BITS Pilani presentation

BITS Pilani
Pilani Campus

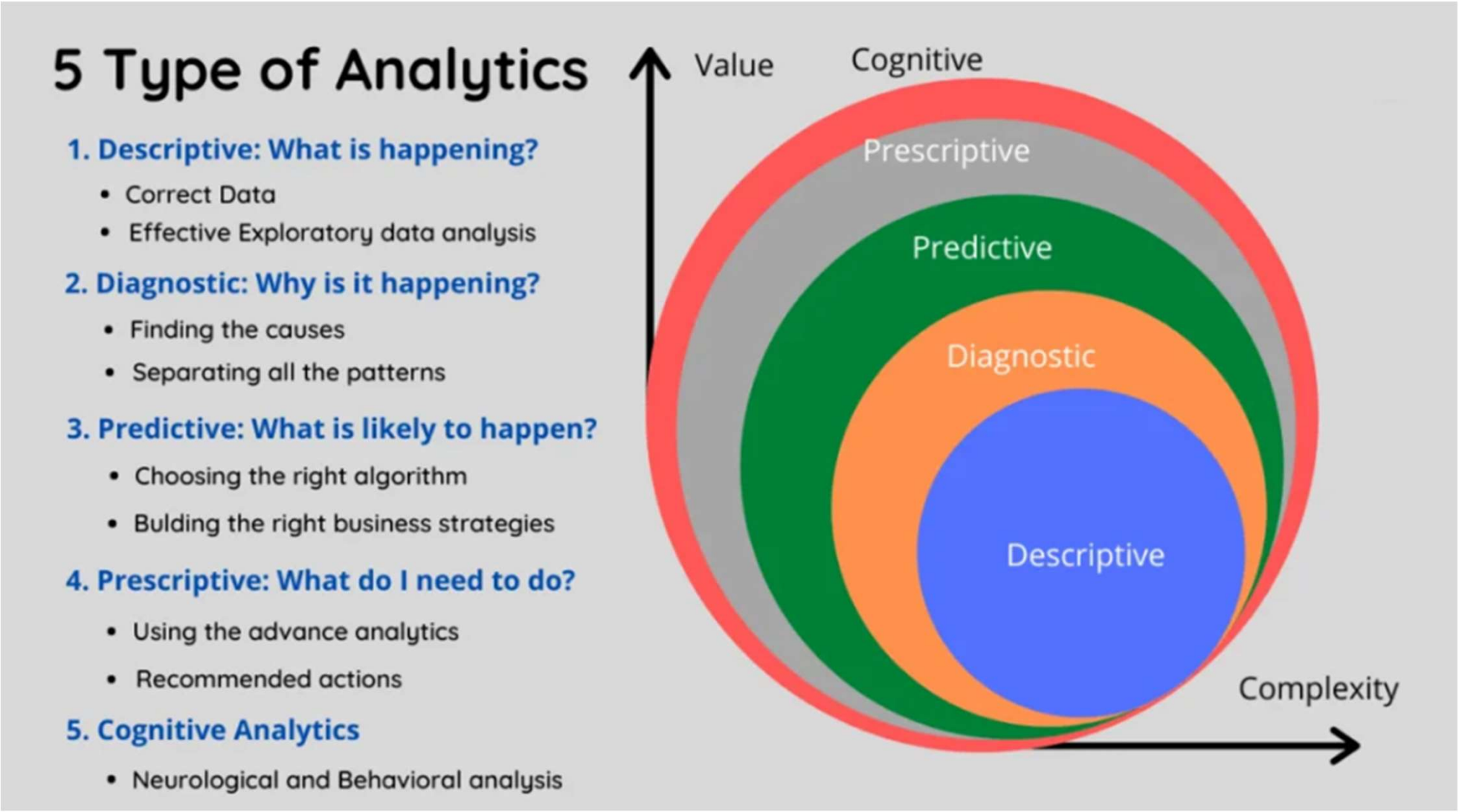
Dr. Vivek V. Jog
Dept. Of Computer Engineering

