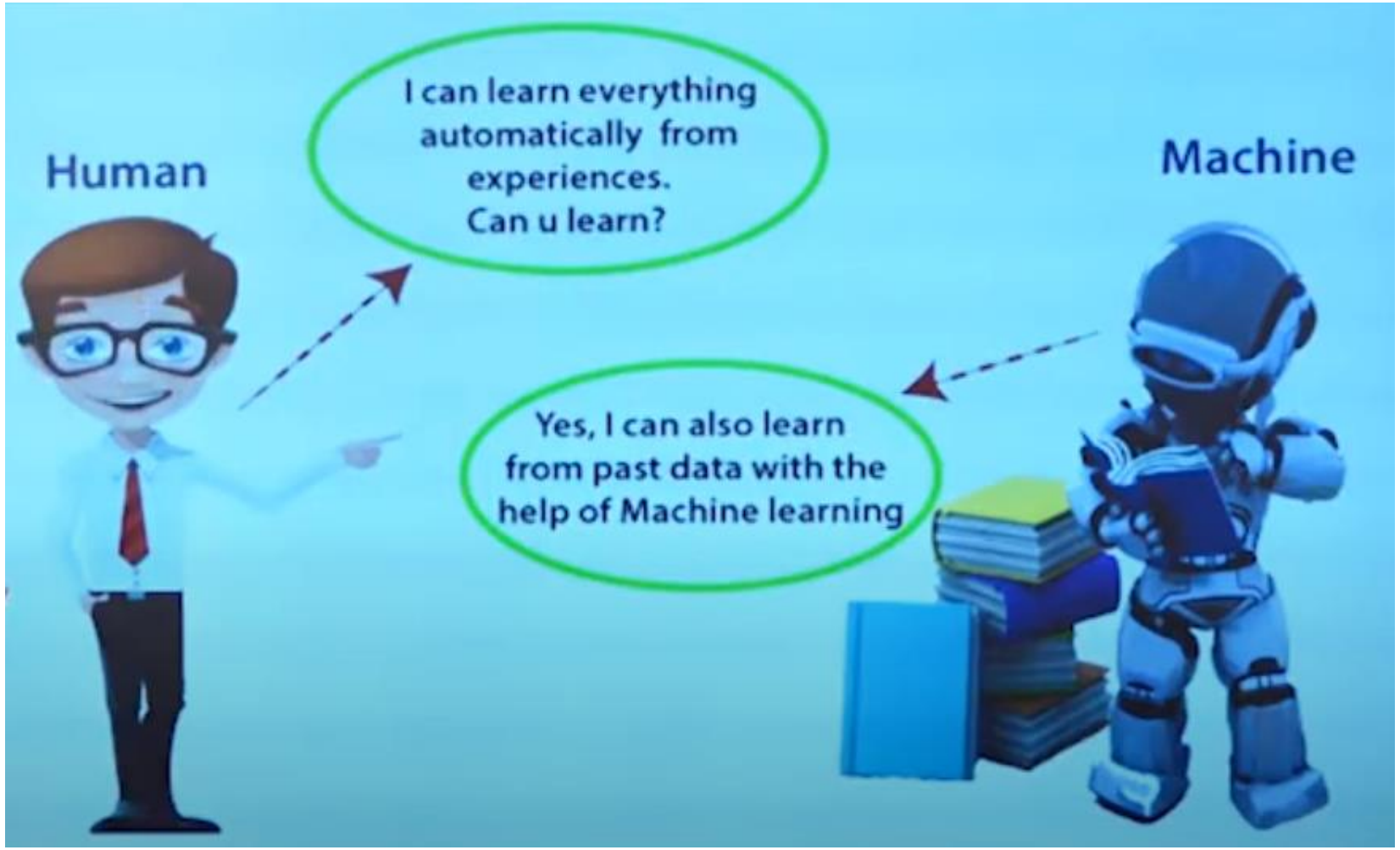
- It is generally easy for a person to differentiate
  - the sound of a human voice, from that of a violin;
  - a handwritten numeral "3," from an "8";
  - the aroma of a rose, from that of an onion.
- However, it is difficult for a programmable computer to solve these kinds of perceptual problems.
- These problems are difficult because each pattern usually contains a large amount of information, and the recognition problems typically have an inconspicuous, high-dimensional, structure.

Pattern Recognition(PR)

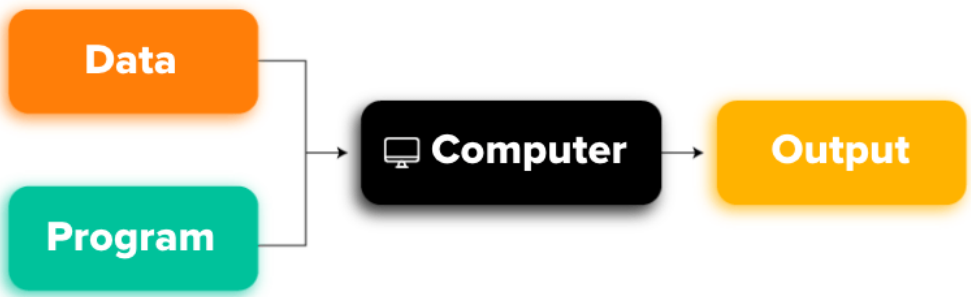
- Identifying and analyzing the patterns in the data is known as pattern recognition.
- Machine Learning is the complex set of algorithms that allows to automatically identify patterns in the data.



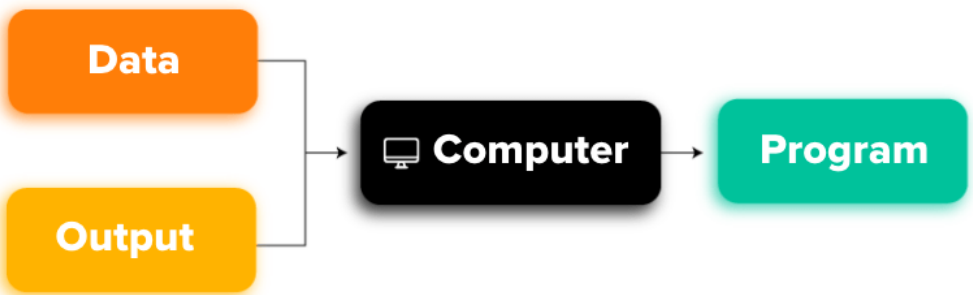
Machine Learning



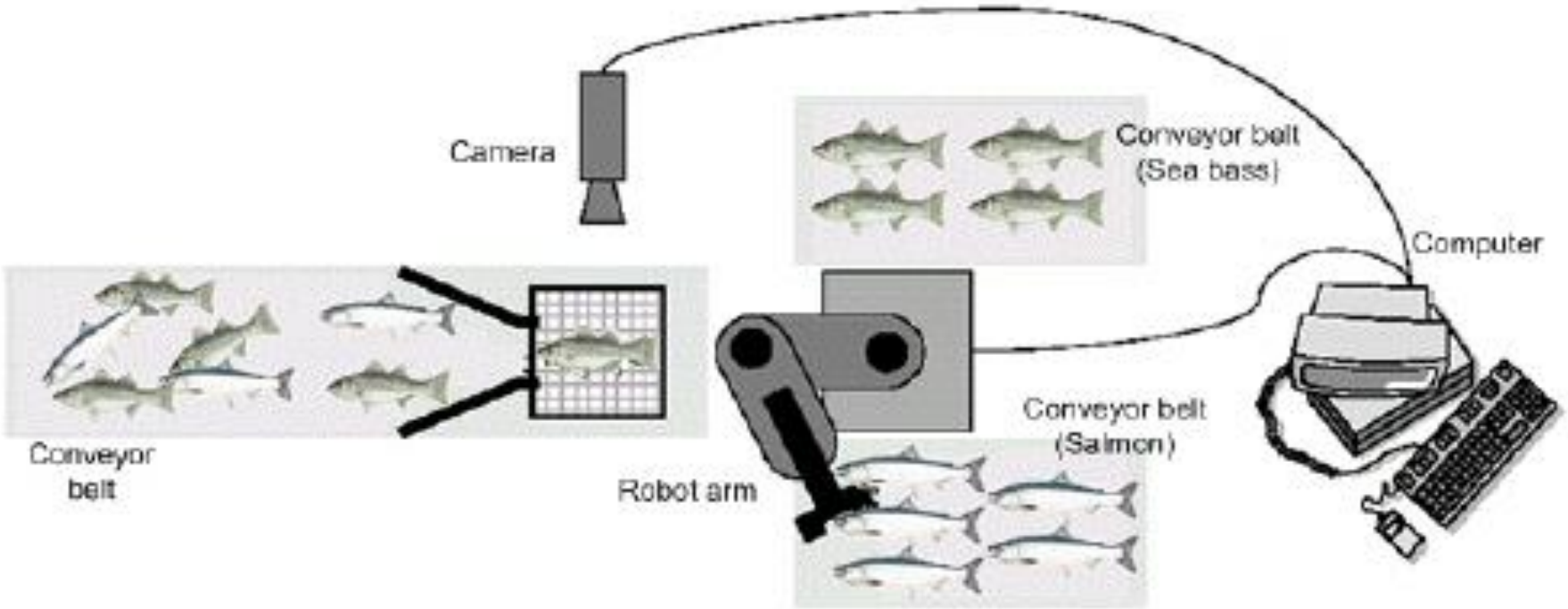
TRADITIONAL PROGRAMMING



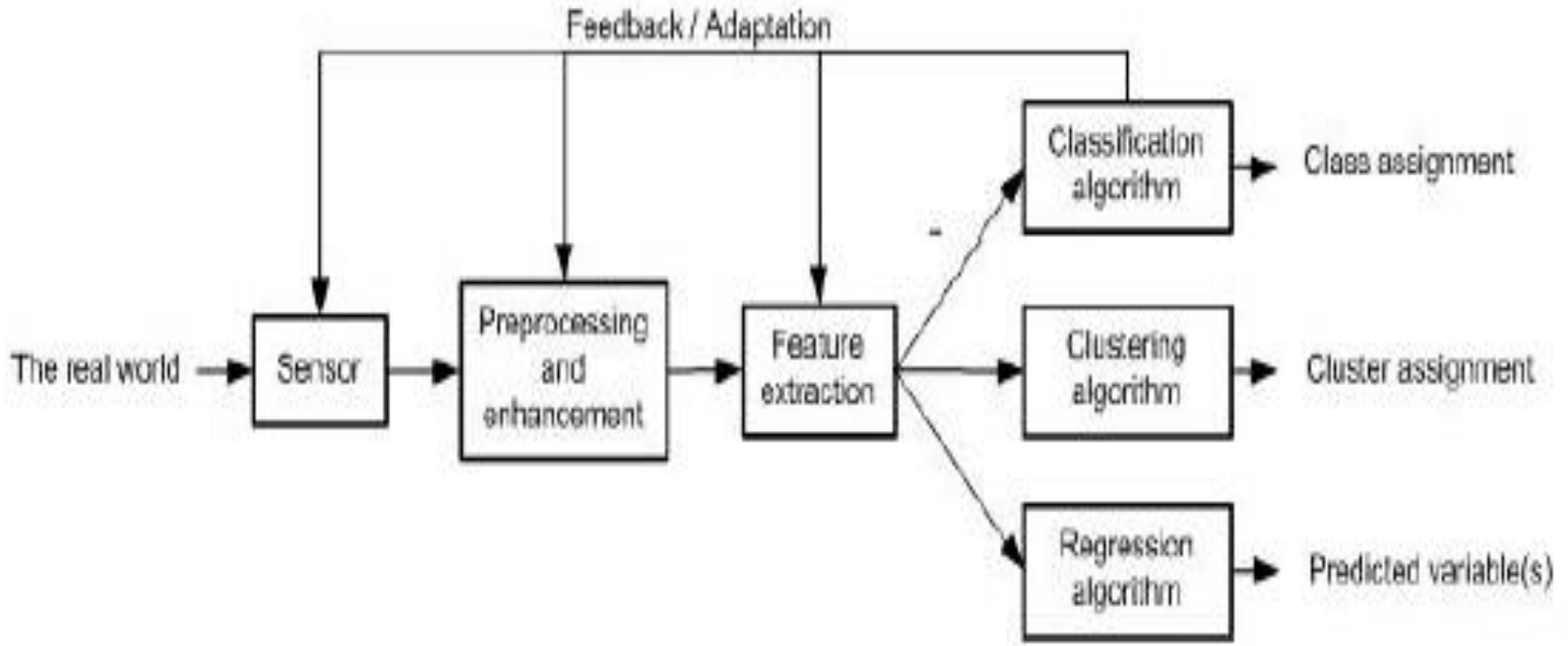
MACHINE LEARNING



PR for fish classification

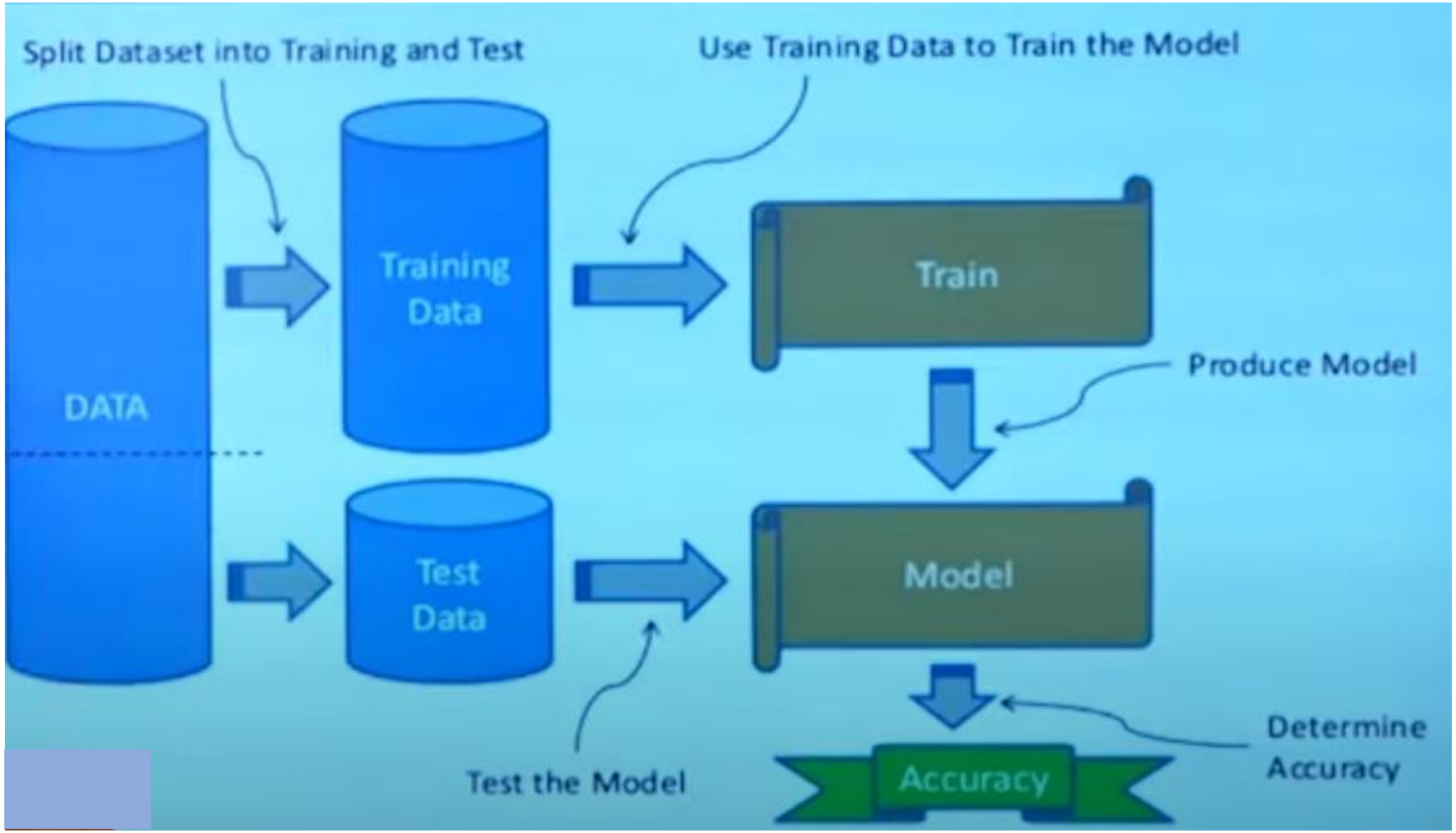


Typical PR system



- A pattern recognition system needs some input from the real world that it perceives with sensors. Such a system can work with any type of data: images, videos, numbers, or texts.
- Having received some information as the input, the algorithm performs **preprocessing**. That is segmenting something interesting from the background.
  - For example, when you are given a group photo and a familiar face attracts your attention, this is preprocessing.
- Preprocessing is tightly connected with **enhancement**.
  - By this term, researchers understand an increase in the ability of a human or a system to recognize patterns even when they are vague.
  - Imagine you are still looking at the same group photo but it is 20 years old.
  - To make sure that the familiar face in the photo is really the person you know, you start comparing their hair, eyes, and mouth. This is when enhancement steps into the game.
- The next component is **feature extraction**. The algorithm uncovers some characteristic traits that are similar to more than one data sample.
- The result of a pattern recognition system will be either a class assignment (if we used classification), or cluster assignment (in case of clustering), or predicted values (if you apply regression).

Machine learning Pipeline



Training and Learning in Pattern Recognition

- Learning is a phenomenon through which a system gets trained and becomes adaptable to give results in an accurate manner.
- Learning is the most important phase as to how well the system performs on the data provided to the system depends on which algorithms are used on the data.
- The entire dataset is divided into two categories, one which is used in training the model i.e. Training set, and the other that is used in testing the model after training, i.e. Testing set.