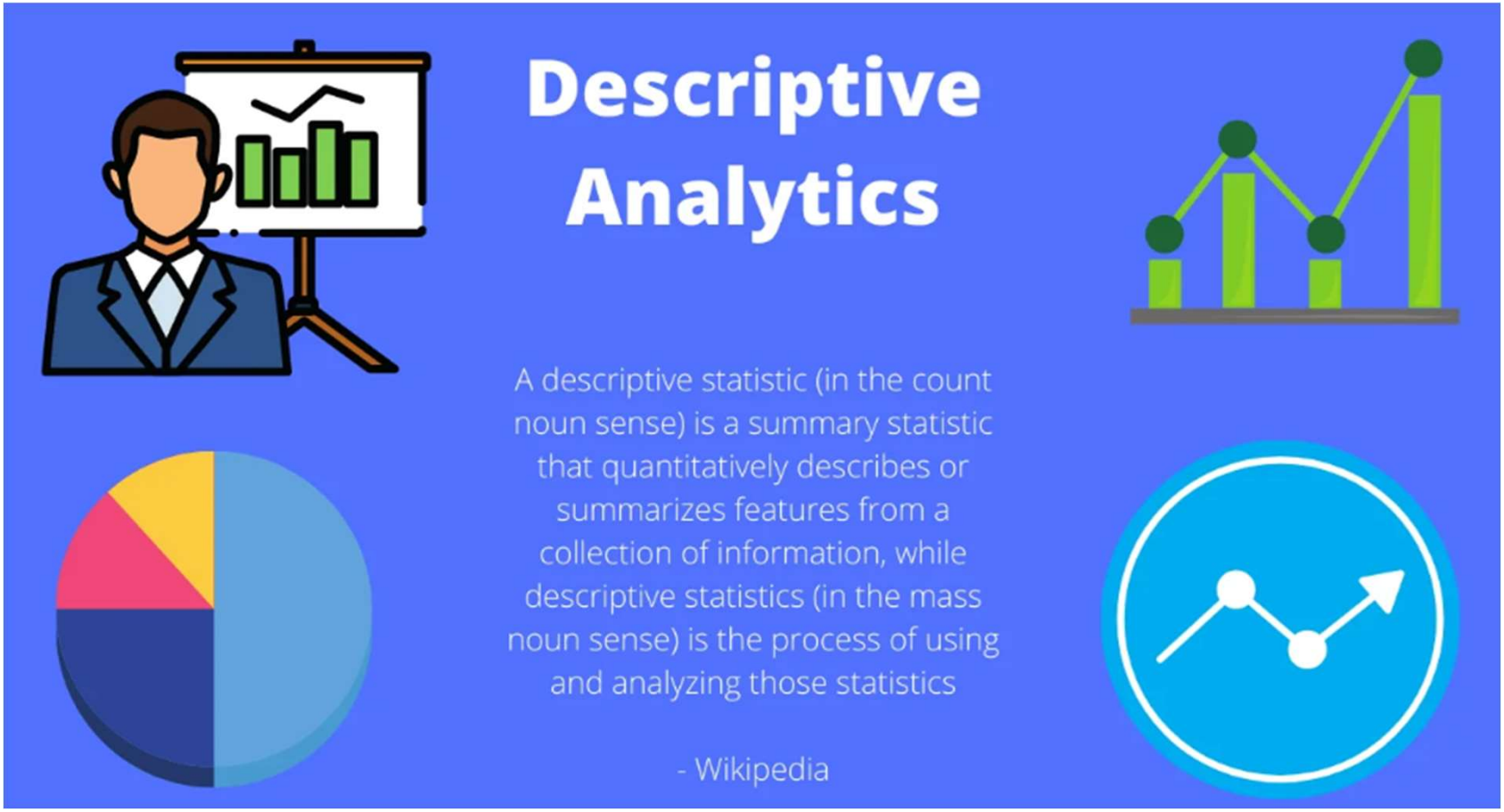
innovate achieve lead

BITS Pilani
Pilani Campus

Big Data Systems (S1-24_CCZG522)
Lecture No.3

Data Analytics types

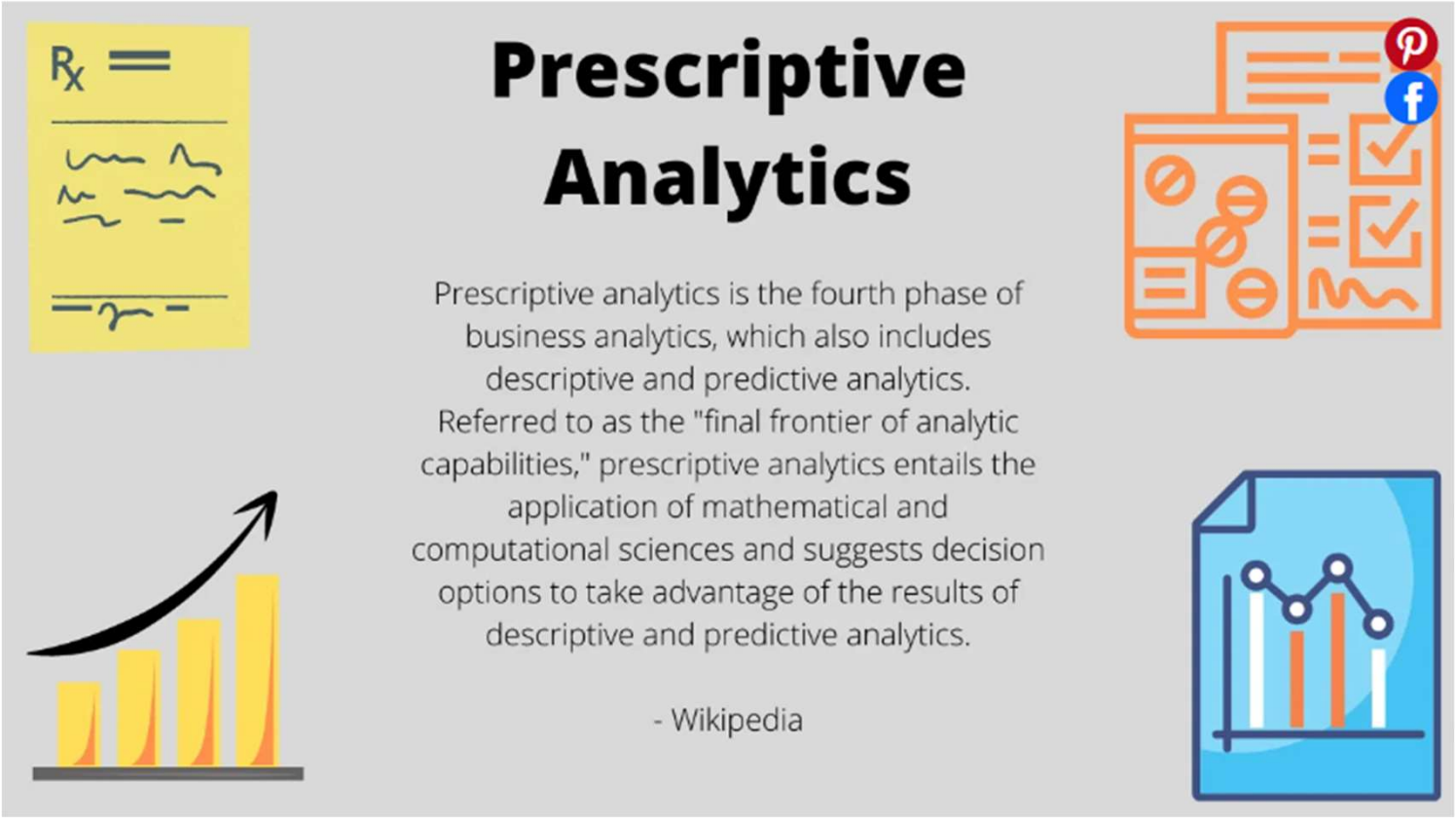




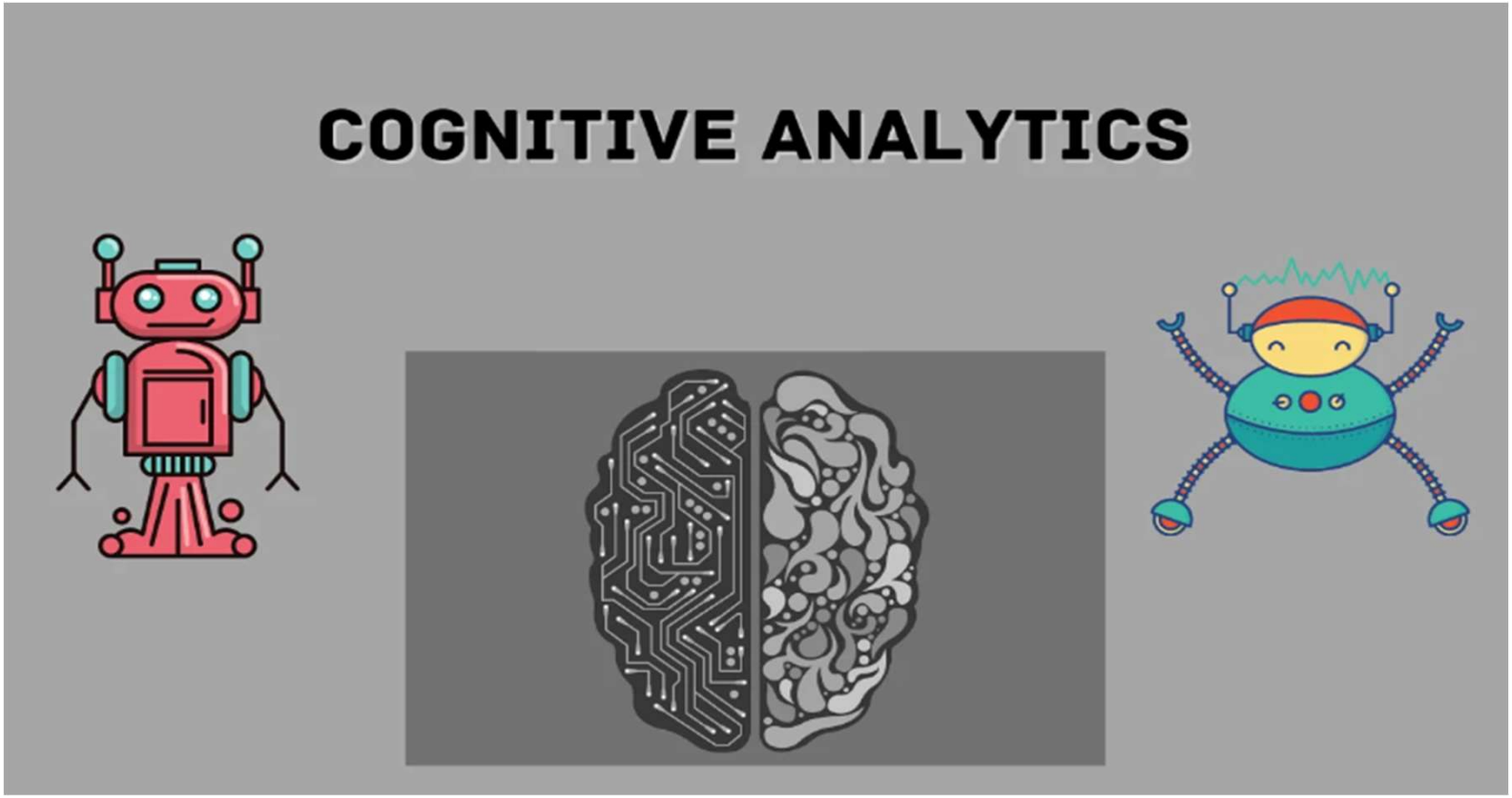
Descriptive analytics used in businesses commonly assesses historical data for finding the trends, customer patterns, areas to do improvement, etc.

Diagnostic analytics can do multiple operations on data like data discovery, data mining, and different type of bivariant data analysis like correlation, etc.

> Predictive analytics comprises a variety of statistical methods from data mining, predictive modeling, machine learning, Deep learning, and past data to make future predictions or unknown events.



Prescriptive modeling is typically not just one individual action in analytics, but it is a combination of other actions.

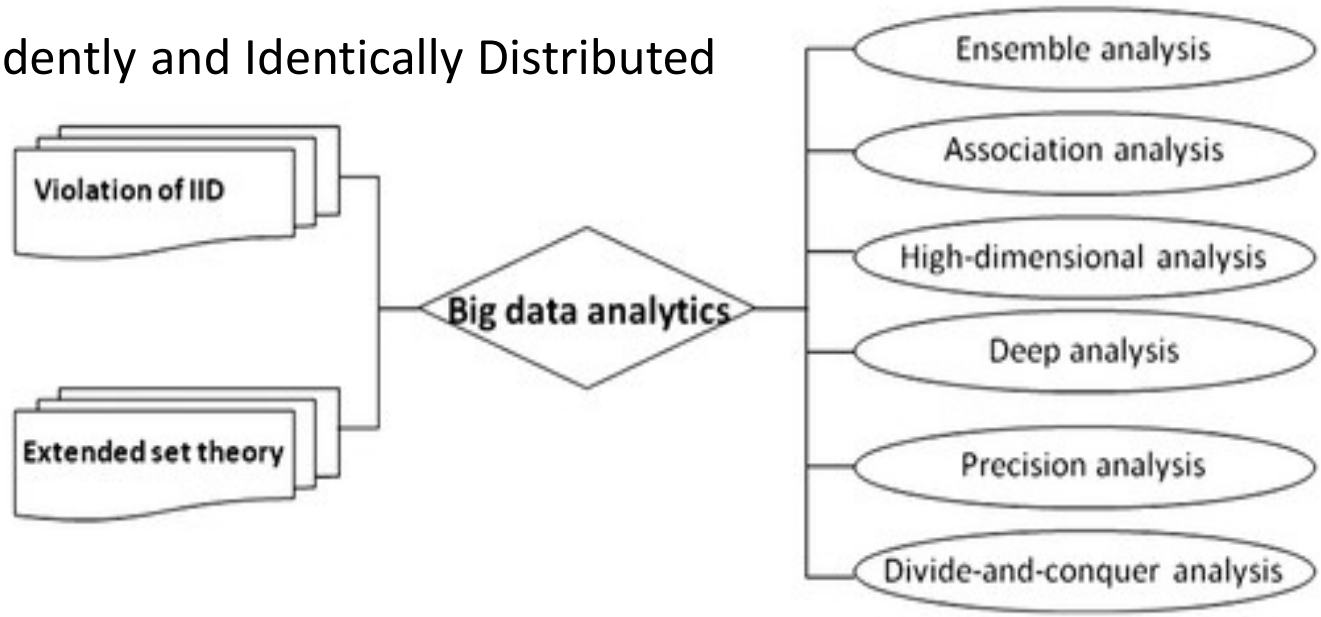


The chief advantage of using (cognitive analytics) intellectual investigation over conventional enormous information examination is that such datasets don't should be pre-labeled.

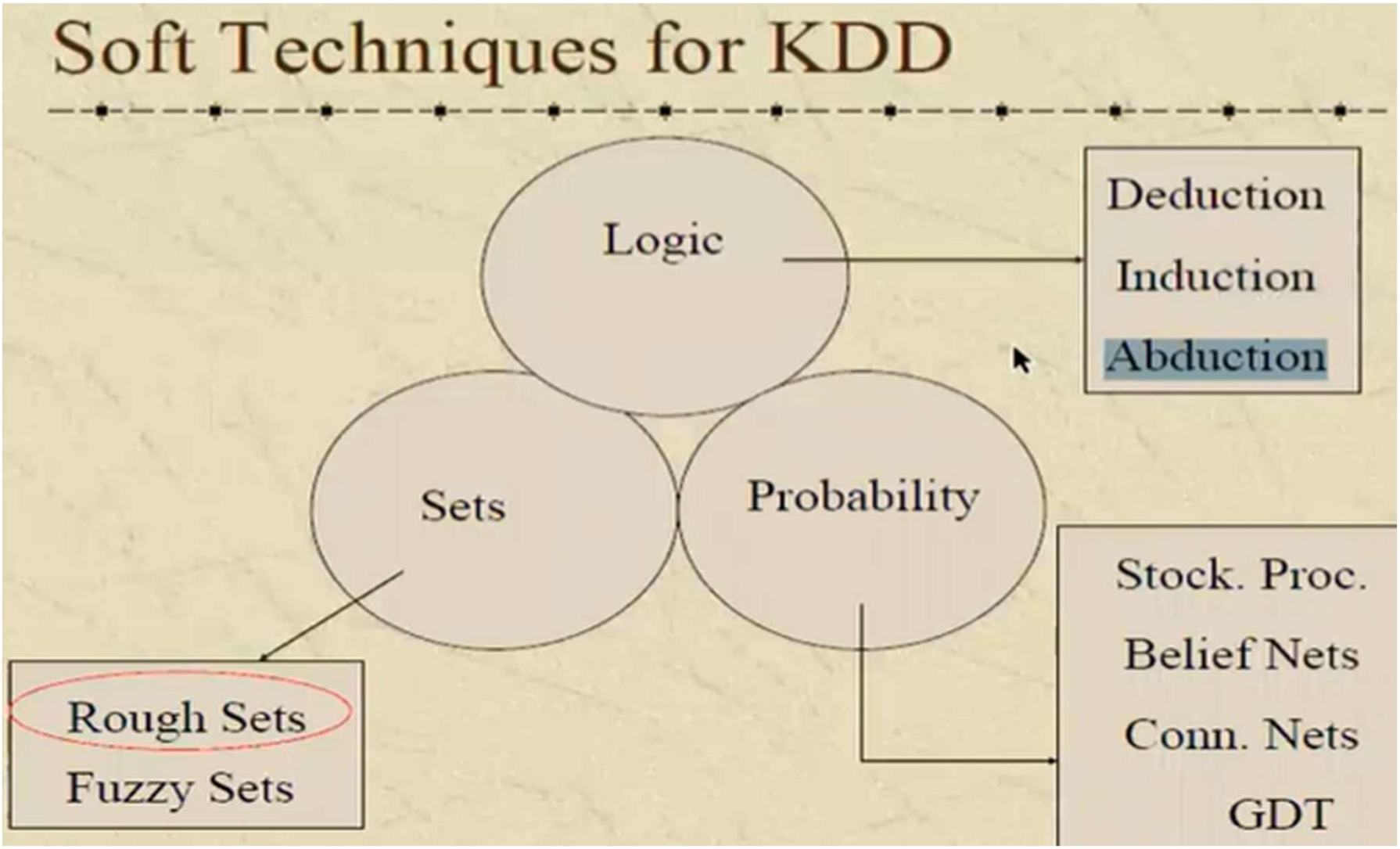
Analytics Types based on two theoretical Ideas

Two theoretical breakthroughs and six techniques in big data analytics.

Independently and Identically Distributed



Exyended Set Theory



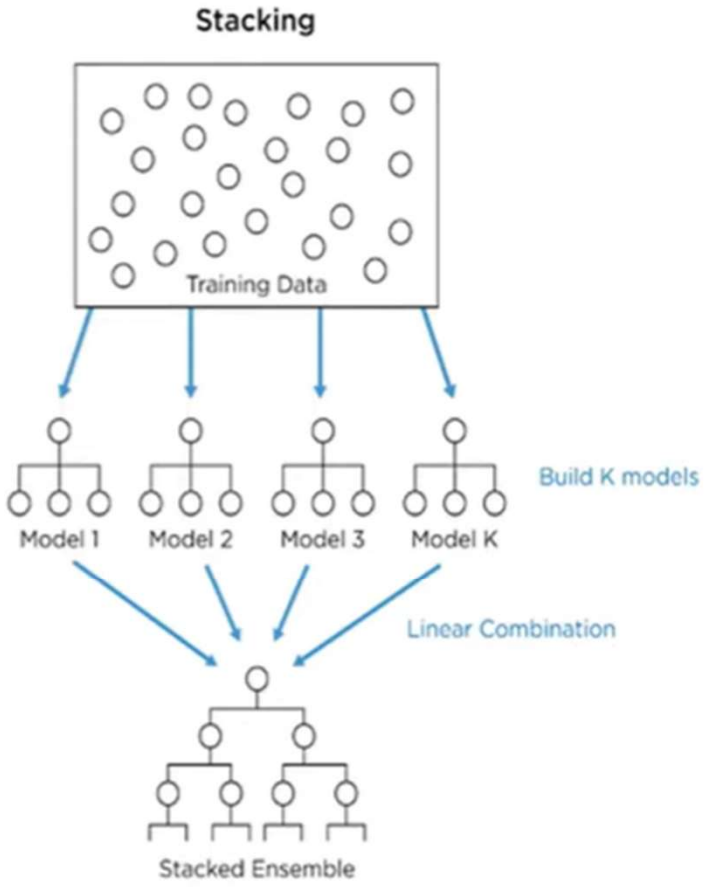
Ensemble-Techniques

Ensemble-Stacking

Advanced Ensemble Techniques

1. **Stacking**
2. **Blending**
3. **Bagging**
4. **Boosting**

- **Stacking:** It combines predictions from multiple (base-level) models to build a new model (meta-model). This meta-model is used for making predictions on the test set.
- Base level algorithms are trained based on a complete training data-set using k-fold cross validation.
- Meta model is trained on the prediction combination of all base level model as feature.
- Stacking is useful when the results of the individual algorithms are very different.



Ensemble-Blending

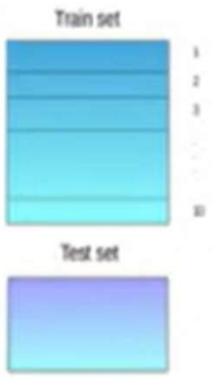
innovate achieve lead

Blending Ensemble Technique

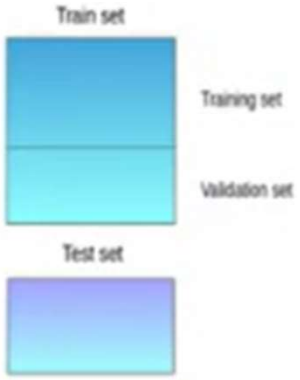
Blending follows the same approach as stacking but uses only a holdout (validation) set from the train set to make predictions.

In other words, unlike stacking, the predictions are made on the holdout set only. The holdout set and the predictions are used to build a model which is run on the test set.

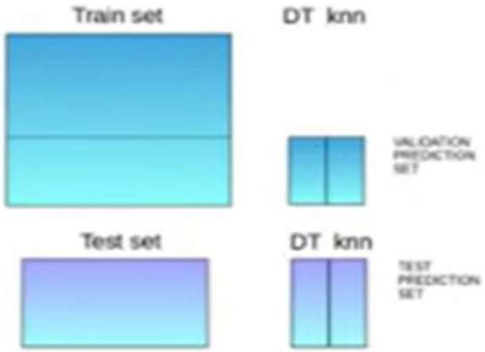
Step 1:
First Data is divided into train and test set.



Step 2:
The train set is split into training and validation sets.



Step 3:
Base models are fitted on the training set and predictions are made on the validation set and test set.



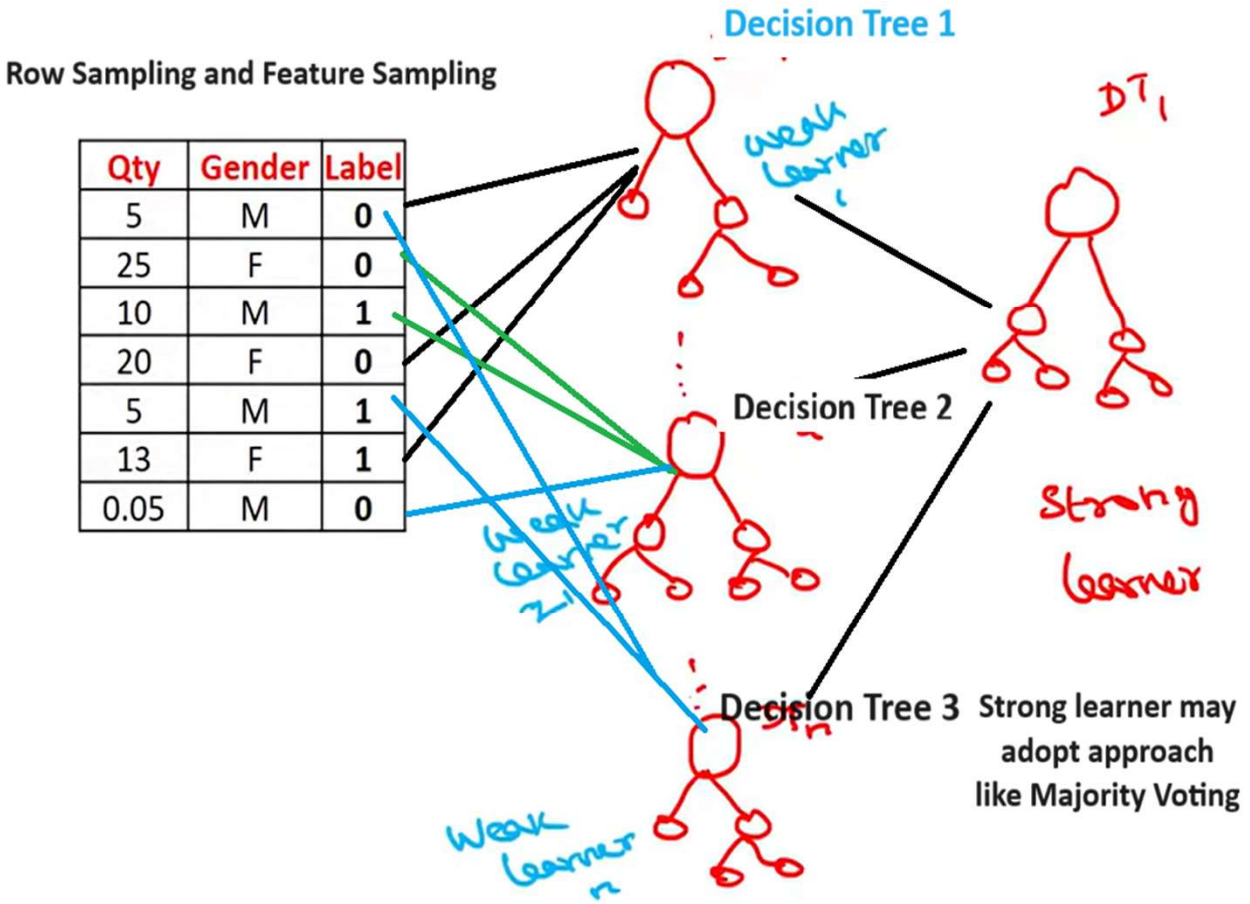
Step 4:
The validation set and its predictions are used as features to build a new model and this model is used to make final predictions on the test features.