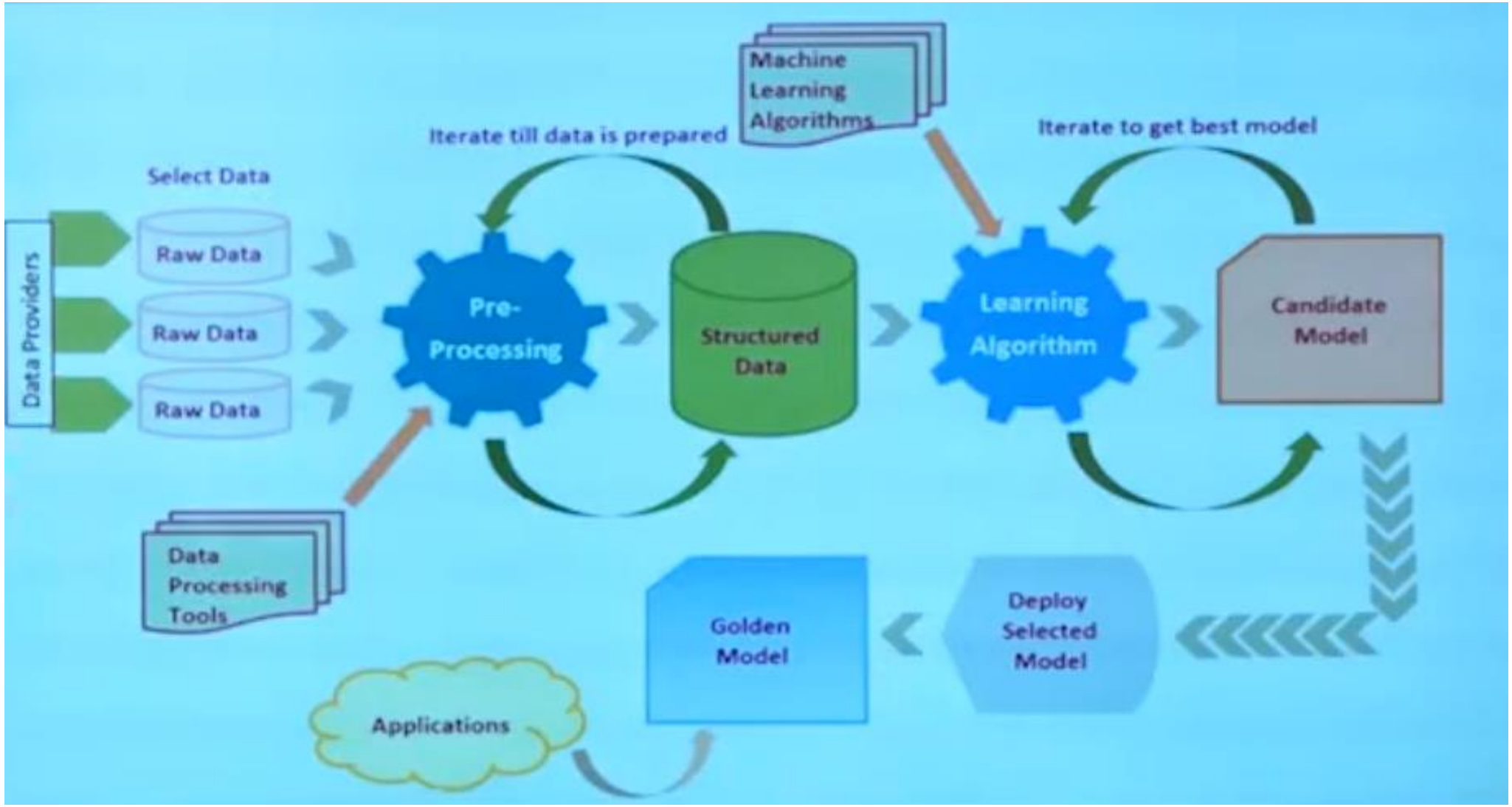
Training set:

- The training set is used to build a model.
- It consists of the set of data that are used to train the system.
- Training rules and algorithms are used to give relevant information on how to associate input data with output decisions.
- The system is trained by applying these algorithms to the dataset, all the relevant information is extracted from the data, and results are obtained.
- Generally, 80% of the data of the dataset is taken for training data.

Testing set:

- Testing data is used to test the system.
- It is the set of data that is used to verify whether the system is producing the correct output after being trained or not.
- Generally, 20% of the data of the dataset is used for testing.
- Testing data is used to measure the accuracy of the system.
- For example, a system that identifies which category a particular flower belongs to is able to identify seven categories of flowers correctly out of ten and the rest of others wrong, then the accuracy is 70 %

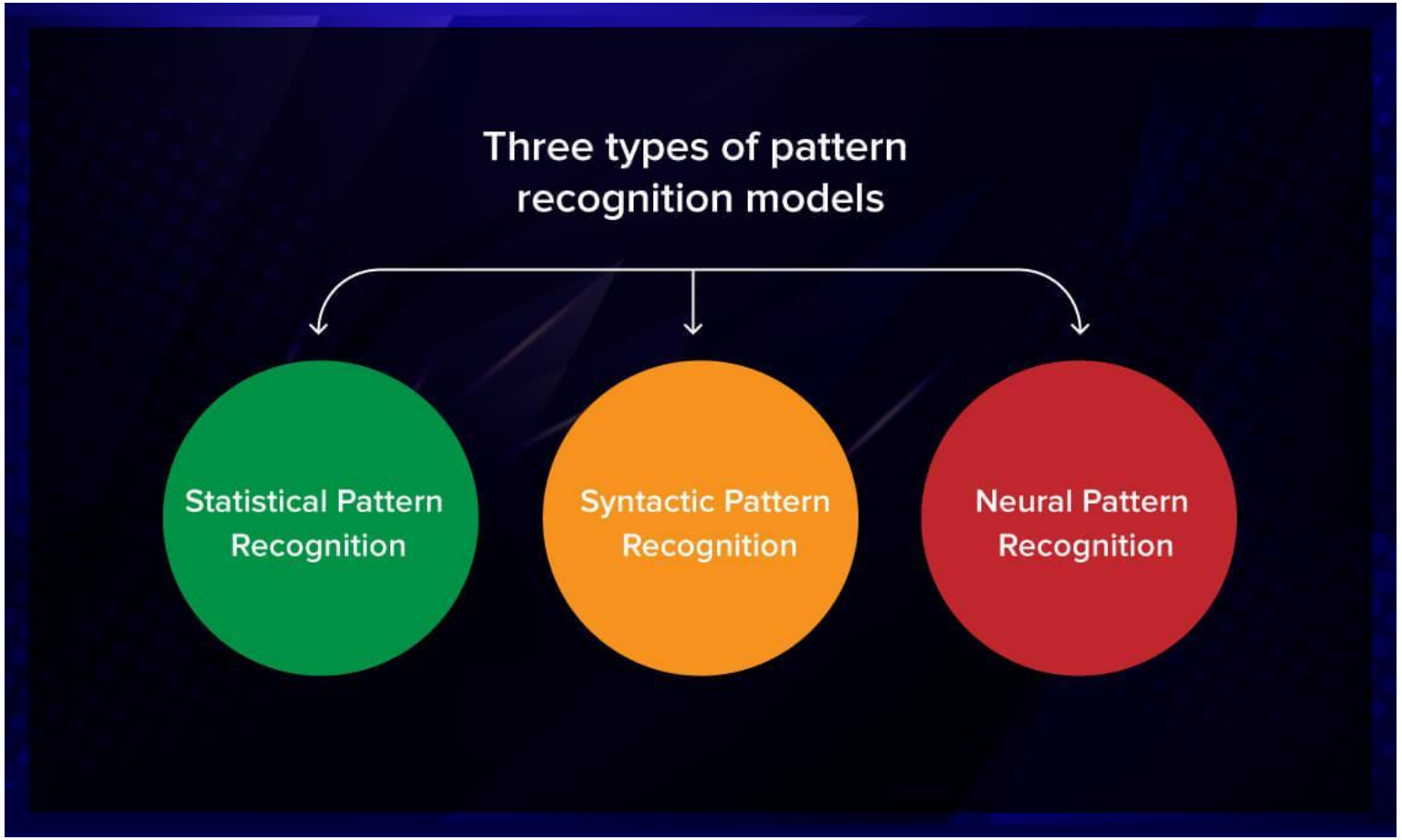
Typical PR Workflow



Real-time Examples

- A pattern is a physical object or an abstract notion.
- While talking about the classes of animals, a description of an animal would be a pattern.
- While talking about various types of balls, then a description of a ball is a pattern.
  - In the case balls considered as pattern, the classes could be football, cricket ball, table tennis ball, etc. Given a new pattern, the class of the pattern is to be determined.
  - The choice of attributes and representation of patterns is a very important step in pattern classification.
  - A good representation is one that makes use of discriminating attributes and also reduces the computational burden in pattern classification.
- An obvious representation of a pattern will be a **vector**.
  - Each element of the vector can represent one attribute of the pattern.
  - The first element of the vector will contain the value of the first attribute for the pattern being considered.
- **Example:** While representing spherical objects, (25, 1) may be represented as a spherical object with 25 units of weight and 1 unit diameter.
- The class label can form a part of the vector.
- If spherical objects belong to class 1, the vector would be (25, 1, 1), where
  - the first element represents the weight of the object,
  - the second element, the diameter of the object and
  - the third element represents the class of the object.

How does pattern recognition work?



Types of Pattern Recognition model

- **Statistical Pattern Recognition**

  This type of pattern recognition refers to statistical historical data when it learns from examples: it collects observations, processes them, and learns to generalize and apply these rules to new observations.

- **Syntactic Pattern Recognition**

  It is also called structural pattern recognition because it relies on simpler subpatterns called primitives (for example, words). The pattern is described in terms of connections between the primitives, for example, words form sentences and texts.

- **Neural Pattern Recognition**

  In neural pattern recognition, artificial neural networks are used. They can learn complex nonlinear input-output relations and adapt themselves to the data.

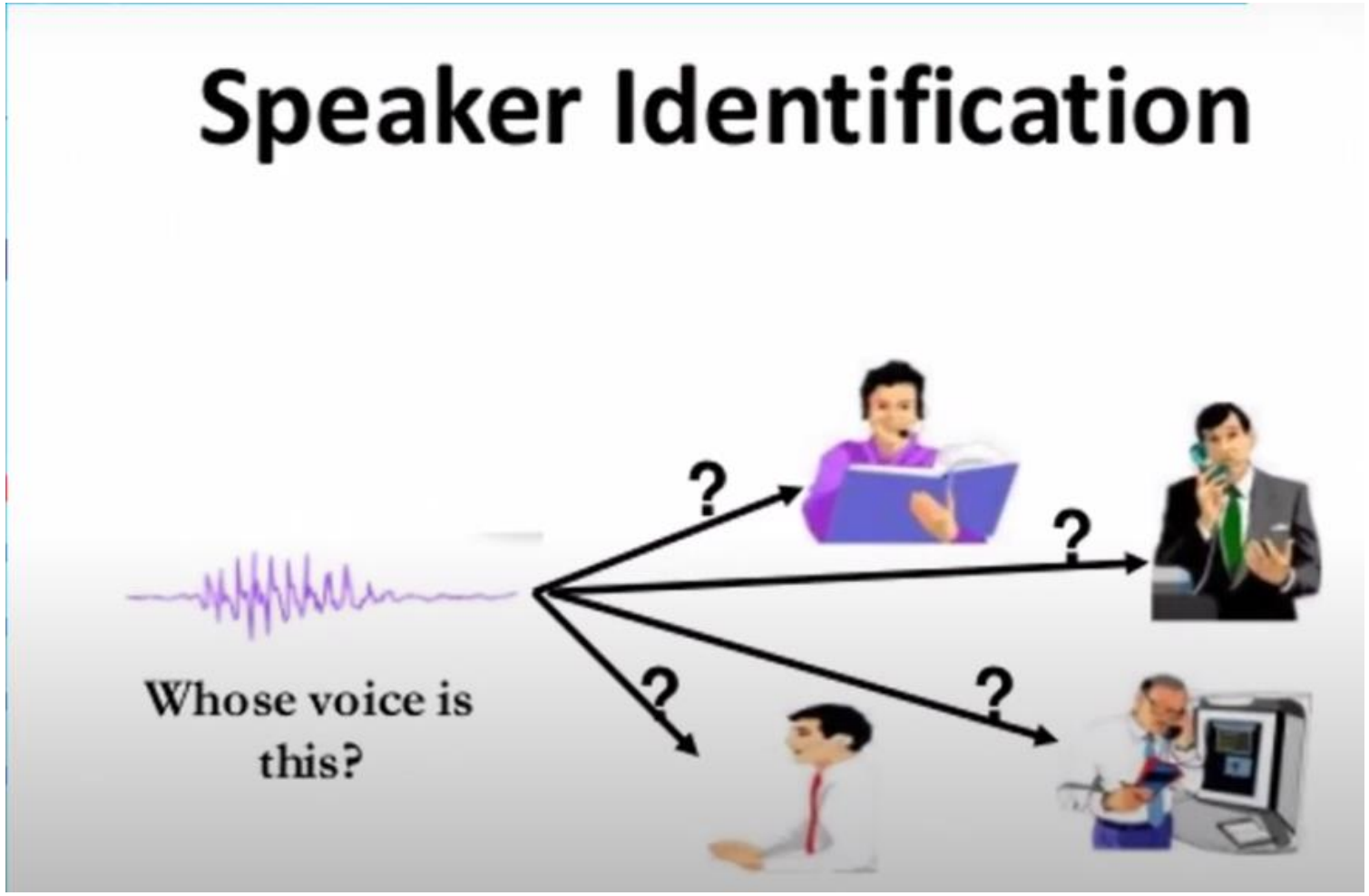
Classification

- **Given a sample the objective of the model is to identify suitable class label**
- **Comes under Supervised learning**
- **Model predicts an integer value**

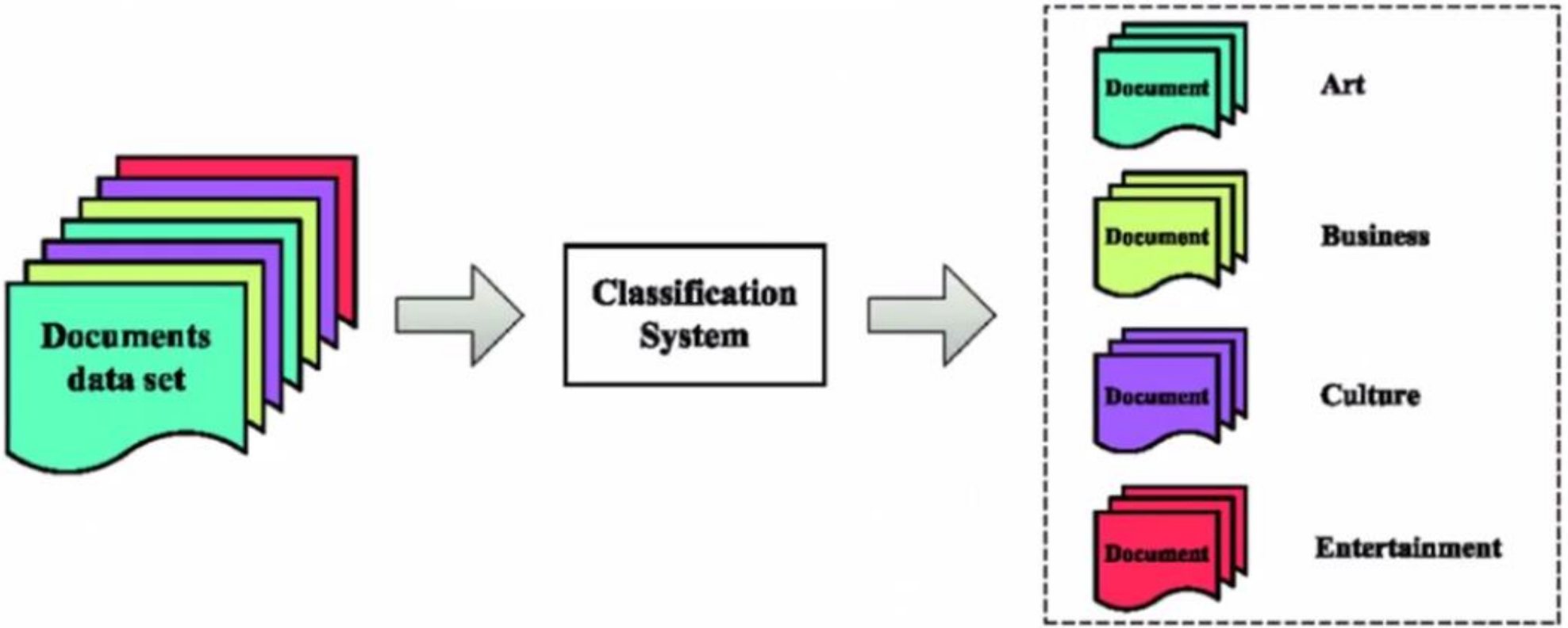
Examples:

- Speaker recognition
- Tumor classification
- Face Recognition
- Document classification