Bootstrap approach

Bagging Ensemble Technique

- Bagging is combining the results of multiple models (for example, all decision trees) to get a generalized result.
- If **all the models** utilizes the **same set of data** and combine it, will it be useful?
  - Mostly, All models provides the **same result** due to the **same dataset as input**.

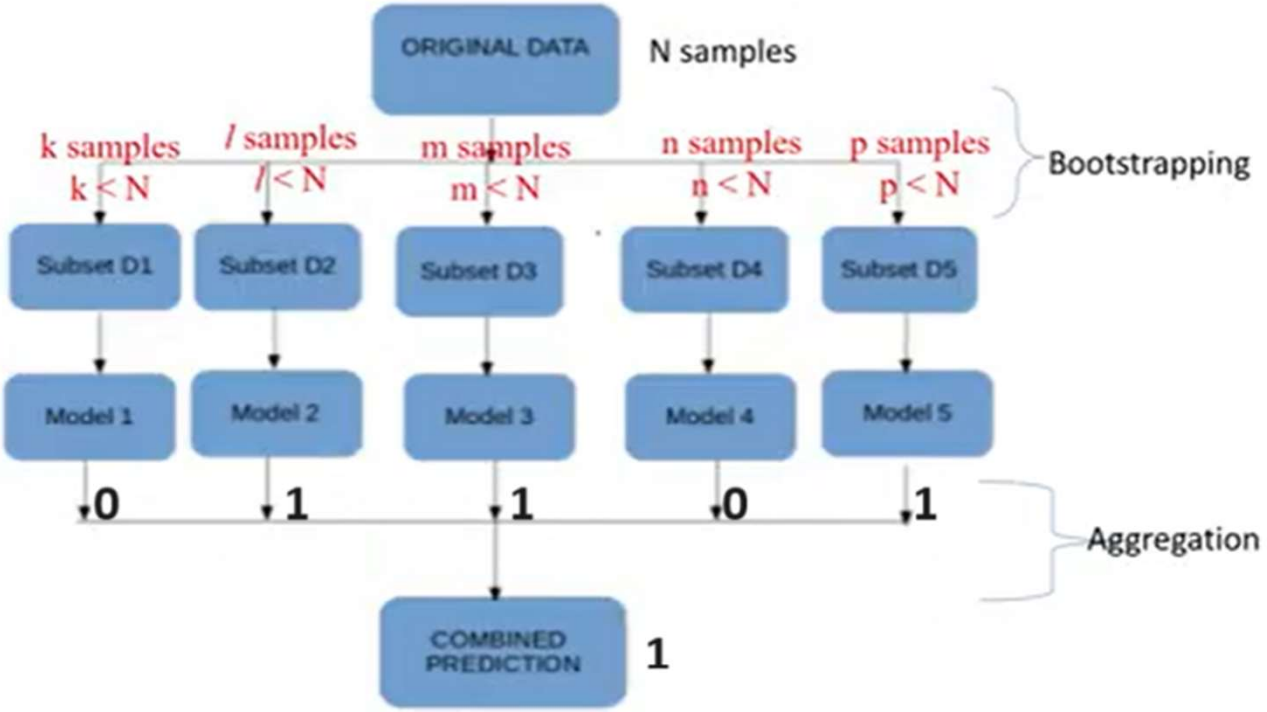
- **Bootstrapping** solves this problem.
- In Bootstrapping, **divide dataset** into **few subsets** with data examples replacement (row sampling).
- **Bagging /Bootstrap Aggregating uses** these subsets (bags) to get better distribution (complete set).
- The size of subsets less than the original dataset.



Contd… approach

Bagging Ensemble Technique

- A base model (weak model) is created on each of these subsets (subsets are selected using sampling process).
- The models run in parallel, and independent of each other.
- The final predictions are determined by combining the predictions from all the models.



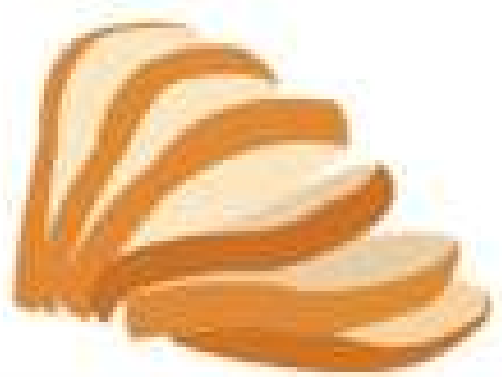
Boosting Ensemble

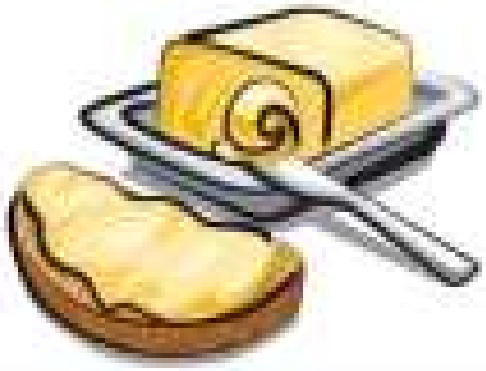
innovate achieve lead

Boosting Ensemble Technique

- If a data point is incorrectly predicted by the first model, and then the next model (probably all models), will combining the predictions provide better results?
- It is solved by **boosting.**
- **Boosting** is a sequential process, where each subsequent model attempts to correct the errors of the previous model.
- The succeeding models are dependent on the previous model.

Association Type
Market Basket Analysis





Bread and Butter

High-Dimensional Type

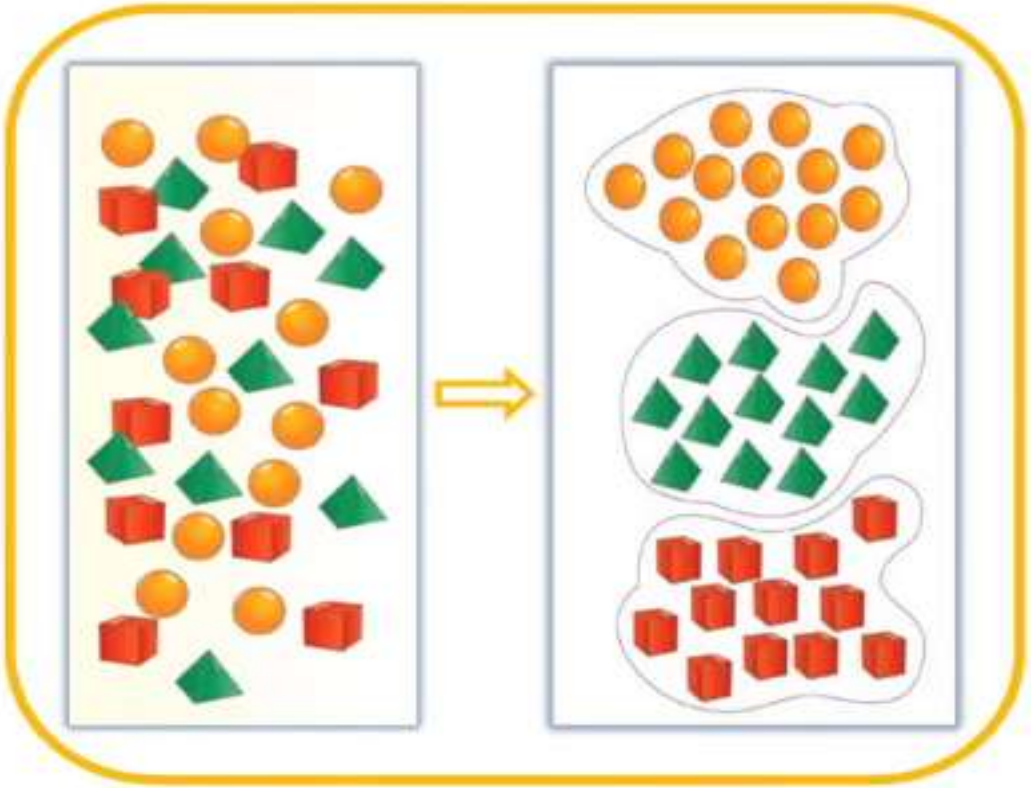
Traditional Clustering

What is clustering?

- Unsupervised learning method to classify unlabeled data into different groups

Traditional clustering methods

- *K*-means clustering
- Hierarchical clustering
- Density-based clustering

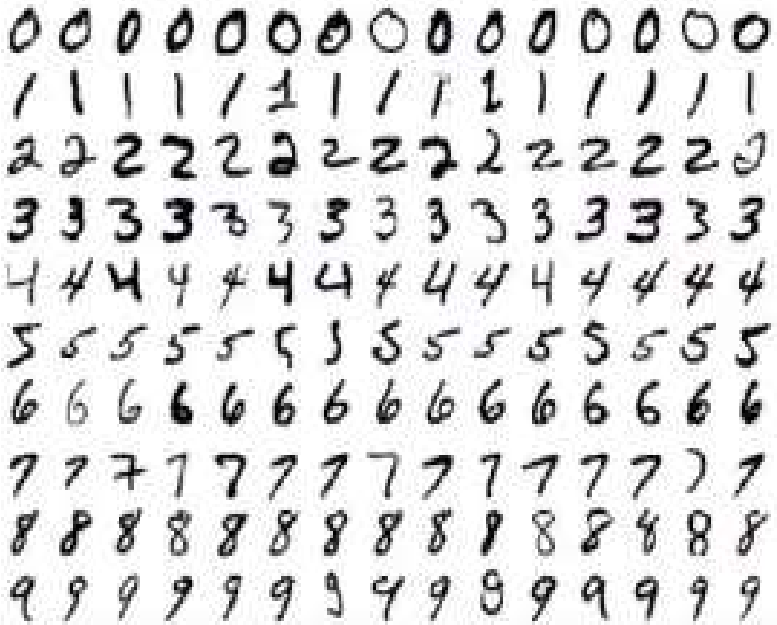


K-means approach

Points which are close to each other within the threshold distance are grouped together to form a cluster

In high dimensional cluster all points will become sparse and thus it becomes impossible to compute their closeness
Example: stars in galaxy