Speaker Recognition

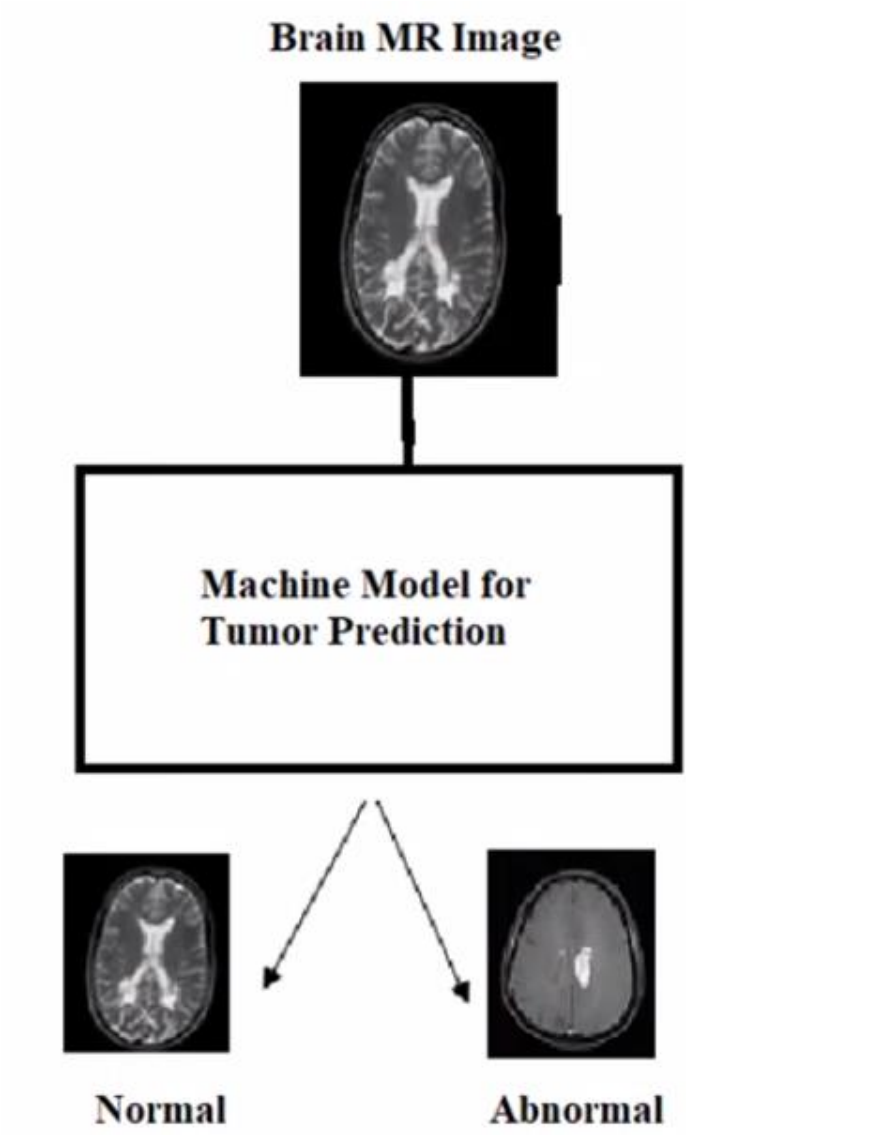


Document classification

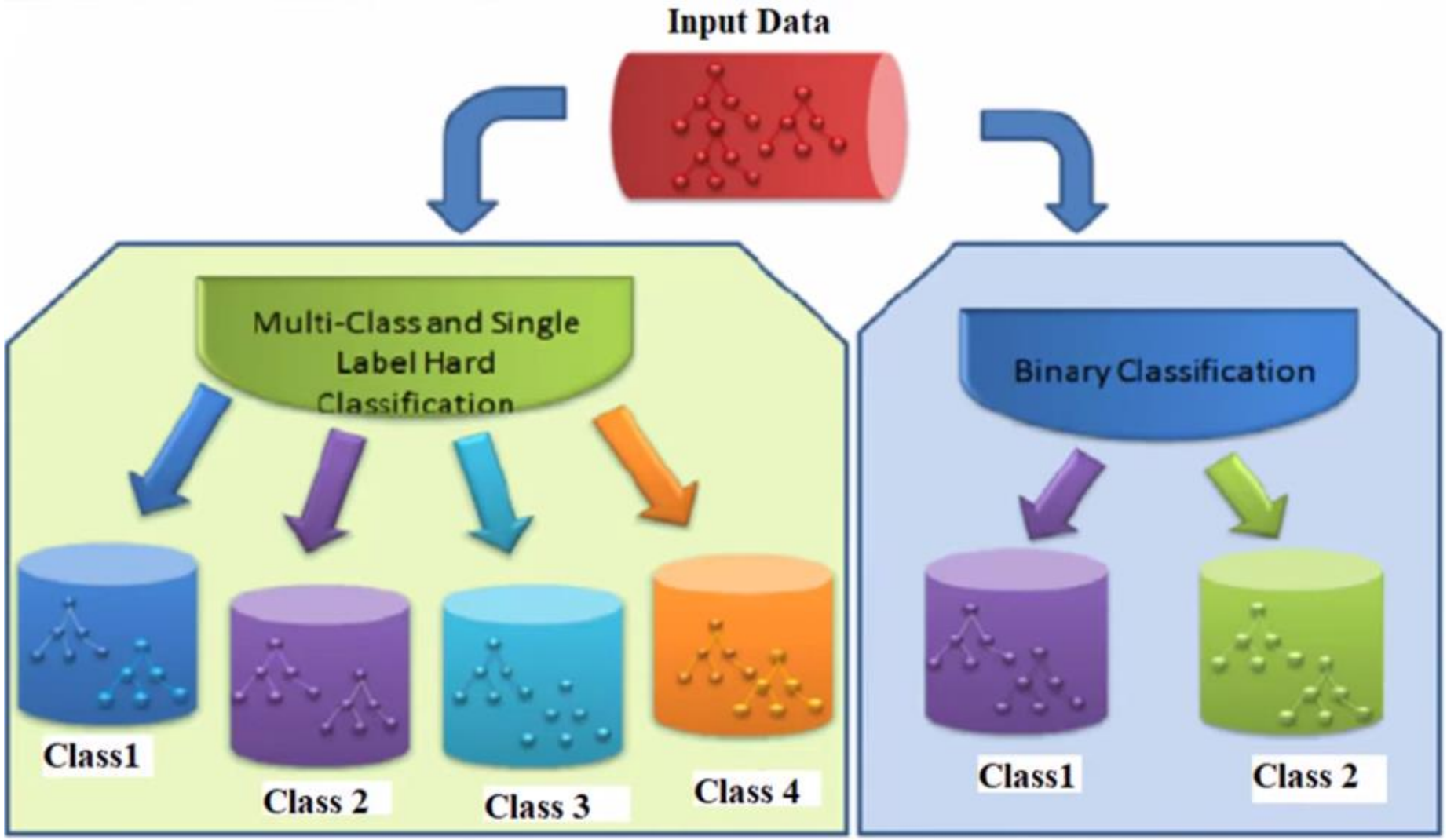


A simple example algorithm framework for text categorization.

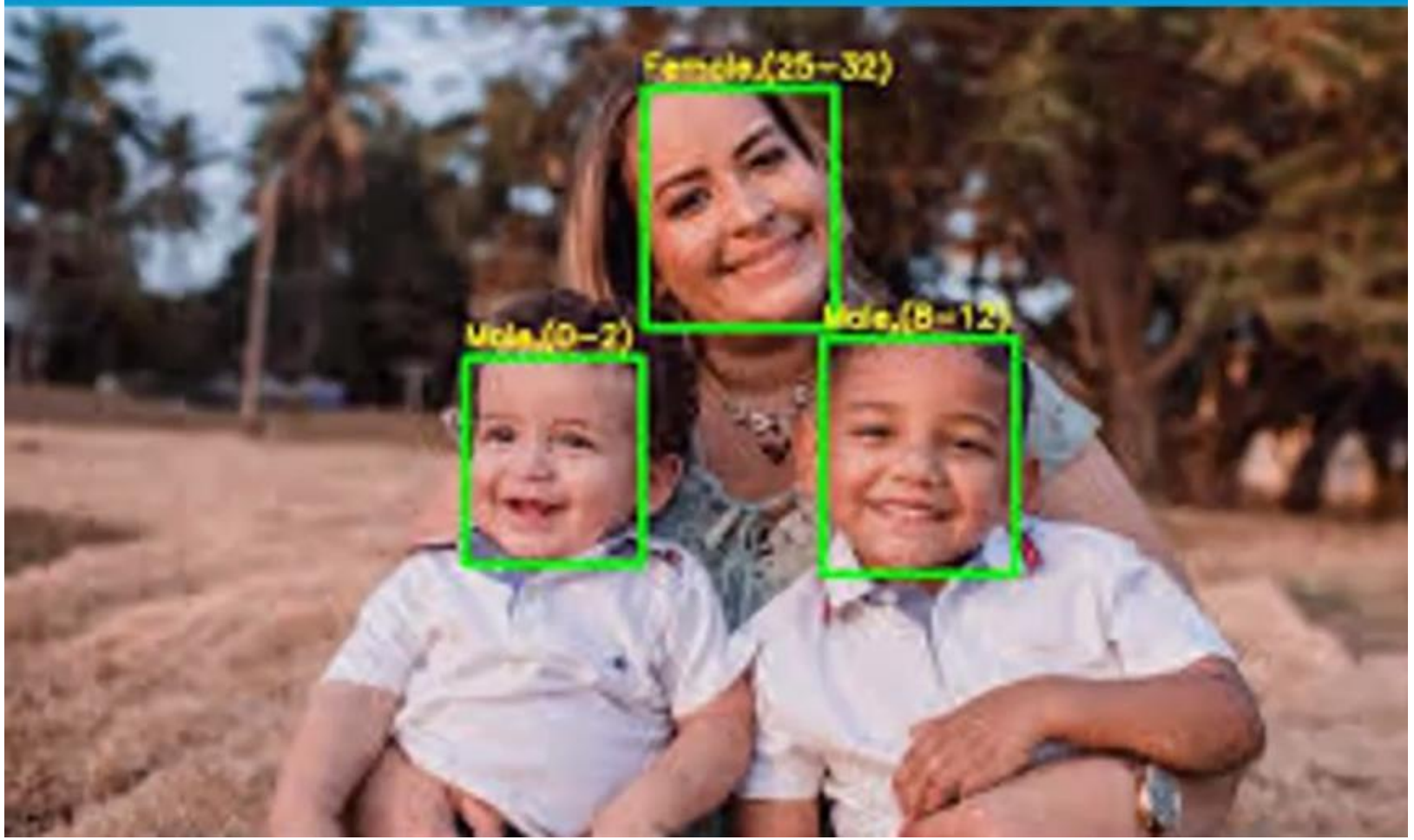
Tumor Classification



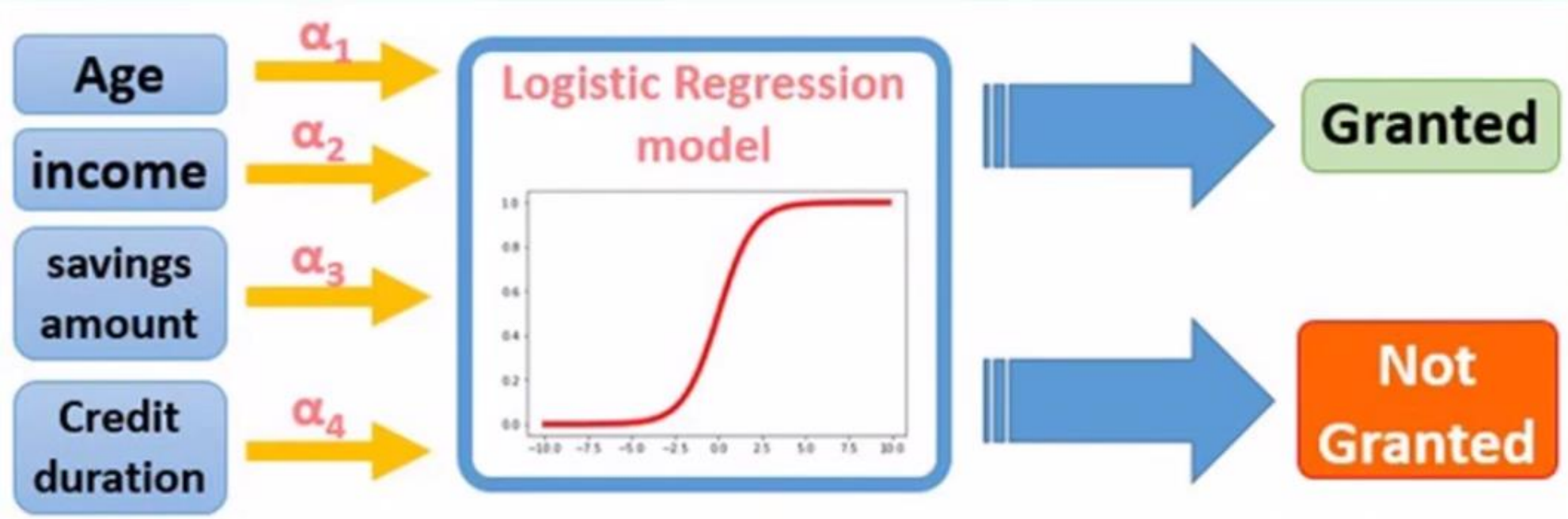
Types of Classification



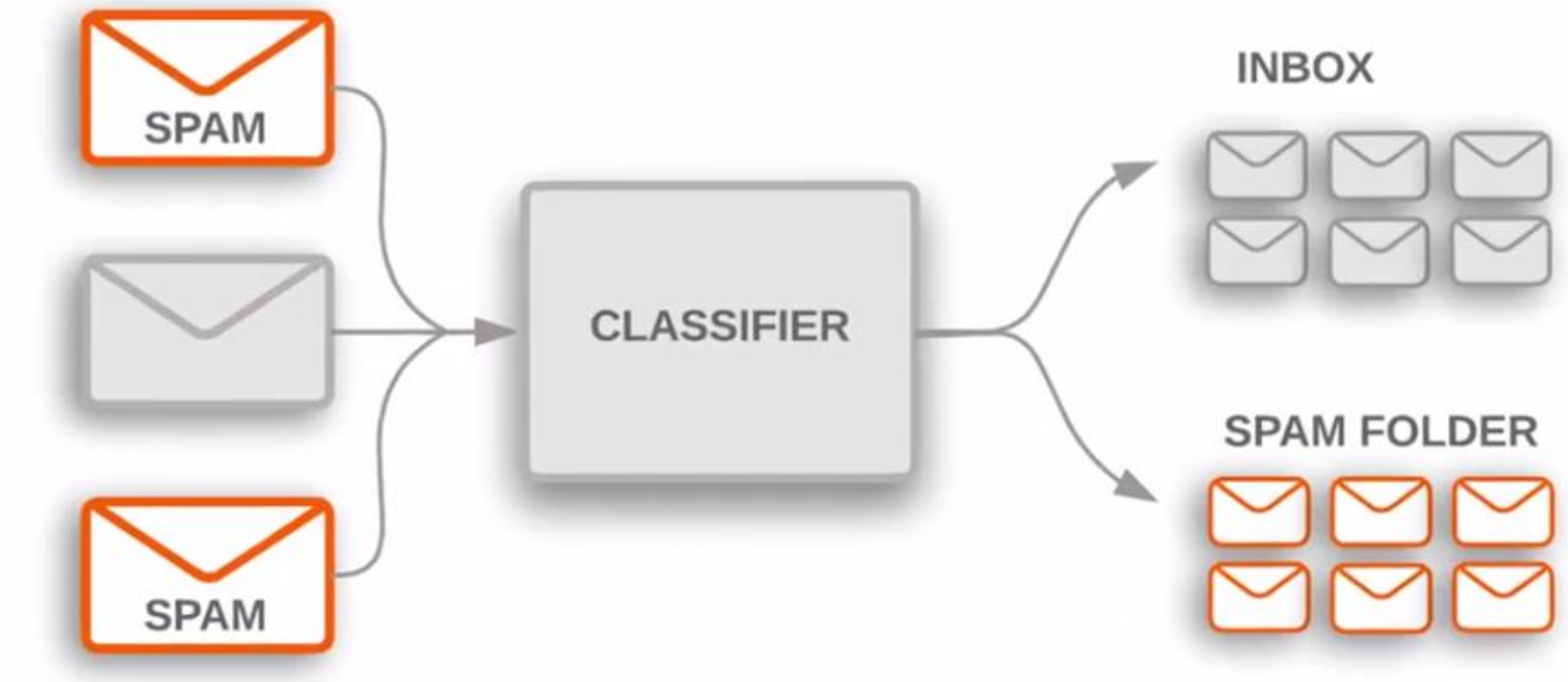
Gender Classification



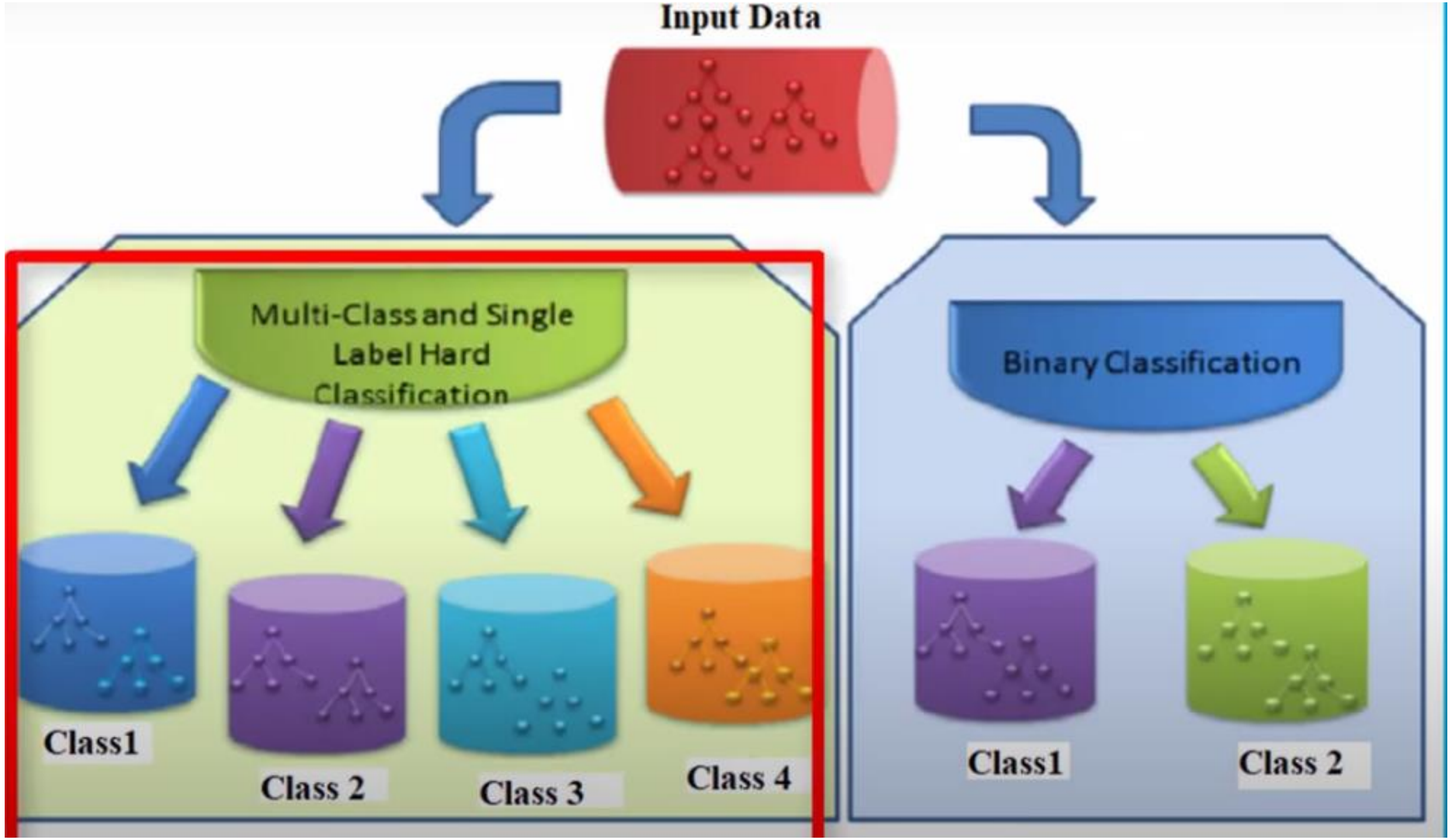
Loan approval System



SPAM/Normal Mail Prediction



Multiclass Classification

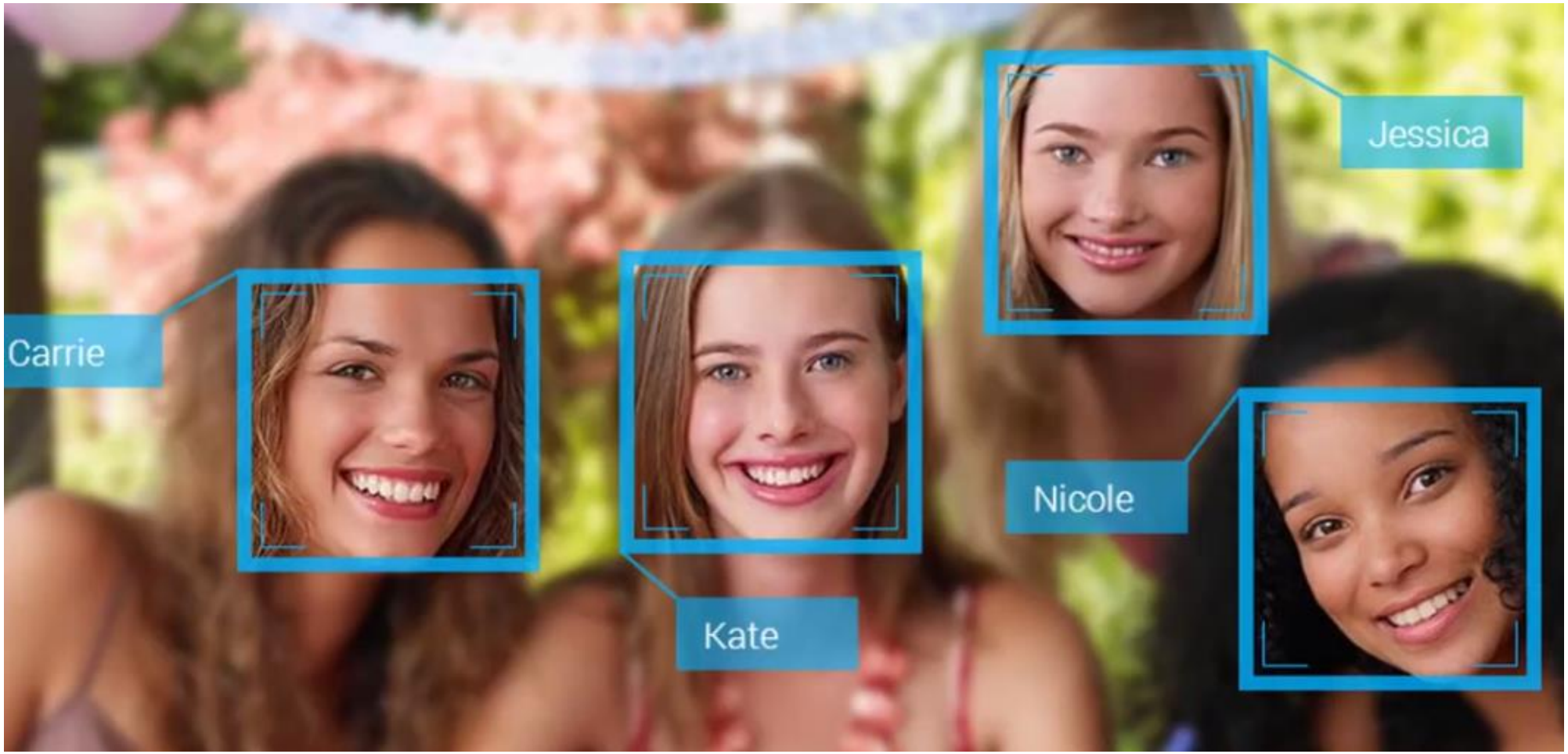


Multi-Class Classification

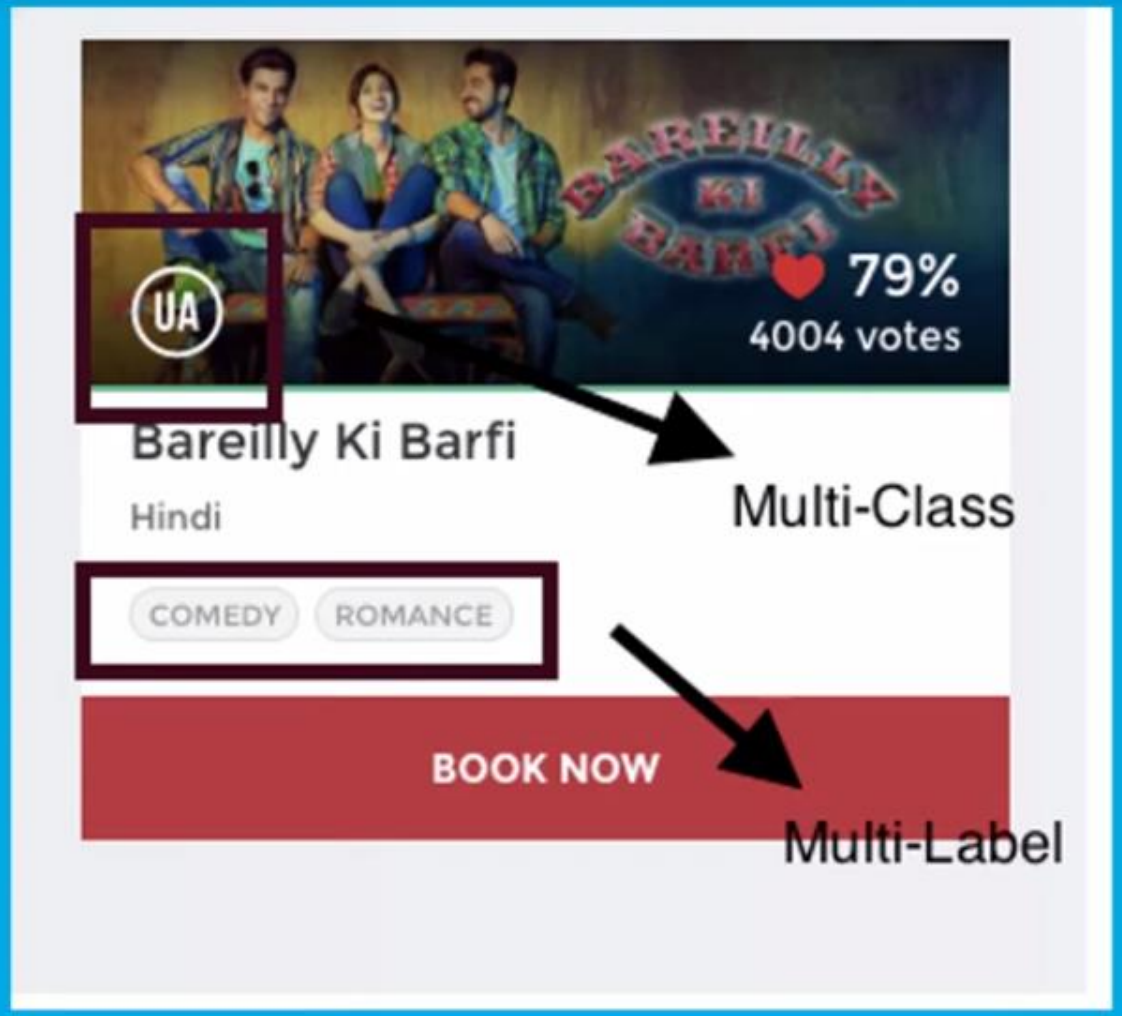
Data & Labels

0
1
2
3
4
5
6
7
8
9

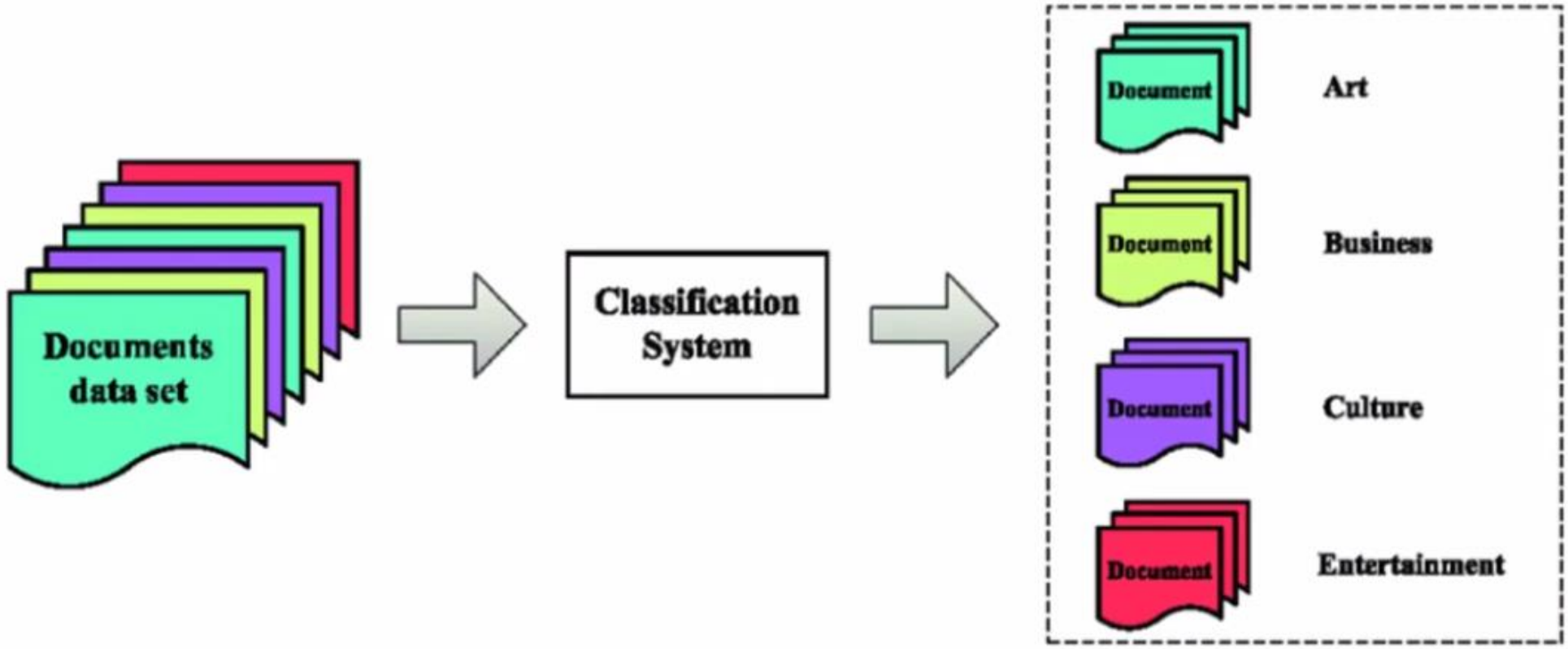
Multiclass classification



Multi-Label Classification



Multi Label Classification



A simple example algorithm framework for text categorization.

Binary Classification

- Spam
- Not spam

Multiclass Classification

- Dog
- Cat
- Horse
- Fish
- Bird
- ...

Multi-label Classification

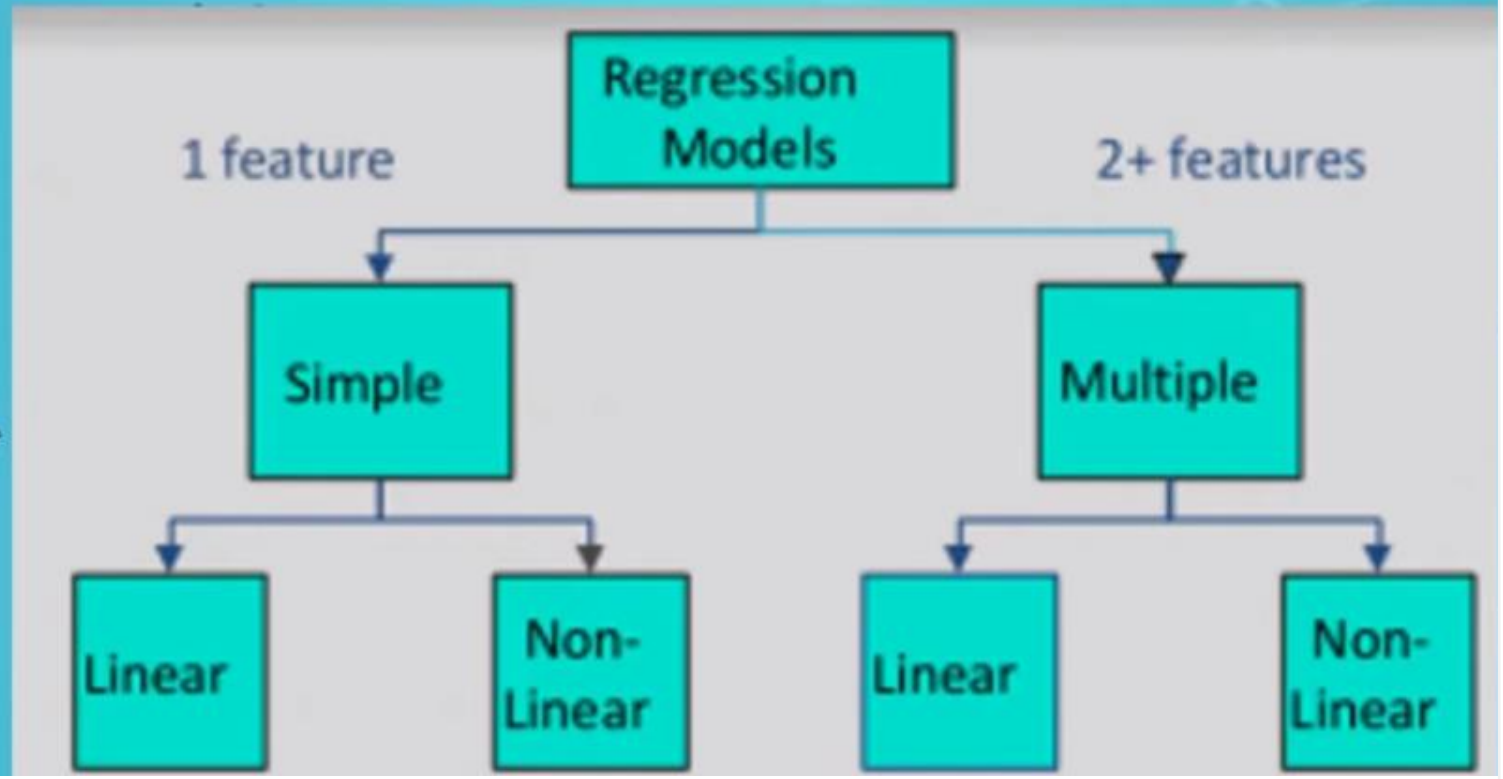
- Dog
- Cat
- Horse
- Fish
- Bird
-

Regression

- Similar to Classification, in Regression also the objective to predict a value
- The value to be predicted is a real or continuous value, such as “salary” or “weight” or “temp”

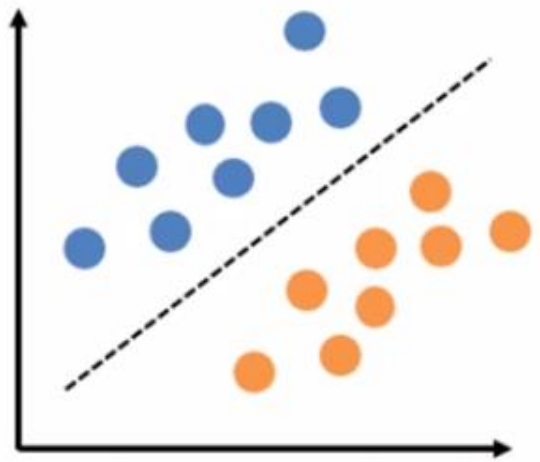
Regression Analysis

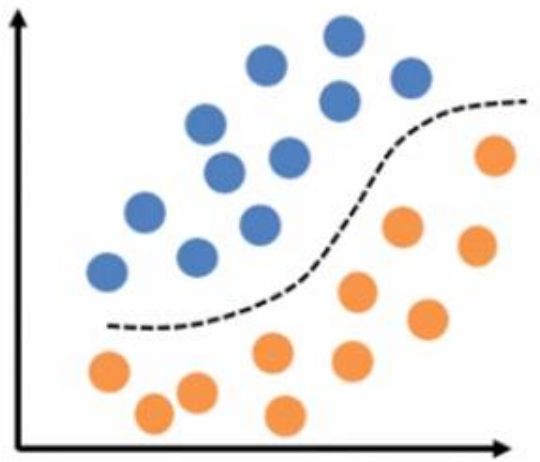
- Similar to Classification, in Regression also the objective to predict a value
- The value to be predicted is a real or continuous value, such as “salary” or