Too Heavy



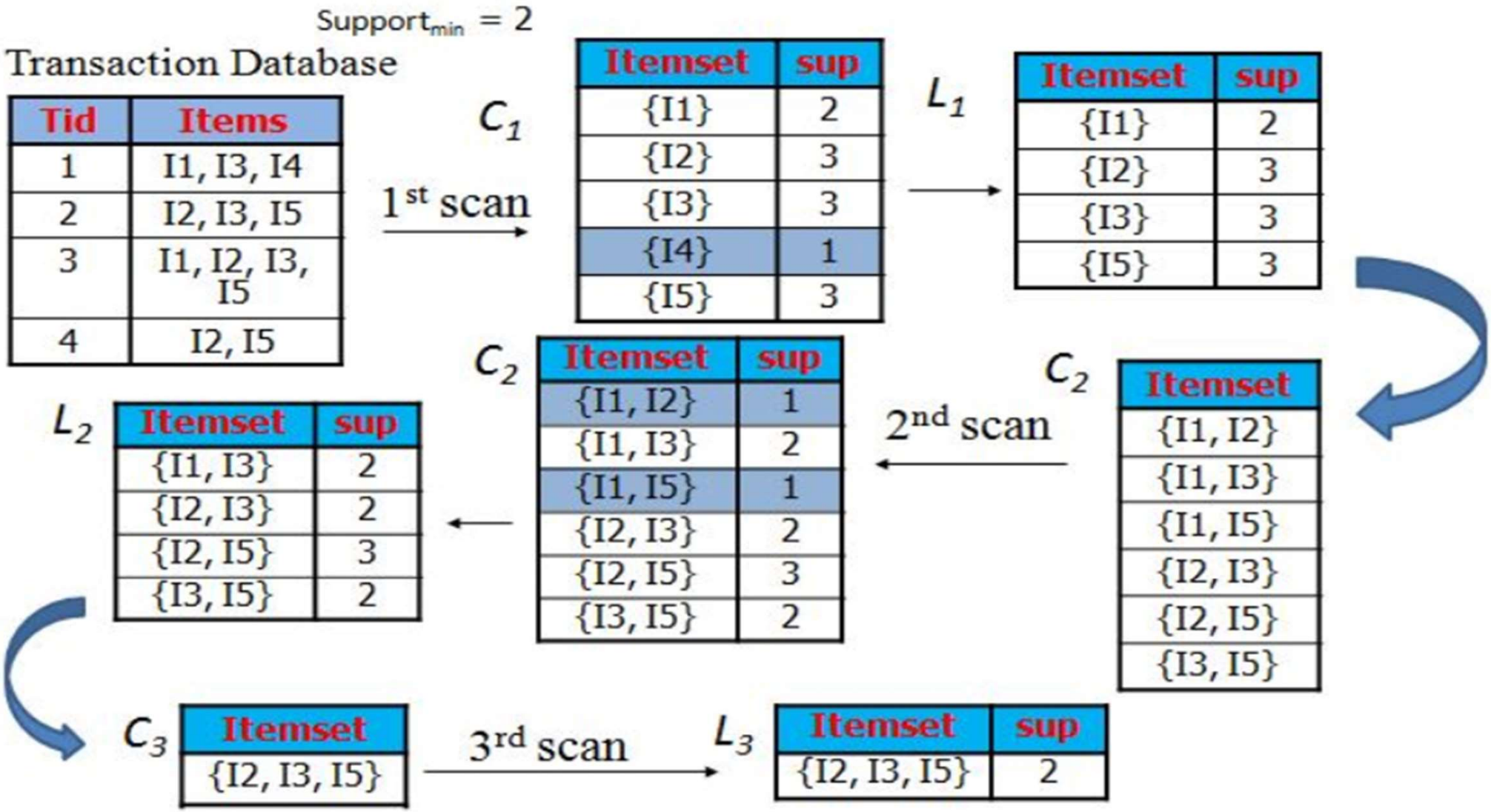
MNIST

28 * 28 = 784 features

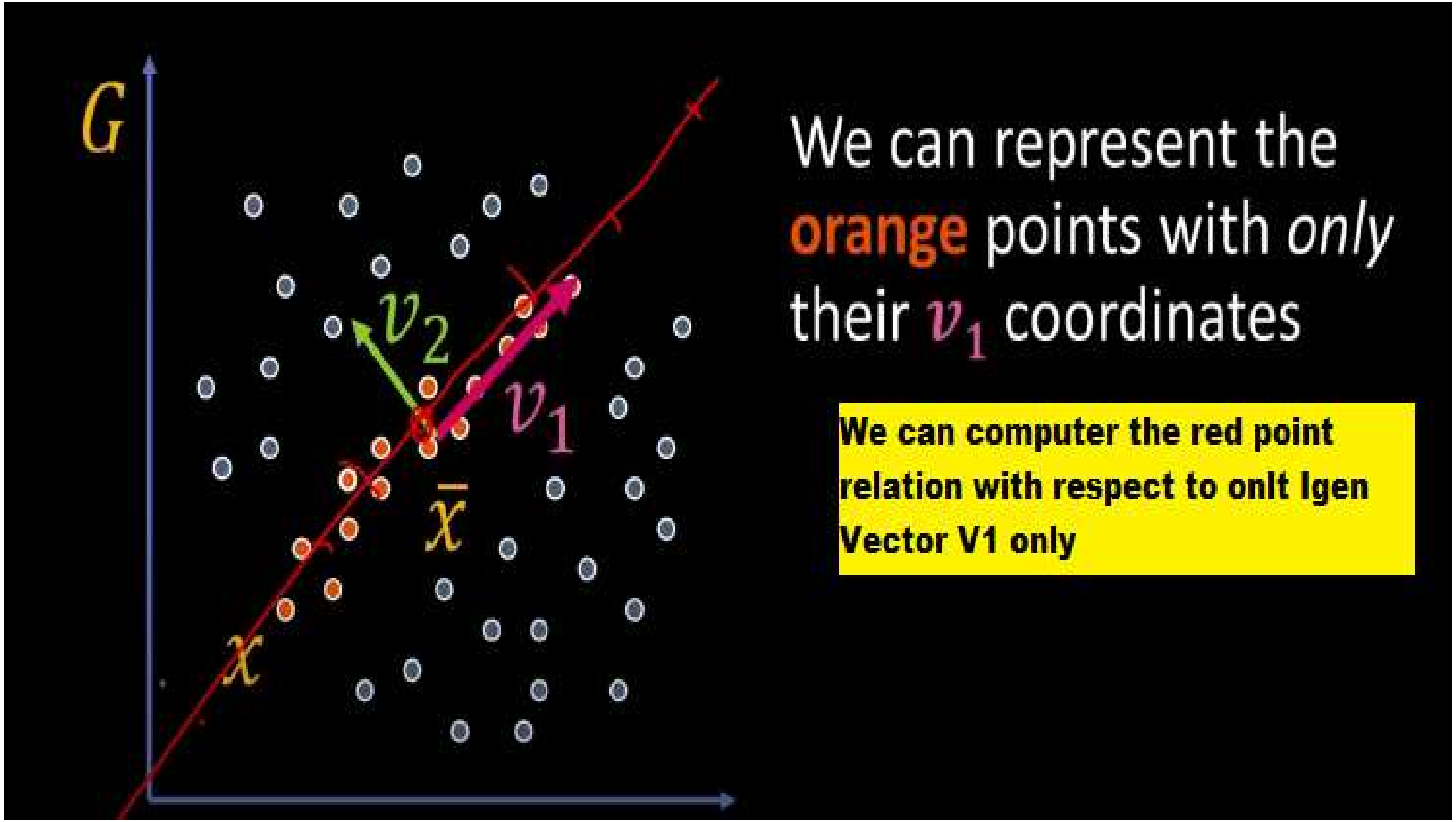
CIFAR-10

32 * 32 * 3 = 3072 features

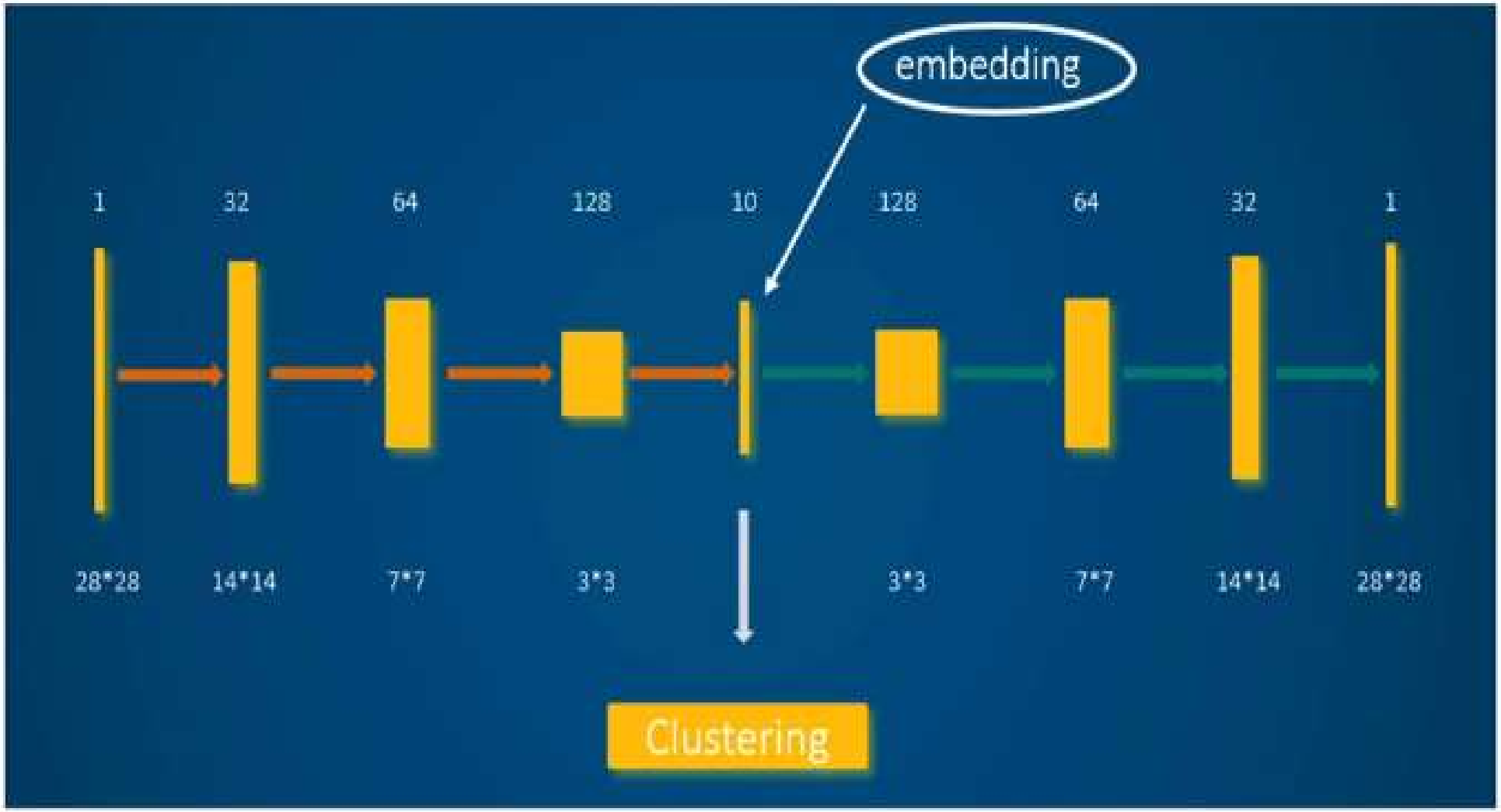
The Apriori Algorithm—An Example



Solution : Dimension reduction



Solution : Dimension reduction



Deep Analysis Type

Precision Analysis Type

Precision analysis is used to evaluate the veracity of data from the perspective of data utility and data quality.

The fundamental difference between precision and traditional medicine is their treatment approaches. Traditional medicine often relies on generalized protocols based on the average responses of large populations. This can lead to effective treatments for some patients but less so for others. In contrast, precision medicine utilizes big data analytics to identify patterns and correlations within diverse patient datasets, enabling custom data visualization to illustrate specific biomarkers and genetic variations that influence disease progression and treatment response. As a result, precision medicine can offer therapies precisely calibrated to the patient's genetic profile and health conditions, significantly improving the chances of successful outcomes.

Divide & Conquer Analysis Type

Divide and Conquer Approach:

Divide:

Task Division: Split the large dataset into smaller chunks or partitions. Each chunk can be processed independently. For instance, if you have a dataset containing billions of lines of text, you might divide it into thousands of smaller files or segments.

Conquer:

Local Processing: Process each chunk independently to compute the word counts. This involves reading the text, tokenizing it into words, and counting occurrences for each word in the chunk. This step is usually performed in parallel across different machines or nodes in a distributed computing environment.

Combine:

Aggregation: After processing all chunks, combine the word counts from each chunk to get the final word count for the entire dataset. This involves merging the results from all chunks and summing up the word counts for each unique word.

Analytics Contd…

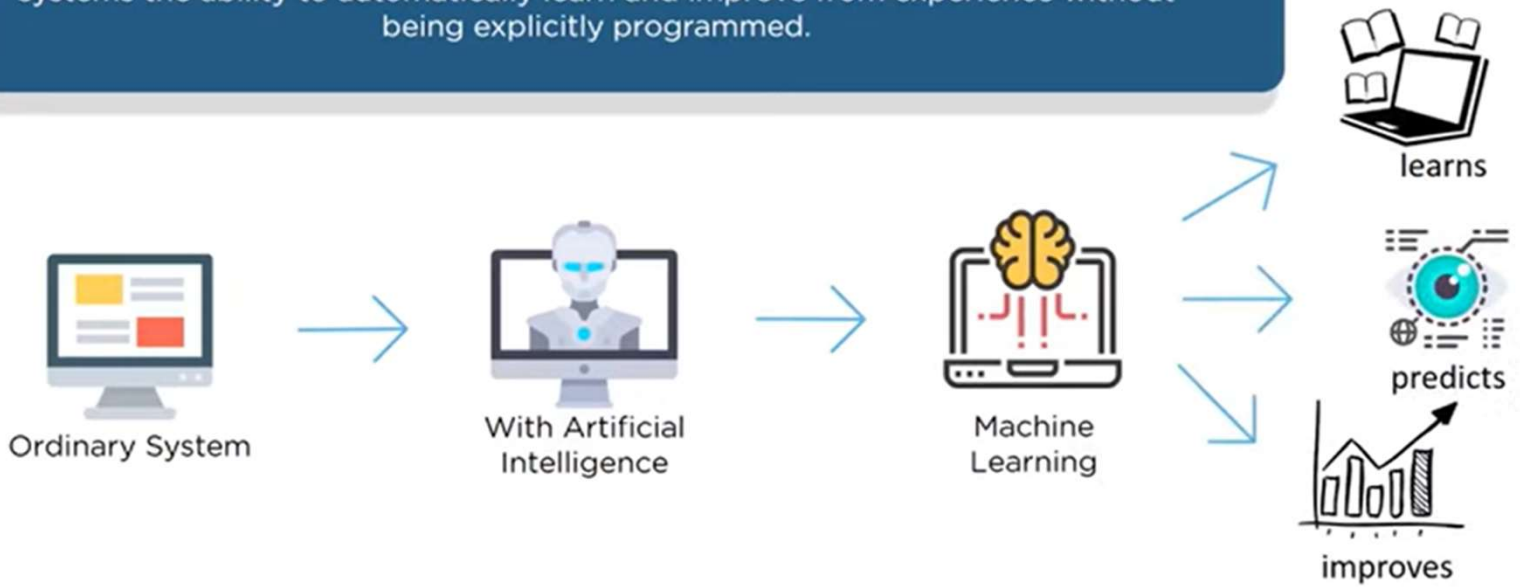
What is Data Analytics?

Data analytics is the process of deriving useful insights from data, It is a subset of data science, although the terms are often used interchangeably.

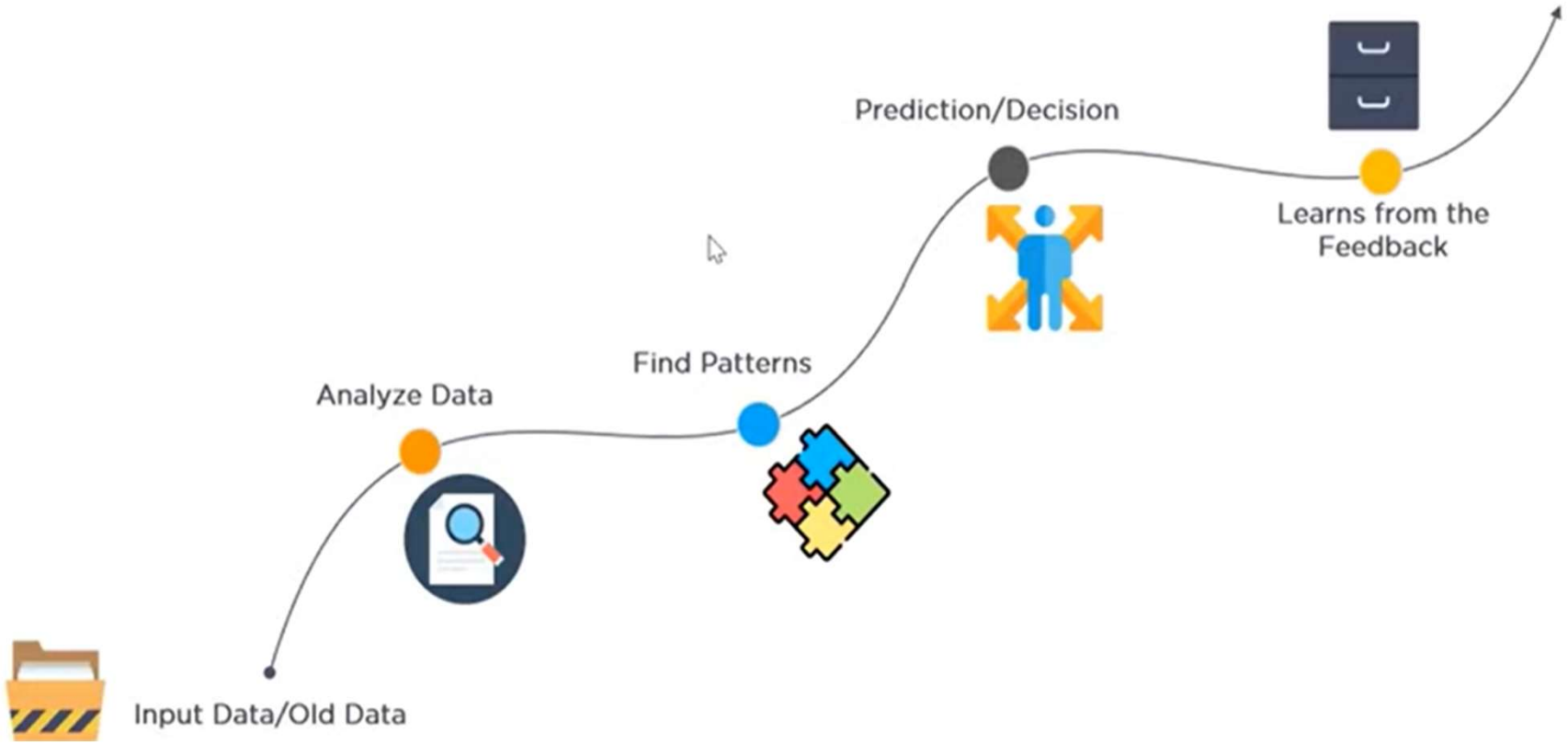
> Data analytics itself breaks down into several sub-areas:
>
> 1. Statistical analysis,
>
> 2. machine learning,
>
> 3. Business intelligence (BI).