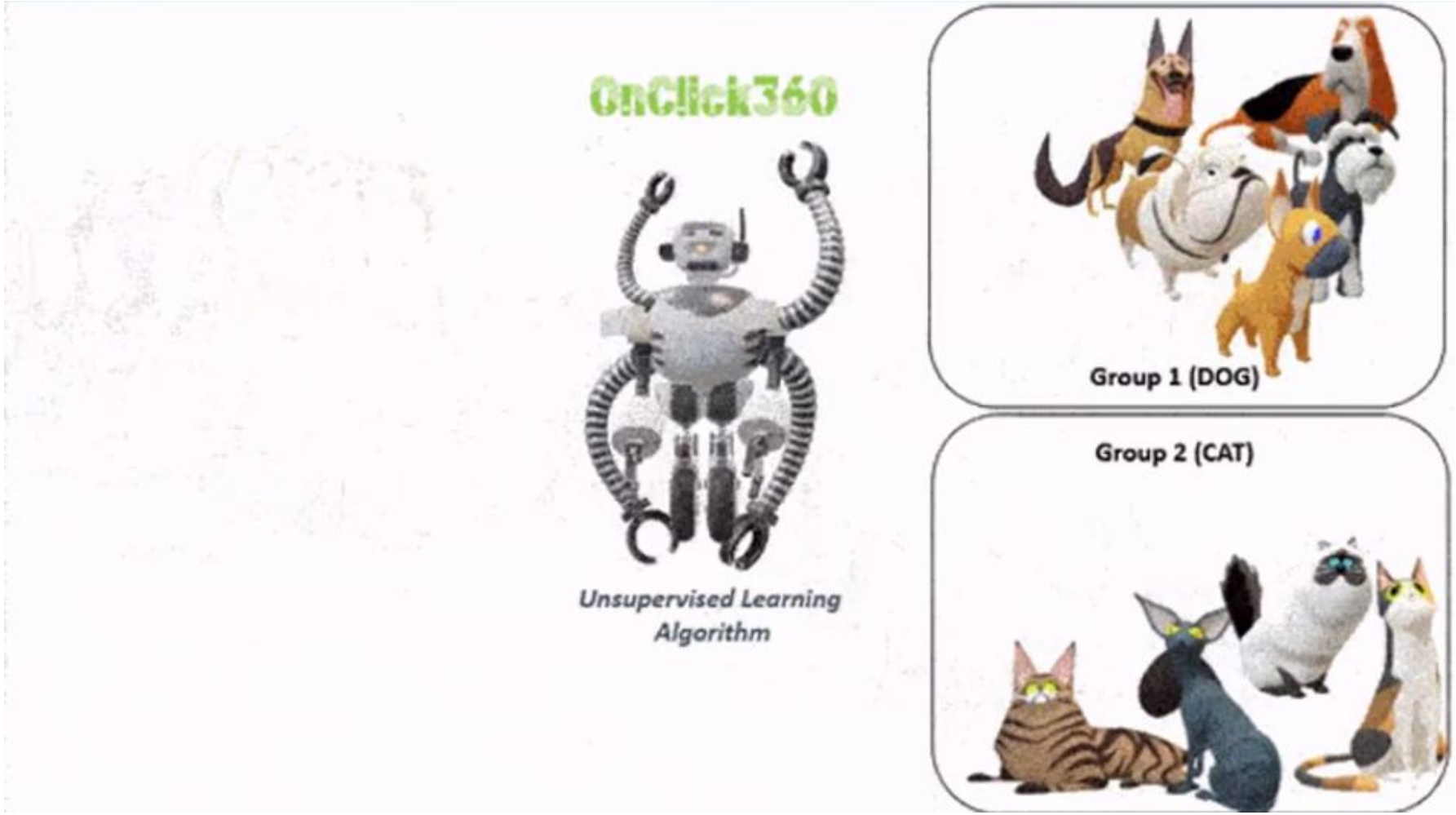
Regression Analysis

- Similar to Classification, in Regression also the objective to predict a value
- The value to be predicted is a real or continuous value, such as “salary” or





Clustering


OnClick360
Unsupervised Learning Algorithm
Group 1 (DOG)
Group 2 (CAT)

Clustering

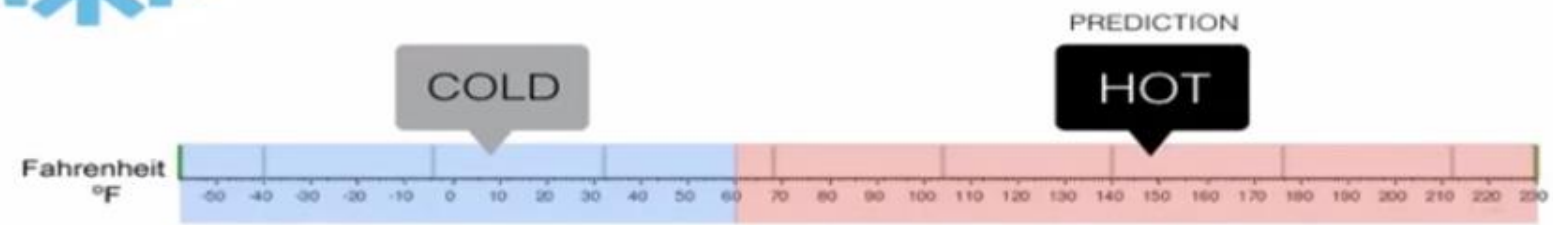
Classification vs Regression

Regression

What is the temperature going to be tomorrow?

Classification

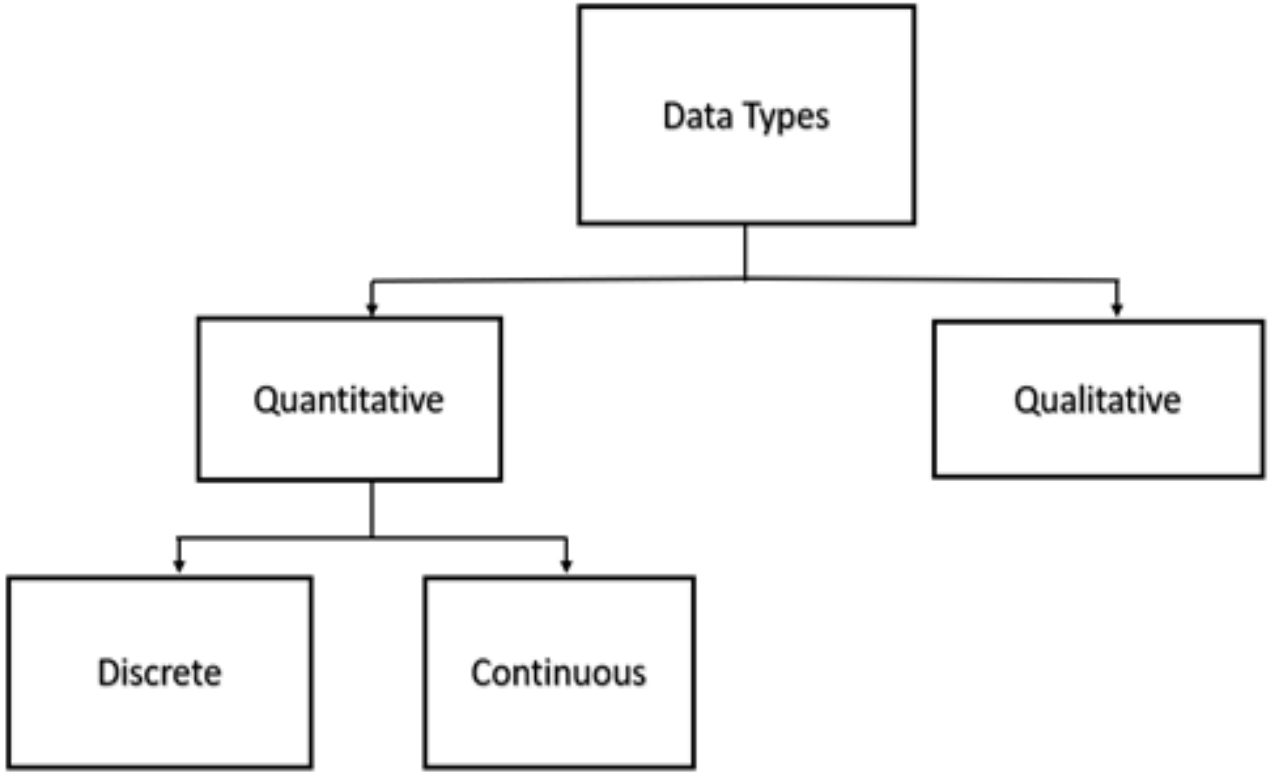
Will it be Cold or Hot tomorrow?



Case Studies

- what is tomorrow's temperature?
- what is the value of the stock?
- how many runs can be taken by Virat in tomorrow's test match?
- will it rain today or not?
- Whether the Covid test result comes Positive or negative?
- Is this picture is of a male or female?
- This mail is spam or important or promotion?
- Is this picture a cat or a dog or a elephant?
- given the balls, group them?
- Given the set of transactions identify the abnormal ones?
- Given the satellite image mark the forest area
- Given a set of faces, group them into males and females?
- Given an MR image of brain identify the tumor effected regions

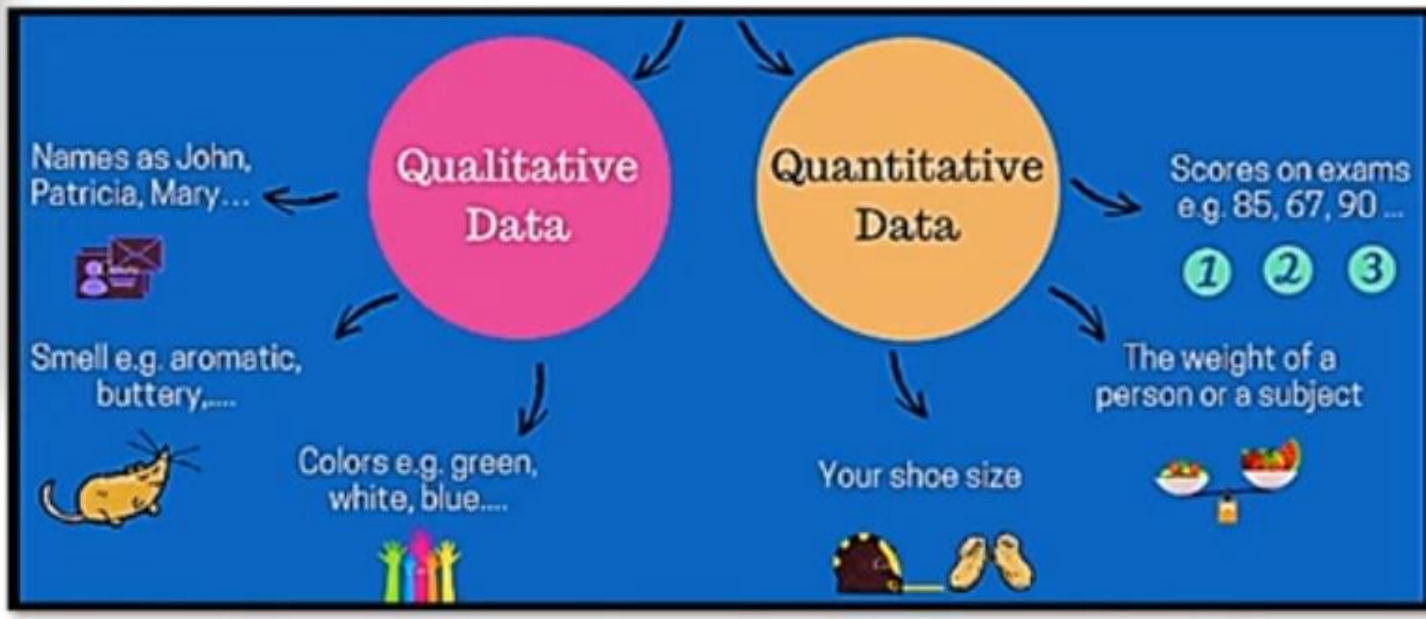
Data Types In Machine Learning



Types of Data (Qualitative and Quantitative)

QUALITATIVE DATA VERSUS QUANTITATIVE DATA

QUALITATIVE DATA	QUANTITATIVE DATA
Data type that consists of descriptive statements	Data type that can be measured and expressed numerically
Text-based	Number-based
Statistical analysis is harder	Statistical analysis is easier
Collected using interviews, written documents, observations	Collected using surveys, observations, experiments, and interviews



Quantitative Data Type

- This Type Of Data Type Consists Of Numerical Values. Anything Which Is Measured By Numbers.
- E.G., Profit, Quantity Sold, Height, Weight, Temperature, Etc.

This Is Again Of Two Types

A.) Discrete Data Type: –

The Numeric Data Which Have Discrete Values Or Whole Numbers. This Type Of Variable Value If Expressed In Decimal Format Will Have No Proper Meaning. Their Values Can Be Counted.

E.G.: – No. Of Cars You Have, No. Of Marbles In Containers, Students In A Class, Etc.

No. of Laptops

No. of Cars

B.) Continuous Data Type: –

The Numerical Measures Which Can Take The Value Within A Certain Range. This Type Of Variable Value If Expressed In Decimal Format Has True Meaning. Their Values Can Not Be Counted But Measured. The Value Can Be Infinite

E.G.: – Height, Weight, Time, Area, Distance, Measurement Of Rainfall, Etc.

Height

Time

Qualitative Data Type:

- These Are The Data Types That Cannot Be Expressed In Numbers. This Describes Categories Or Groups And Is Hence Known As The Categorical Data Type.

- This Can Be Divided Into:-

A. Structured Data:

This Type Of Data Is Either Number Or Words. This Can Take Numerical Values But Mathematical Operations Cannot Be Performed On It. This Type Of Data Is Expressed In Tabular Format.

E.G.) Sunny=1, Cloudy=2, Windy=3 Or Binary Form Data Like 0 Or1, Good Or Bad, Etc.

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

B. Unstructured Data:

This Type Of Data Does Not Have The Proper Format And Therefore Known As Unstructured Data.This Comprises Textual Data, Sounds, Images, Videos, Etc.

Scales of Measurement
(Nominal, Ordinal, Interval, Ratio)

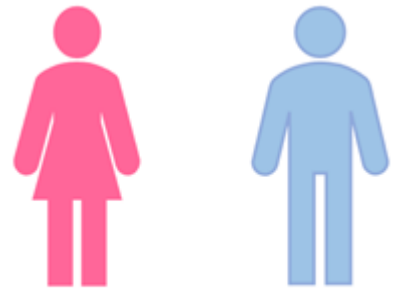


Fig: Gender (Female, Male),

1. Nominal:

- "Nominal" scales could simply be called "labels."
- In this scale, categories are nominated names (hence "nominal").
- There is no inherent order between categories.
- Put simply, one cannot say that a particular category is superior/ better than another.

Example:

- **Gender (Male/ Female):-** One cannot say that Males are better than Females, or vice-versa.
- **Blood Groups (A/B/O/AB):-** One cannot say that group A is superior to group O, for instance.
- **Religion (Hindu/ Muslim/ Christian/ Buddhist, etc.):-** Here, too, the categories cannot be arranged in a logical order. Each category can only be considered as equal to the other.

Scales of Measurement
(Nominal, Ordinal, Interval, Ratio)

2. Ordinal:

- The various categories can be logically arranged in a meaningful order.

Example:

1. **Ranks (1st/ 2nd/ 3rd, etc.)**: The ranks can be arranged in either ascending or descending order without difficulty.
2. **Liker scale (Strongly Disagree/ Disagree/ Neutral/ Agree/ Strongly Agree)**: The ordering is flexible- the order can easily be reversed without affecting the interpretation- (Strongly Agree/ Agree/ Neutral/ Disagree/ Strongly Disagree). Again, the difference between categories is not uniform.

Scales of Measurement
(Nominal, Ordinal, Interval, Ratio)

3. Interval:

- The values (not categories) can be ordered and have a meaningful difference.

Example:

The Celsius scale: 50 C is hotter than 40 C (order). However, 20 C is not half as hot as 40 C and vice versa.

Scales of Measurement
(Nominal, Ordinal, Interval, Ratio)

4. Ratio:

- Data measurement scales because they tell us about the order, they tell us the exact value between units.
- Ratio scales provide a wealth of possibilities when it comes to statistical analysis.

This Device Provides Two Examples of Ratio Scales (height and weight)

Example:

1. **Weight**: 100 kg is twice as heavy as 50 kg; the difference between 45 kg and 55 kg is the same as that between 105 kg and 100 kg; values can be arranged in an order (ascending/ descending).
2. **Height**: 100 cm is taller than 50 cm; this difference is the same as that between 150 cm and 100 cm, or 200 cm and 150 cm; 100 cm is twice as tall as 50 cm; the values can be arranged in a particular manner (ascending/ descending).

Feature Selection

Need for Feature Selection

To train a model, we collect huge quantities of data to help the machine learn better. Consider a table which contains information on old cars. The model decides which cars must be crushed for spare parts

Model	Year	Miles	Owner

Need for Feature Selection

But not all this data will be useful to us. Some classes or a part of the data may not contribute much to our model and can be dropped.