What is Machine Learning?

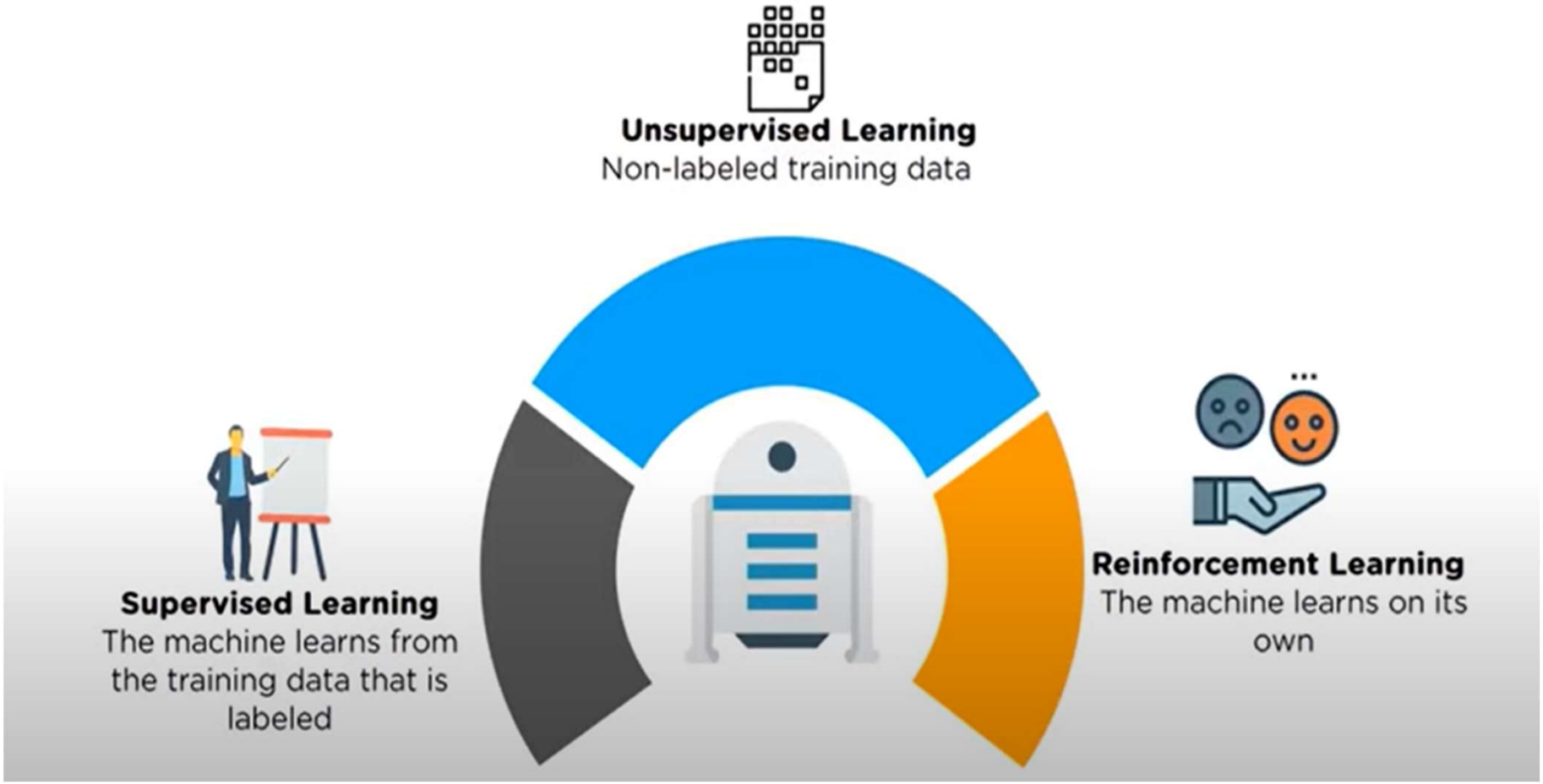
Machine Learning is an application of Artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.