Model	Year	Miles	Owner

Need for Feature Selection

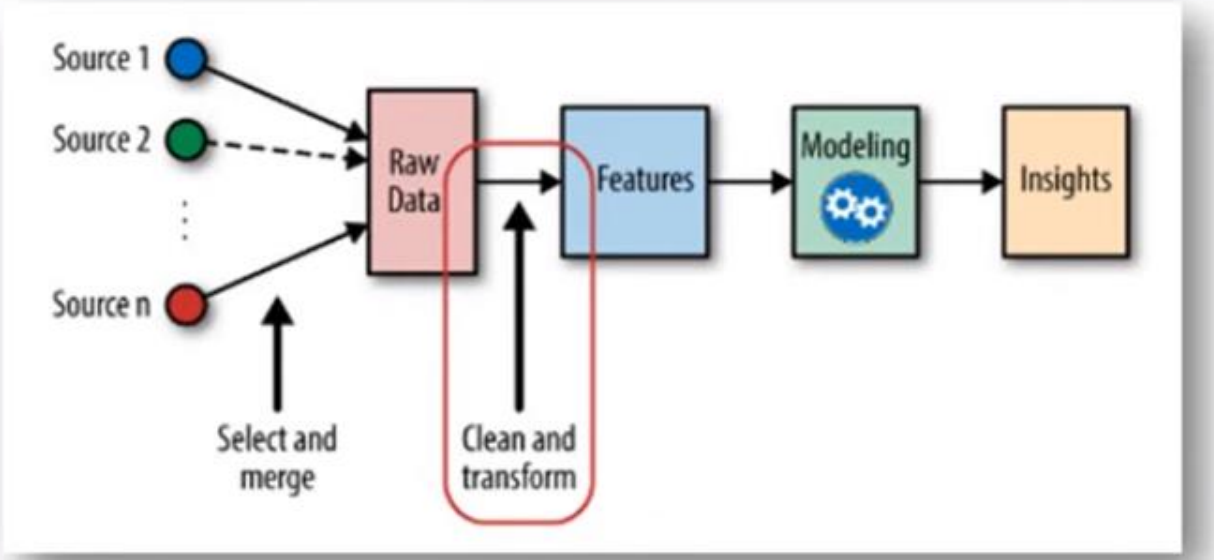
Having too much unnecessary data can cause the model to be slow. The model may also learn from this irrelevant data and be inaccurate

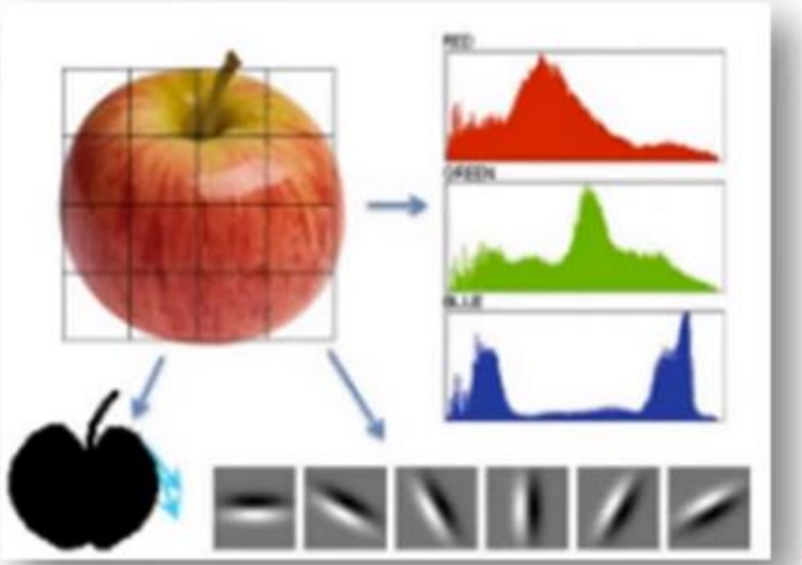
What is Feature Selection?

Feature Selection is the process of reducing the input variable to your model by using only relevant data and getting rid of noise in data

About Features in Machine Learning

- Feature engineering refers to the process of using **domain knowledge to select and transform the most relevant variables from raw data** when **creating a predictive model** using machine learning or statistical modelling.
- The goal of feature engineering and selection is to **improve the performance of machine learning (ML) algorithms.**



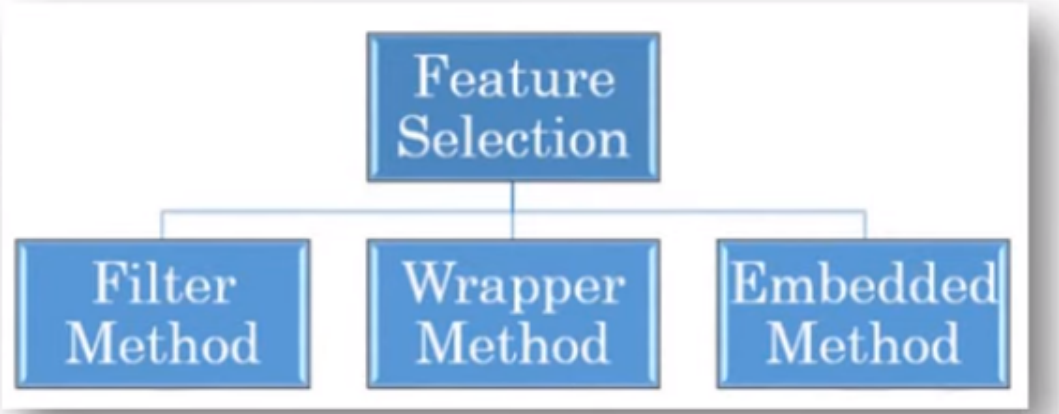


Feature Selection

- Feature selection techniques find a smaller subset of many dimensional data set to create data model.
- The complexity depends on size of **data sample N**, **number of input dimensions d** & **selected dimensions k**.
- It is a technique of finding k features of the d dimensions that gives us the most information & discard the other **(d-k) dimensions**.

Feature Selection Methods:

1. Filter Method
2. Wrapper Method
3. Embedded Method

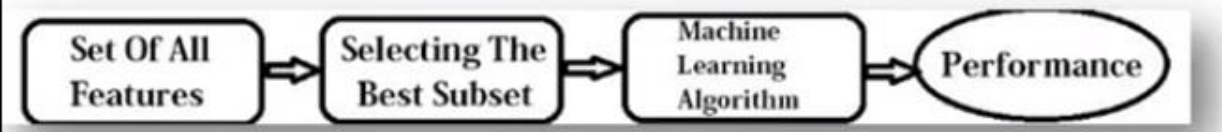


Example: Student Model

1. Filter Method:

- In this method, features are **filtered** based on general characteristics (some metric such as correlation) of the dataset such **correlation with the dependent variable**.
- It is **faster** and usually the better approach when the number of features are huge.
- **Avoids over fitting** but sometimes may fail to select best features.

1. **Information Gain:** Find meaningful information & attribute features.
2. **Chi-Square Taste:** Observed & expected count.
3. **Variance Threshold**: Remove all feature.

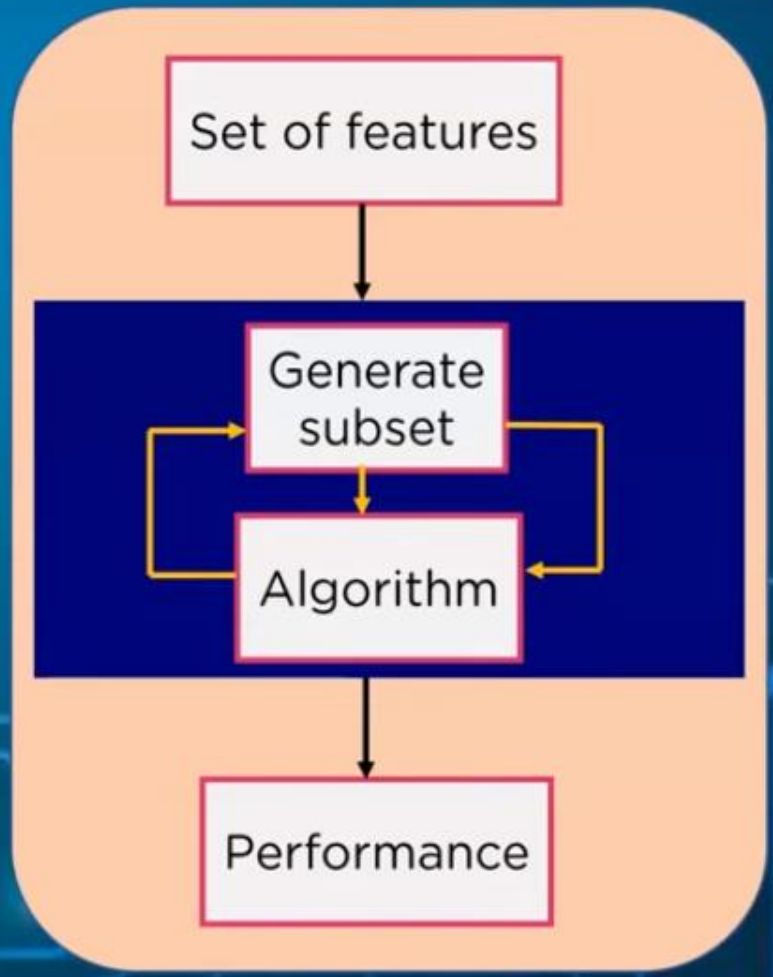


ATTRIBUTES OR FEATURES					TARGET FEATURES
A (%)	B (Roll No)	C (Maths M)	D (C++ M)	E (Name)	T (Target Variable)
75	1	78	60	Neha	Pass
35	2	20	21	Ajay	Fail
90	3	87	89	Rahul	Pass
39	4	25	22	Gauri	Fail

Feature Selection Methods

Wrapper Method

We split our data into subsets and train a model using this. Based on the output of the model, we add and subtract features and train the model again



Example: Student Model

2. Wrapper Method:

- In wrapper method, the feature selection algorithm exits as a **wrapper** around the **predictive model algorithm and uses the same model to select best features**.
- Though computationally **expensive and prone to over fitting**, gives better performance.
- Select **optimal features** from dataset.

M1= A+B

M2=A+C+D

M2= A+E

M4 = B+E

1. **Genetic Algorithms:** Find a subset of features.
2. **Recursive Feature Elimination:**

   Removes the weakest feature
3. **Sequential Feature Selection:**

   Highest number of features are added.

ATTRIBUTES OR FEATURES					TARGET FEATURES
A (%)	B (Roll No)	C (Maths M)	D (C++ M)	E (Name)	T (Target Variable)
75	1	78	60	Neha	Pass
35	2	20	21	Ajay	Fail
90	3	87	89	Rahul	Pass
39	4	25	22	Gauri	Fail

Example: Student Model

3. Embedded Method:

- Embedded methods **combine the qualities of filter and wrapper methods.**
- It's implemented by algorithms that have their **own feature selection methods** in them.

M1= A+B
M2=A+C+D
M2= A+E
M4 = B+E

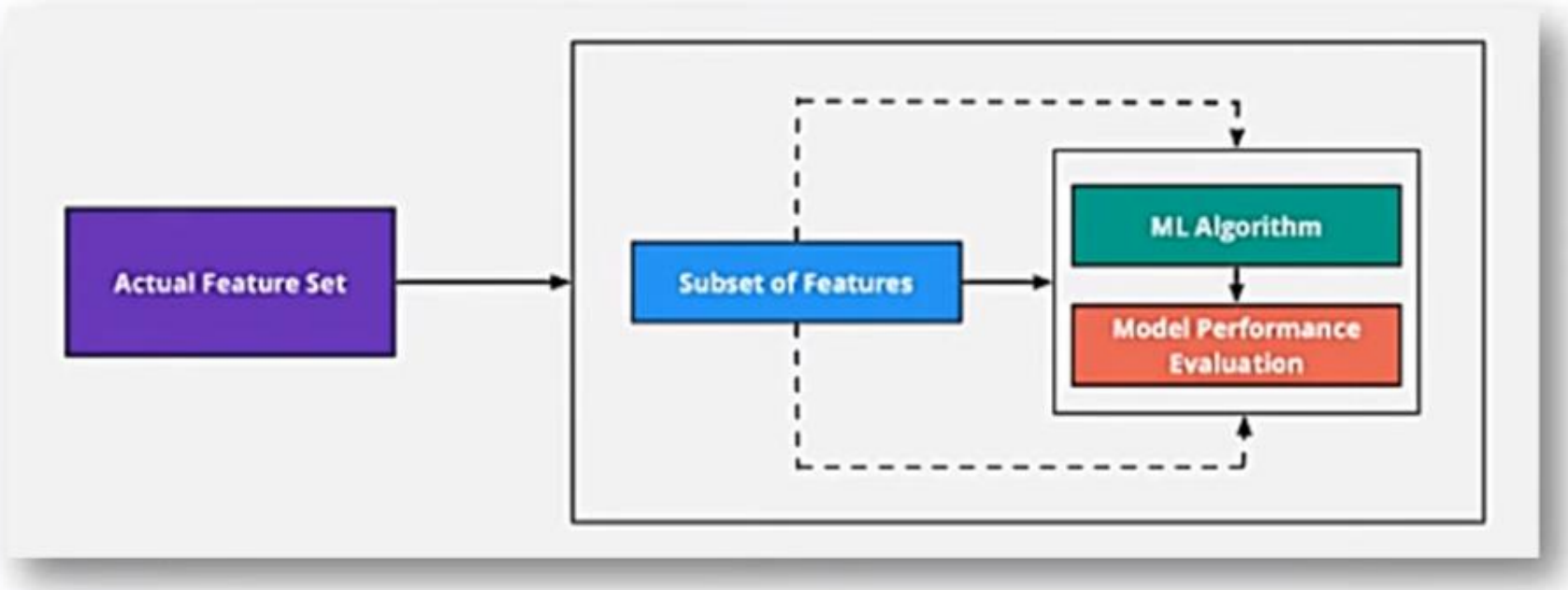
Advantages of Embedded Methods:

- They take into consideration the interaction of features like **wrapper** methods do.
- They are **faster** like filter methods.
- They are **more accurate** than filter methods.
- They find the **feature subset** for the algorithm being trained.
- They are much less prone to **over-fitting**.

ATTRIBUTES OR FEATURES					TARGET FEATURES
A (%)	B (Roll No)	C (Maths M)	D (C++ M)	E (Name)	T (Target Variable)
75	1	78	60	Neha	Pass
35	2	20	21	Ajay	Fail
90	3	87	89	Rahul	Pass
39	4	25	22	Gauri	Fail

Benefits of Feature Selection Methods

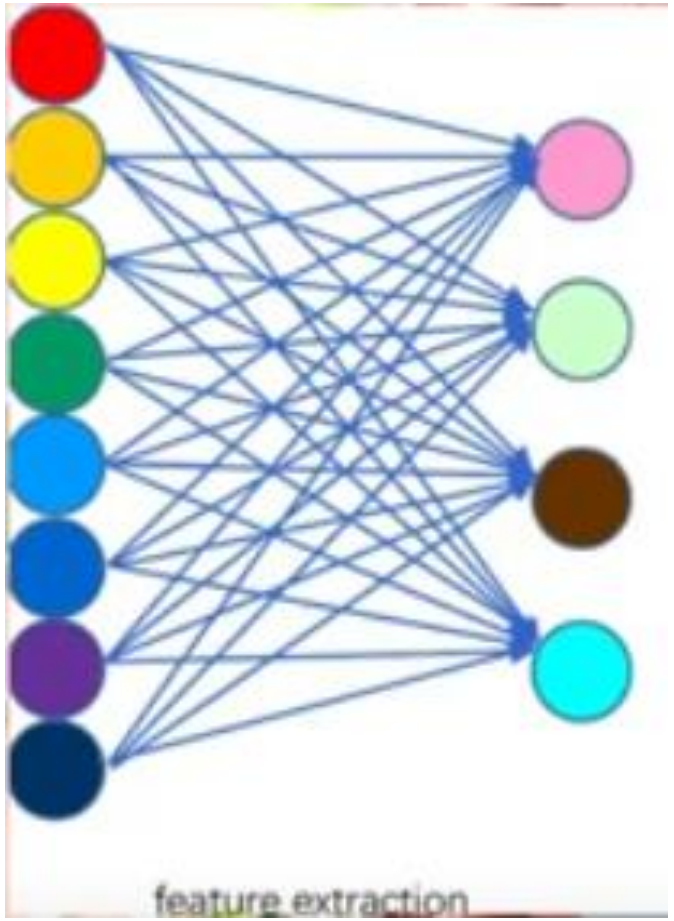
- **Reduces Overfitting**: Less redundant data means **less opportunity to make decisions** based on noise.
- **Improves Accuracy**: Less misleading data means **modelling accuracy improves**.
- **Reduces Training Time**: fewer data points reduce algorithm complexity and **algorithms train faster.**



Feature Extraction

Feature Extraction aims to reduce the number of features in a dataset by creating new features from the existing ones (and then discarding the original features).

- These new reduced set of features should then be able to summarize most of the information contained in the original set of features.
- In this way, a summarised version of the original features can be created from a combination of the original set.



When we actually work in real world machine learning problem then we rarely get data in shape of CSV so we have to extract the useful features from the raw data.

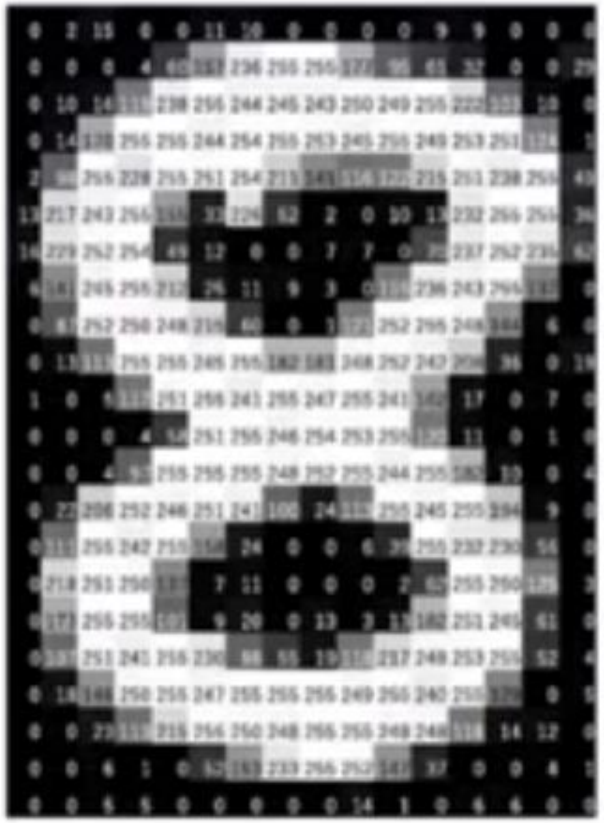
Some of the popular types of raw data from which features(new feature creation) can be extracted.

- Texts
- Images
- Geospatial data
- Date and time
- Web data
- Sensors Data

Feature Extraction in Machine Learning (Image)

0	2	15	0	0	11	10	0	0	0	0	9	9	0	0	0
0	0	0	4	60	157	236	255	255	177	95	61	32	0	0	29
0	10	16	119	238	255	244	245	243	250	249	255	222	103	10	0
0	14	170	255	255	244	254	255	253	245	255	249	253	251	124	1
2	98	255	228	255	251	254	211	141	116	122	215	251	238	255	49
13	217	243	255	155	33	226	52	2	0	10	13	232	255	255	36
16	229	252	254	49	12	0	0	7	7	0	70	237	252	235	62
6	141	245	255	212	25	11	9	3	0	115	236	243	255	137	0
0	87	252	250	248	215	60	0	1	121	252	255	248	144	6	0
0	13	113	255	255	245	255	182	181	248	252	242	208	36	0	19
1	0	5	117	251	255	241	255	247	255	241	162	17	0	7	0
0	0	0	4	58	251	255	246	254	253	255	120	11	0	1	0
0	0	4	97	255	255	255	248	252	255	244	255	182	10	0	4
0	22	206	252	246	251	241	100	24	113	255	245	255	194	9	0
0	111	255	242	255	158	24	0	0	6	39	255	232	230	56	0
0	218	251	250	137	7	11	0	0	0	2	62	255	250	125	3
0	173	255	255	101	9	20	0	13	3	13	182	251	245	61	0
0	107	251	241	255	230	98	55	19	118	217	248	253	255	52	4
0	18	146	250	255	247	255	255	255	249	255	240	255	129	0	5
0	0	23	113	215	255	250	248	255	255	248	248	118	14	12	0
0	0	6	1	0	52	153	233	255	252	147	37	0	0	4	1
0	0	5	5	0	0	0	0	0	14	1	0	6	6	0	0

Machines store images in the form of a matrix of numbers.



These numbers, or the pixel values, denote the intensity or brightness of the pixel. Smaller numbers (closer to zero) represent black, and larger numbers (closer to 255) denote white.

Feature Extraction

- Feature extraction involves transforming **high dimensional data into space of fewer dimensional.**
- Features extraction is a technique of finding a **new set of k dimensions that are combinations of original d dimensions.**
- The best known & most widely used feature extraction method is **Principle Component Analysis (PCA).**

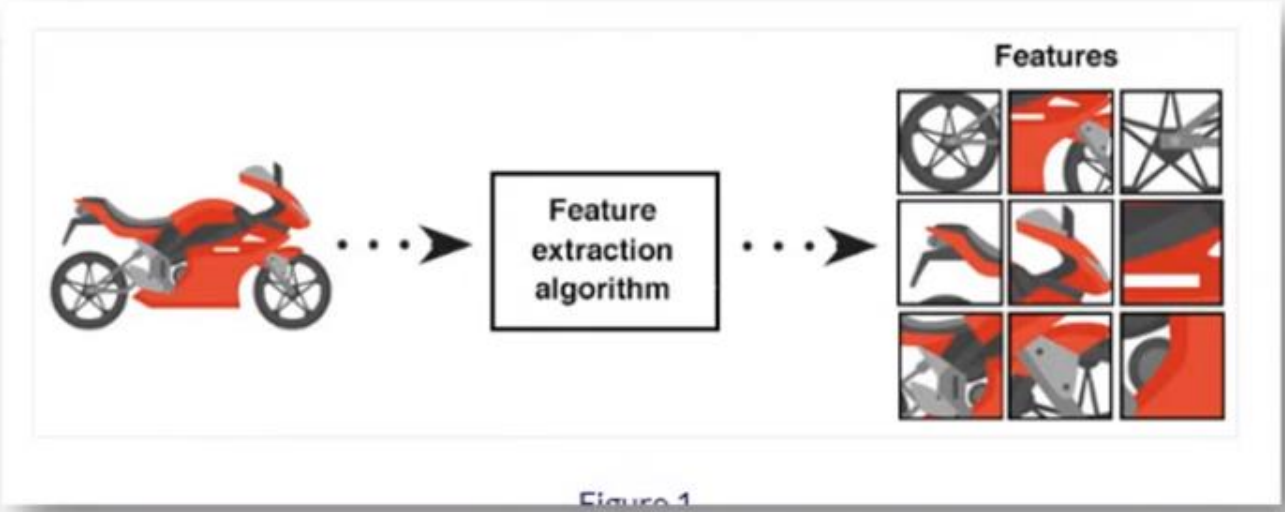


Figure 1

Here are four ways feature extraction enables machine learning models to better serve their intended purpose:

Reduces redundant data

- Feature extraction cuts through the noise, removing redundant and unnecessary data. This frees machine learning programs to focus on the most relevant data.

Improves model accuracy

- The most accurate machine learning models are those developed using only the data required to train the model to its intended business use. Including peripheral data negatively impacts the model's accuracy.

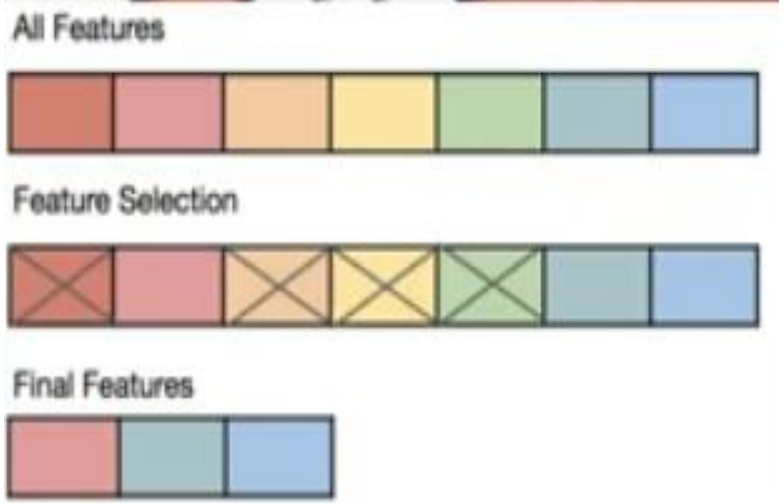
Boosts speed of learning

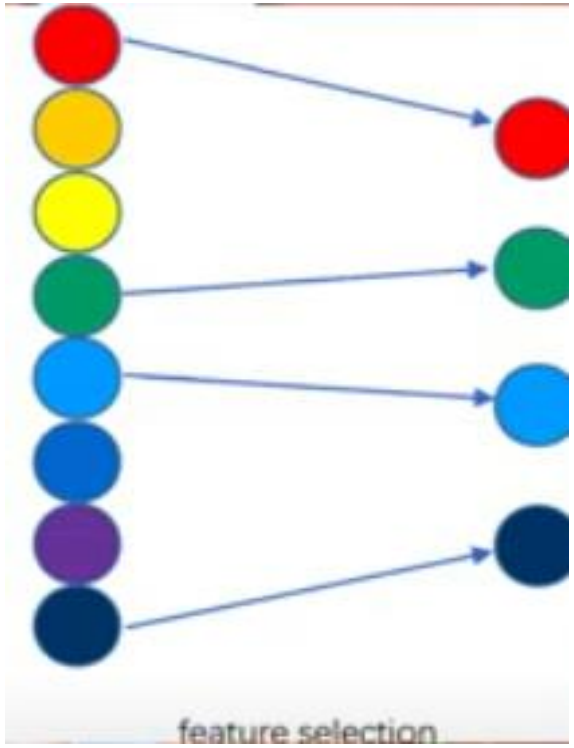
- Including training data that doesn't directly contribute to solving the business problem bogs down the learning process. Models trained on highly relevant data learn more quickly and make more accurate predictions.

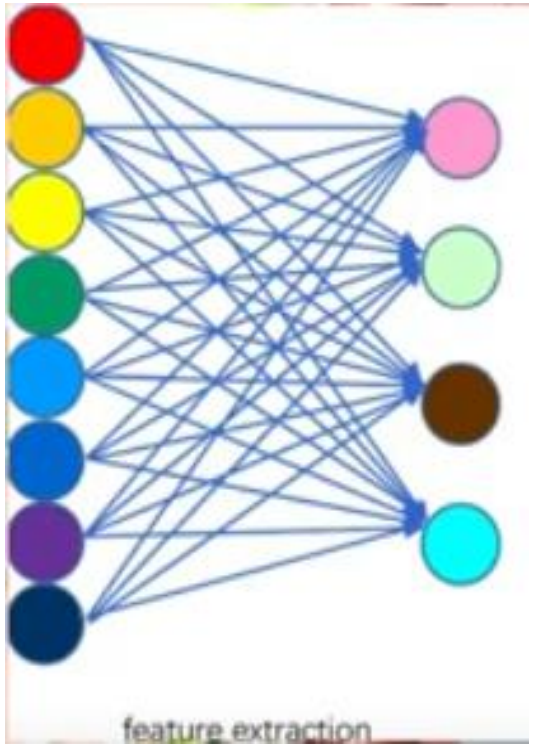
More-efficient use of compute resources

- Pruning out peripheral data boosts speed and efficiency. With less data to sift through, compute resources aren't dedicated to processing tasks that aren't generating additional value.

Feature Selection VS Extraction







- The difference between Feature Selection and Feature Extraction is that feature selection aims instead to rank the importance of the existing features in the dataset and discard less important ones (no new features are created).

Concepts of Probability

Motivation

- Uncertainty arises through:
- Noisy measurements
- Finite size of data sets
- Ambiguity: The word bank can mean (1) a financial institution, (2) the side of a river, or (3) tilting an airplane. Which meaning was intended, based on the words that appear nearby?
- Probability theory provides a consistent framework for the quantification and manipulation of uncertainty
- Allows us to make optimal predictions given all the information available to us, even though that information may be incomplete or ambiguous

- In our day to day life, we use the concept of probability in many places.
- Probability represents the certainty factor.
- Certainty is the rate that you would assign to an event to happen.
- Say, you are rolling a dice and you say that the certainty with which a 6 shows up on the dice is ⅙. It means there's a 16.67% chance that a 6 shows up on the dice.
- That's the certainty you allot to that particular event. This, in turn, is known as **probability**, or precisely, in our case, it's called **frequentist probability**.

- The **frequentist** probability denotes the frequency with which the event can happen amongst many trials/events.
- Rolling a dice is frequentist as ⅙ means that out of infinitely many trials of rolling a dice, there's a 1/6th chance that 6 is going to show up.

- Not all scenarios are frequency related as in our previous assumption.
- If we consider a machine learning problem in which we estimate the probability of inflation or deflation of the price of fuel, we wouldn’t be thinking this in the perspective of repetition, as seen in the frequentist probability scenario.
- Instead, we say that this event could occur with a certain probability/certainty.

- The latter phenomenon is called **Bayesian probability**.
- Rather than considering the frequency with which an event repeats, we quantify our belief.
- Consider the statement — there's a 32% chance that a diabetic patient is going to develop heart failure. This statement isn't prone to repetition where we create infinite replicas of the patient's symptoms. We instead quantify with a 32% certainty that heart failure could happen.

Foundation rules

- p(A) denotes the probability that the event A is true.
- For example, A might be the logical statement ‘Brazil is going to win the next football world cup final’.
- The expression $0 \leq p(A) \leq 1$ denotes that the probability of this event happening lies between 0 and 1,
- where $p(A) = 0$ means the event will definitely not happen, and
- $p(A) = 1$ means the event will definitely happen.

Sample Space

Sample Space: The set of all possible outcomes of an experiment is called the sample space and is denoted by Ω.
Individual elements are denoted by ω and are termed elementary outcomes.
Examples:

- (Finite) A single roll of an ordinary die. Here, $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- (Countable) Infinite number of coin tosses in order to study, say, the number of tosses before 5 consecutive heads are observed. Here, $\Omega = \{H, T\}^{\infty}$.
- (Uncountable) Speed of a vehicle measured with infinite precision. Here, $\Omega = \mathbb{R}$.

Event

Event: An event is any collection of possible outcomes of an experiment, that is, any subset of Ω.

In most experiments we are generally more interested in observing the occurrence of particular events rather than the elementary outcomes. For example, on rolling a die, we may be interested in observing whether the outcome was even (event $E = \{2, 4, 6\}$) or odd (event $O = \{1, 3, 5\}$).

Set Theory Notations

$$A \subset B \Leftrightarrow x \in A \Rightarrow x \in B$$

$$A = B \Leftrightarrow A \subset B \text{ and } B \subset A$$

$$A \cup B = \{x : x \in A \text{ or } x \in B\}$$

$$A \cap B = \{x : x \in A \text{ and } x \in B\}$$

$$A^c = \{x : x \notin A\}$$

Properties of Set Operations

Commutativity

$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

Associativity

$$A \cup (B \cup C) = (A \cup B) \cup C$$
$$A \cap (B \cap C) = (A \cap B) \cap C$$

Distributivity

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

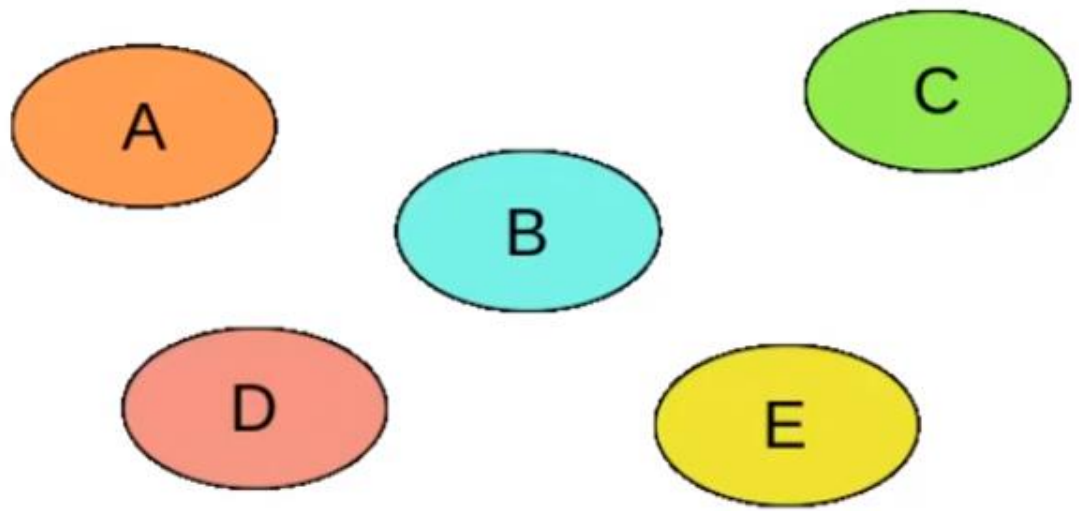
DeMorgan's Laws

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

Disjoint Events

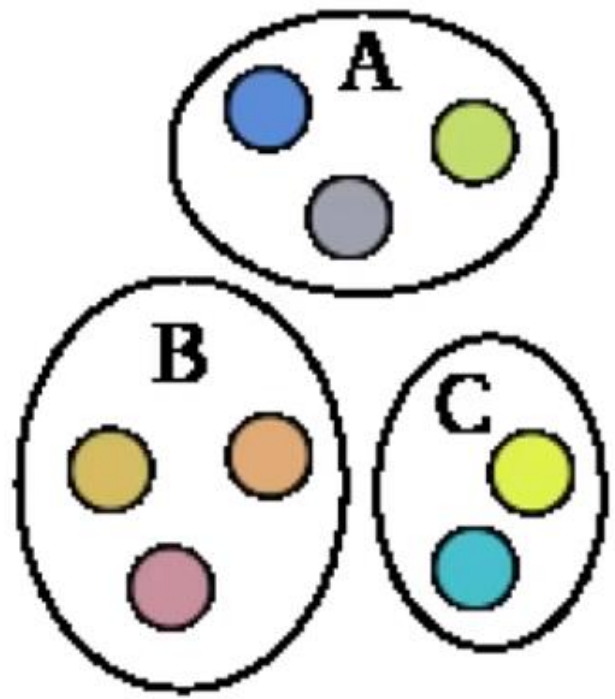
Two events A and B are disjoint (or mutually exclusive) if $A \cap B = \phi$.

A sequence of events $A_1, A_2, A_3, \ldots$ are pair-wise disjoint if $A_i \cap A_j = \phi$ for all $i \neq j$.



Partition

If $A_1, A_2, \ldots$ are pair-wise disjoint and $\cup_{i=1}^{\infty} A_i = \Omega$, then the collection $A_1, A_2, \ldots$ forms a partition of Ω.



Sigma Algebra

Given a sample space Ω, a σ-algebra is a collection $\mathcal{F}$ of subsets of Ω, with the following properties:

(a) $\Phi \in \mathcal{F}$.
(b) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
(c) If $A_i \in \mathcal{F}$ for every $i \in \mathbb{N}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

A set A that belongs to $\mathcal{F}$ is called an $\mathcal{F}$-measurable set (event).

Example: Consider $\Omega = \{1, 2, 3\}$.
$\mathcal{F}_1 = \{\phi, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$.
$\mathcal{F}_2 = \{\phi, \{1, 2, 3\}\}$.

Sample Space Size Considerations

For any Ω (countable or uncountable) 2^{Ω} is always a σ-algebra.

For example, for $\Omega = \{H, T\}$, a feasible σ-algebra is the power set, i.e., $\mathcal{F} = \{\phi, \{H\}, \{T\}, \{H, T\}\}$.

However, if Ω is uncountable, then probabilities cannot be assigned to every subset of 2^{Ω}.

Probability Measure & Probability Space

A probability measure $\mathcal{P}$ on $(\Omega, \mathcal{F})$ is a function $\mathcal{P} : \mathcal{F} \to [0, 1]$ satisfying

(a) $\mathcal{P}(\phi) = 0, \quad \mathcal{P}(\Omega) = 1;$
(b) if $A_1, A_2, \ldots$ is a collection of pair-wise disjoint members of $\mathcal{F}$, then

$$\mathcal{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathcal{P}(A_i)$$

The triple $(\Omega, \mathcal{F}, \mathcal{P})$, comprising a set Ω, a σ-algebra $\mathcal{F}$ of subsets of Ω, and a probability measure $\mathcal{P}$ on $(\Omega, \mathcal{F})$, is called a **probability space**.

Example

Consider a simple experiment of rolling an ordinary die in which we want to identify whether the outcome results in a prime number or not.

$\Omega = \{1, 2, 3, 4, 5, 6\}$

$\mathcal{F} = \{\phi, \{1, 4, 6\}, \{2, 3, 5\}, \{1, 2, 3, 4, 5, 6\}\}$

$\mathcal{P} : \mathcal{F} \rightarrow [0, 1]$

- $\mathcal{P}(\phi) = 0$
- $\mathcal{P}(\{1, 4, 6\}) = 0.5$
- $\mathcal{P}(\{2, 3, 5\}) = 0.5$
- $\mathcal{P}(\Omega) = 1$

Bonferroni's Inequality

$$P(A \cap B) \geq P(A) + P(B) - 1$$

General form:

$$P(\cap_{i=1}^{n} A_i) \geq \sum_{i=1}^{n} P(A_i) - (n-1)$$

Gives a lower bound on the intersection probability which is useful when this probability is hard to calculate.

Only useful if the probabilities of individual events are sufficiently large.

Boole's Inequality

$$P(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i), \text{ for } \textbf{any} \text{ sets } A_1, A_2, \ldots$$

Gives a useful upper bound for the probability of the union of events.

Conditional Probability

Given two events A and B, if $P(B) > 0$, then the conditional probability that A occurs given that B occurs is defined to be

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Essentially, since event B has occurred, it becomes the new sample space.

Conditional probabilities are useful when reasoning in the sense that once we have observed some event, our beliefs or predictions of related events can be updated/improved.

Example

Q. A fair coin is tossed twice. What is the probability that both tosses result in heads given that at least one of the tosses resulted in a heads?

Sol. $\Omega = \{HH, TT, HT, TH\}$
$\mathcal{P}(HH) = \mathcal{P}(TT) = \mathcal{P}(HT) = \mathcal{P}(TH) = 1/4$

$$\begin{aligned}
&\mathcal{P}(HH|\text{at least one toss heads}) \\
&\quad = \mathcal{P}(HH|HT \cup TH \cup HH) \\
&\quad = \frac{\mathcal{P}(HH \cap (HT \cup TH \cup HH))}{\mathcal{P}(HT \cup TH \cup HH)} \\
&\quad = \frac{\mathcal{P}(HH)}{\mathcal{P}(HT \cup TH \cup HH)} \\
&\quad = \frac{1}{3}
\end{aligned}$$

Bayes' Rule

We have:

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(A \cap B)}{\mathcal{P}(B)}$$

$$\mathcal{P}(A \cap B) = \mathcal{P}(A|B)\mathcal{P}(B)$$

$$\mathcal{P}(A \cap B) = \mathcal{P}(B|A)\mathcal{P}(A)$$

$$\mathcal{P}(A|B)\mathcal{P}(B) = \mathcal{P}(B|A)\mathcal{P}(A)$$

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(B|A)\mathcal{P}(A)}{\mathcal{P}(B)} \text{ (Bayes' Rule)}$$

Bayes' Rule

Let $A_1, A_2, \ldots$ be a partition of the sample space, and let B be any subset of the sample space. Then, for each i = 1, 2, ...,

$$\mathcal{P}(A_i|B) = \frac{\mathcal{P}(B|A_i)P(A_i)}{\sum_{j=1}^{\infty} \mathcal{P}(B|A_j)\mathcal{P}(A_j)}$$

Bayes' rule is important in that it allows us to compute the conditional probability $\mathcal{P}(A|B)$ from the "inverse" conditional probability $\mathcal{P}(B|A)$.

Example

Q. To answer a multiple choice question, a student may either know the answer or may guess it. Assume that with probability p the student knows the answer to a question, and with probability q, the student guesses the right answer to a question she does not know. What is the probability that for a question the student answers correctly, she actually knew the answer to the question?

Sol. Let K be the event that the student knows the question, and C be the event that the student answers the question correctly. We have $\mathcal{P}(K) = p$, $\mathcal{P}(\neg K) = 1 - p$, $\mathcal{P}(C|K) = 1$, $\mathcal{P}(C|\neg K) = q$

$$\begin{aligned}
&\mathcal{P}(K|C) \\
&= \frac{\mathcal{P}(C|K)\mathcal{P}(K)}{\mathcal{P}(C)} \\
&= \frac{\mathcal{P}(C|K)\mathcal{P}(K)}{\mathcal{P}(K)\mathcal{P}(C|K)+\mathcal{P}(\neg K)\mathcal{P}(C|\neg K)} \\
&= \frac{p}{p+q(1-p)}
\end{aligned}$$

Independent Events

Two events, A and B, are said to be independent if

$$\mathcal{P}(A \cap B) = \mathcal{P}(A)\mathcal{P}(B)$$

More generally, a family $A_i : i \in I$ is called independent if

$$\mathcal{P}(\cap_{i \in J} A_i) = \prod_{i \in J} \mathcal{P}(A_i)$$

for all finite subsets J of I.

From the above, it should be clear that pair-wise independence does not imply independence.

Conditional Independence

Let A, B, and C be three events with $\mathcal{P}(C) > 0$. The events A and B are called conditionally independent *given* C if

$$\mathcal{P}(A \cap B|C) = \mathcal{P}(A|C)\mathcal{P}(B|C)$$

or equivalently

$$\mathcal{P}(A|B \cap C) = \mathcal{P}(A|C)$$

Example: Assume that admission into the M.Tech. programme at IITM & IITB is based solely on candidate's GATE score. Then

$$\mathcal{P}(IITM|IITB \cap GATE) = \mathcal{P}(IITM|GATE)$$

Random Variable

A random variable is a function $X : \Omega \to \mathbb{R}$, i.e., it is a function from the sample space to the real numbers.

Examples:

- The sum of outcomes on rolling 3 dice.
- The number of heads observed when tossing a fair coin 3 times.

Induced Probability Function

Consider the previous example experiment of tossing a fair coin 3 times. Let X be the number of heads obtained in the three tosses. Enumerating the elementary outcomes, we observe the value of X as

ω	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
$X(\omega)$	3	2	2	2	1	1	1	0

Instead of using the probability measure defined on the elementary outcomes or events, we would ideally like to measure the probability of the random variable taking on values in its range.

x	0	1	2	3
$\mathcal{P}_X(X = x)$	1/8	3/8	3/8	1/8

Induced Probability Function

Let $\Omega = \{\omega_1, \omega_2, ...\}$ be a sample space and $\mathcal{P}$ be a probability measure (function).

Let X be a random variable with range $\mathcal{X} = \{x_1, x_2, ..., x_m\}$.

We define the induced probability function $\mathcal{P}_X$ on $\mathcal{X}$ as

$$\mathcal{P}_X(X = x_i) = \mathcal{P}(\{\omega_j \in \Omega : X(\omega_j) = x_i\})$$

Cumulative Distribution Function

The cumulative distribution function or cdf of a random variable X, denoted by $F_X(x)$, is defined by

$$F_X(x) = \mathcal{P}_X(X \leq x), \text{ for all } x$$

Example:

x	$(-\infty, 0]$	$(-\infty, 1]$	$(-\infty, 2]$	$(-\infty, 3]$	$(-\infty, \infty)$
$F_X(x)$	$1/8$	$1/2$	$7/8$	1	1

1/8+3/8=4/8=1/2

Properties of cdf

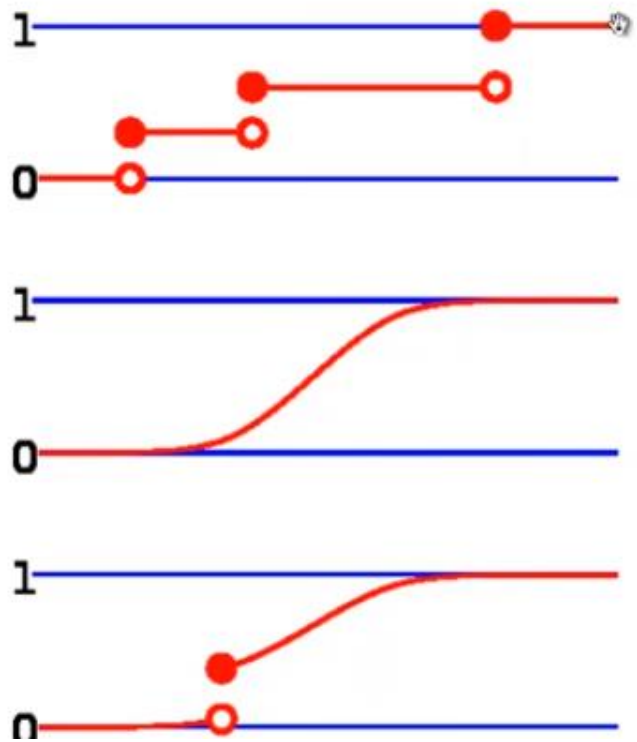
A function $F_X(x)$ is a cdf iff the following three conditions hold:

- (Monotonicity) If $x \leq y$, then $F_X(x) \leq F_X(y)$
- (Limiting values) $\lim_{x \to -\infty} F_X(x) = 0$ and $lim_{x \to \infty} F_X(x) = 1$
- (Right-continuity) For every x, we have $\lim_{y \downarrow x} F_X(y) = F_X(x)$

Continuous & Discrete Random Variables

A random variable X is continuous if $F_X(x)$ is a continuous function of x.

A random variable X is discrete if $F_X(x)$ is a step function of x.



Probability Mass Function

The probability mass function or pmf of a discrete random variable X is given by

$$f_X(x) = \mathcal{P}(X = x), \text{ for all x}$$

Example: For a geometric random variable X with parameter p,

$$f_X(x) = \begin{cases} (1-p)^{x-1}p & \text{for } x = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Properties:

- $f_X(x) \geq 0$, for all x
- $\sum_x f_X(x) = 1$

Probability Density Function

The probability density function or pdf of a continuous random variable is the function $f_X(x)$ which satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt, \text{ for all x}$$

Properties:

- $f_X(x) \geq 0$, for all x
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$

Expectation

The expected value or mean of a random variable X, denoted by $E[X]$, is given by

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx \text{ (continuous RV)}$$

$$E[X] = \sum_{x:\mathcal{P}(x)>0} x f_X(x) = \sum_{x:\mathcal{P}(x)>0} x \mathcal{P}(X = x) \text{ (discrete RV)}$$

Example

Q. Let the random variable X take values -2, -1, 1, 3 with probabilities 1/4, 1/8, 1/4, 3/8 respectively. What is the expectation of the random variable $Y = X^2$?

Sol. The random variable Y takes on the values 1, 4, 9 with probabilities 3/8, 1/4, 3/8 respectively.
Hence,

$$E(Y) = \sum_x xP(Y = x) = 1 \cdot \frac{3}{8} + 4 \cdot \frac{1}{4} + 9 \cdot \frac{3}{8} = \frac{19}{4}$$

Alternatively,

$$E(Y) = E(X^2) = \sum_x x^2 P(X = x) = 4 \cdot \frac{1}{4} + 1 \cdot \frac{1}{8} + 1 \cdot \frac{1}{4} + 9 \cdot \frac{3}{8} = \frac{19}{4}$$

Properties of Expectations

Let X be a random variable and let a, b, c be constants. Then, for functions $g_1(X)$ and $g_2(X)$ whose expectations exist

- $E(ag_1(X) + bg_2(X) + c) = aEg_1(X) + bEg_2(X) + c$
- If $g_1(X) \geq 0$ for all x, then $Eg_1(X) \geq 0$
- If $g_1(X) \geq g_2(X)$ for all x, then $Eg_1(X) \geq Eg_2(X)$
- If $a \leq g_1(X) \leq b$, for all x, then $a \leq Eg_1(X) \leq b$

Moments

For each integer n, the n^{th} moment of X is

$$\mu'_n = EX^n$$

The n^{th} central moment of X is

$$\mu_n = E(X - \mu)^n$$

Variance

The variance of a random variable X is its second central moment.
$VarX = E(X - \mu)^2 = E(X - EX)^2 = EX^2 - (EX)^2$

The positive square root of $VarX$ is the standard deviation of X.

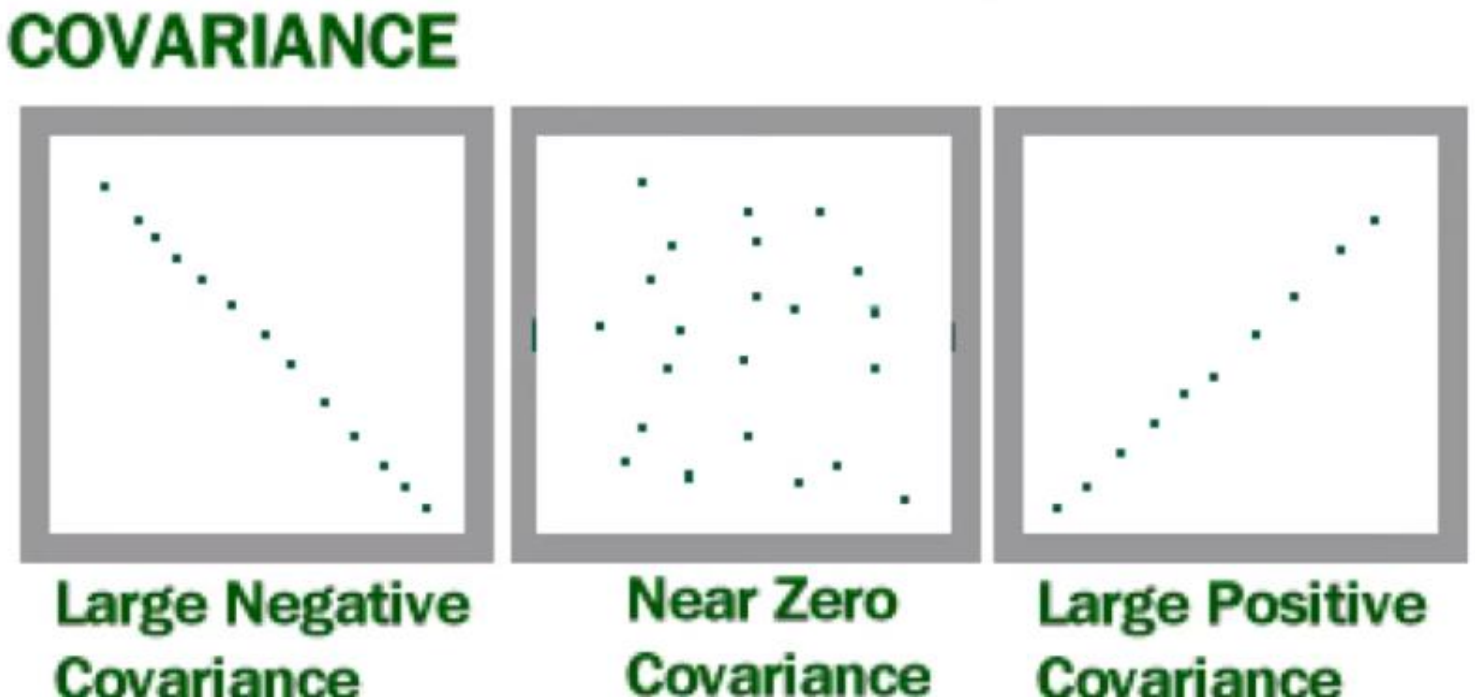
Note: $Var(aX + b) = a^2 VarX$
where a, b are constants

Covariance

The covariance of two random variables, X and Y is

$$cov(X, Y) = E[(X - EX)(Y - EY)]$$

It is a measure of how much two random variables change together.



Correlation

The correlation of two random variables, X and Y is

$$\rho(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)var(Y)}}$$

Note:

- For correlation to be defined, individual variances must be non-zero and finite
- $\rho(X, Y)$ lies between -1 and $+1$

Probability Distributions

Consider two variables X and Y, and suppose we know the corresponding probability mass functions f_X and f_Y

Can we answer the following question:

$\mathcal{P}(X = x \text{ and } Y = y) = ?$

Joint Distributions

To capture the properties of two random variables X and Y, we use the joint PMF

$f_{X,Y} : \mathbb{R}^2 \to [0, 1]$, defined by
$f_{X,Y}(x, y) = \mathcal{P}(X = x, Y = y)$

Marginal Distributions

Suppose we are given the joint PMF

$$f_{X,Y}(x,y) = \mathcal{P}(X = x, Y = y)$$

From this joint PMF, we can obtain the PMF's of the two random variables

$f_X = \sum_y f_{X,Y}(x,y)$ (marginal PMF of R.V. X)
$f_Y = \sum_x f_{X,Y}(x,y)$ (marginal PMF of R.V. Y)

Conditional Distributions

Like joint distributions, we can also consider conditional distributions

$$f_{X|Y}(x|y) = \mathcal{P}(X = x|Y = y)$$

Using conditional probability definition, we have

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

Note that the above conditional probability is undefined if $f_Y(y) = 0$.

Bernoulli Distribution

Consider a random variable X taking one of two possible values (either 0 or 1). Let the PMF of X be given by

$$f_X(0) = \mathcal{P}(X = 0) = 1 - p \qquad (0 \leq p \leq 1)$$
$$f_X(1) = \mathcal{P}(X = 1) = p$$

This describes a Bernoulli distribution

$$E[X] = p$$
$$var(X) = p(1 - p)$$

Binomial Distribution

Consider the situation where we perform n independent Bernoulli trials where

- probability of success (for each trial) $= p$
- probability of failure $= 1 - p$

Let X be the number of successes in the n trials, then we have

$$\mathcal{P}(X = x|n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

where $\binom{n}{x} = \frac{n!}{(n-x)!x!}$
and $0 \leq x \leq n$

$E[X] = np$
$var(X) = np(1-p)$

Geometric Distribution

Suppose we perform a series of independent Bernoulli trials, each with a probability p of success. Let X represent the number of trials before the first success, then we have

$$\mathcal{P}(X = x|p) = (1-p)^{x-1}p \qquad x = 1, 2, 3, ...$$

$$E[X] = 1/p$$
$$var(X) = (1-p)/p^2$$

Uniform Distribution

A continuous random variable X is said to be uniformly distributed on an interval $[a, b]$ if its PDF is given by

$$f_X(x|a,b) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$E[X] = (a+b)/2$
$var(X) = (b-a)^2/12$