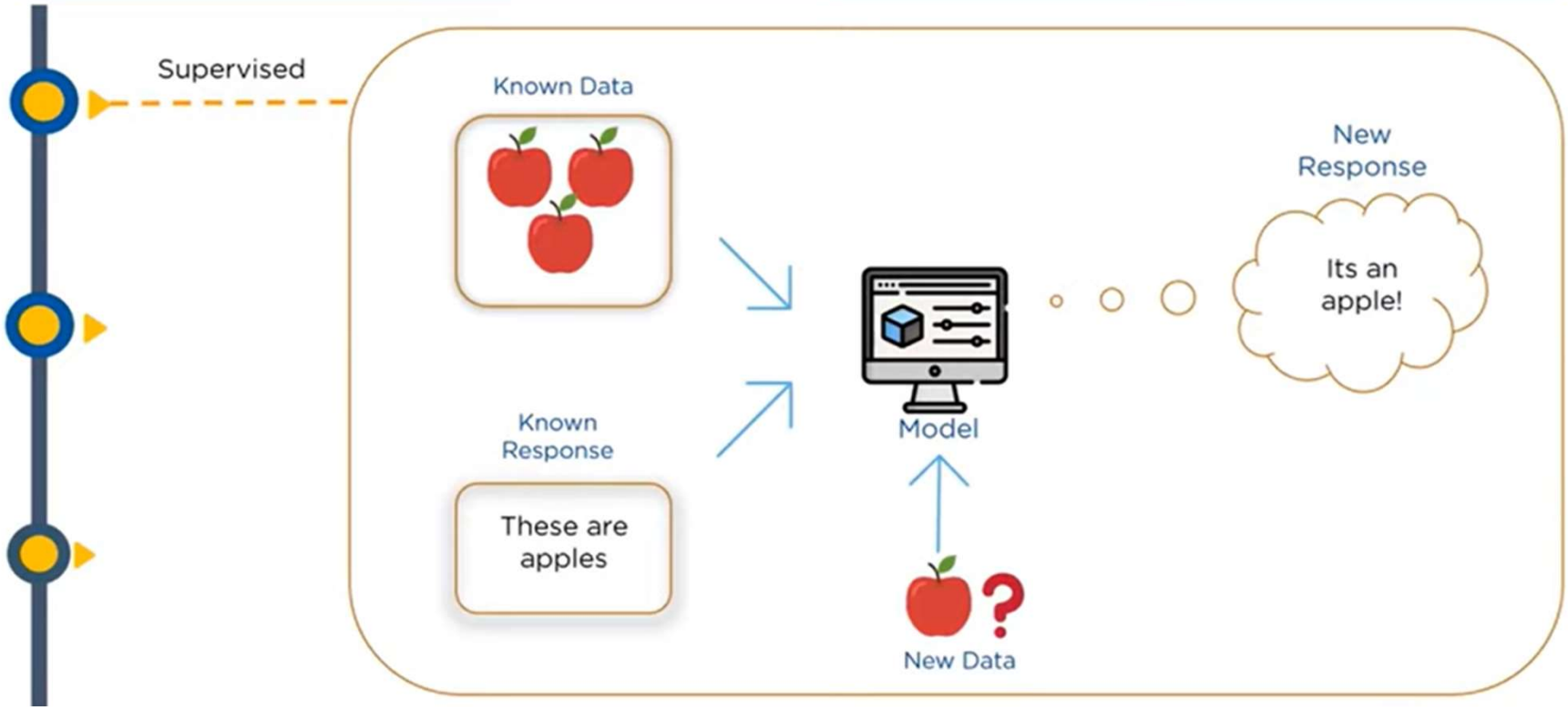
Machine Learning process



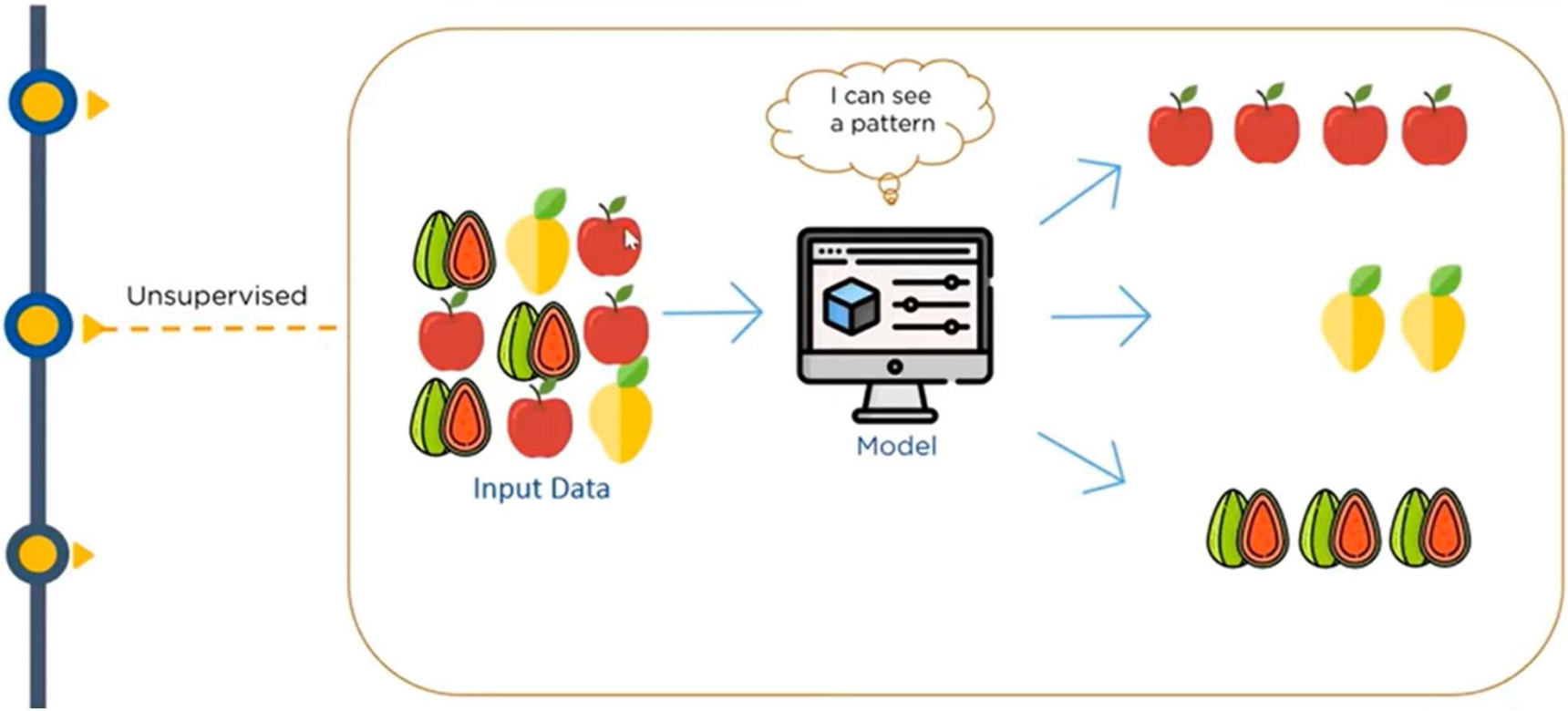
Types of Machine Learning



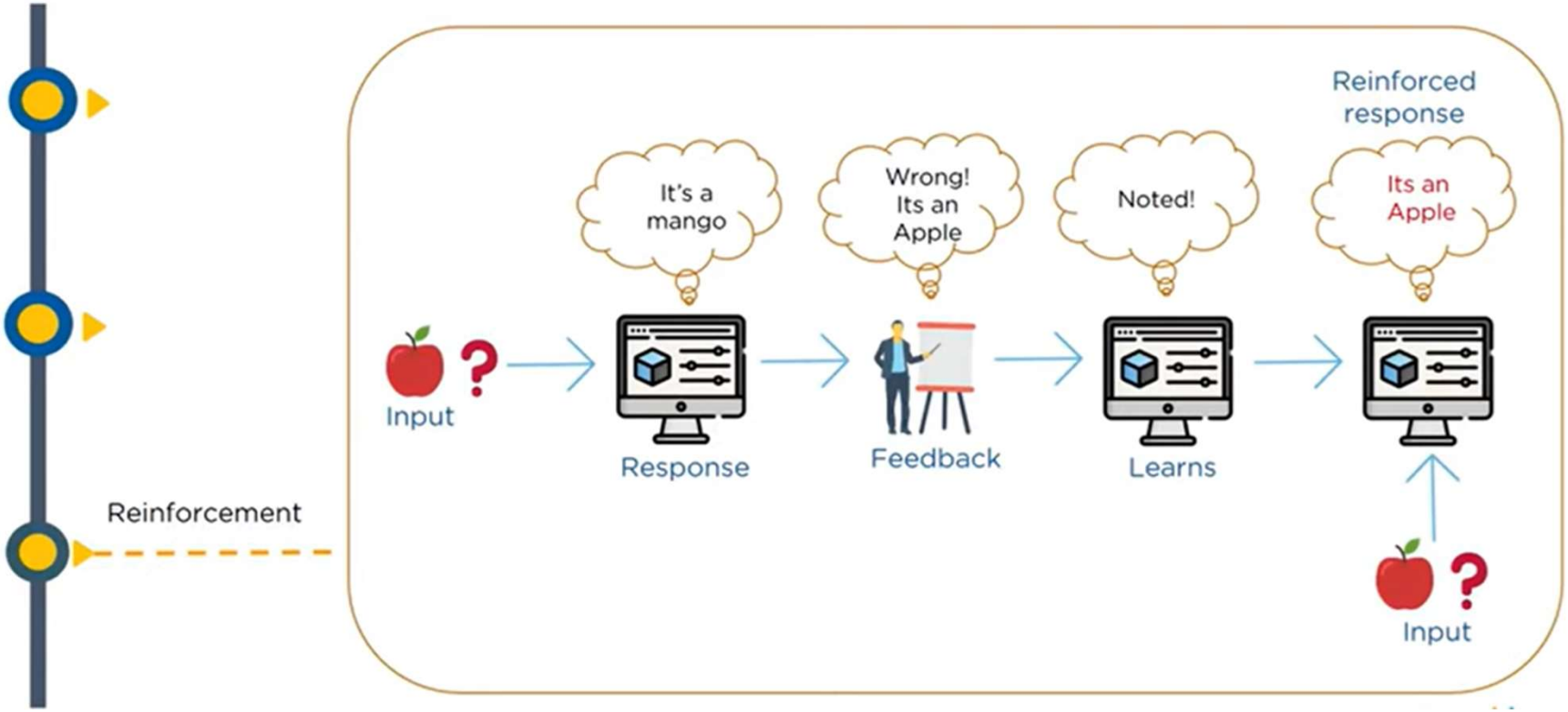
Types of Machine Learning



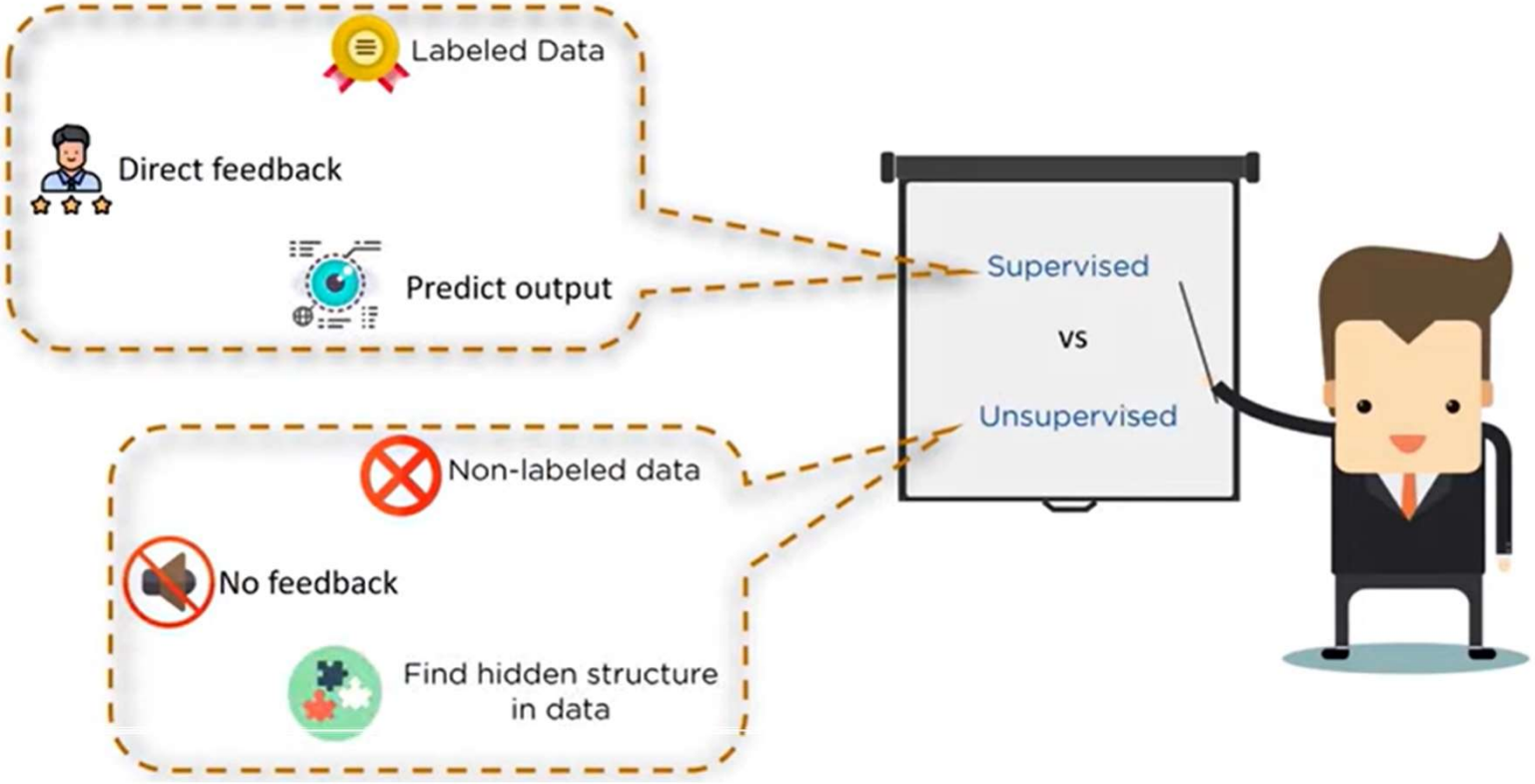
Types of Machine Learning



Types of Machine Learning



Supervised vs Unsupervised



The right Machine Learning solution?

Classification

Used when the output is categorical like 'YES' or 'NO'

Algorithms used

- Decision Tree
- Naïve Bayes
- Random Forest
- Logistic regression
- KNN

Clustering

Used when the data needs to be organized to find patterns in the case of 'product recommendation'

Algorithms used

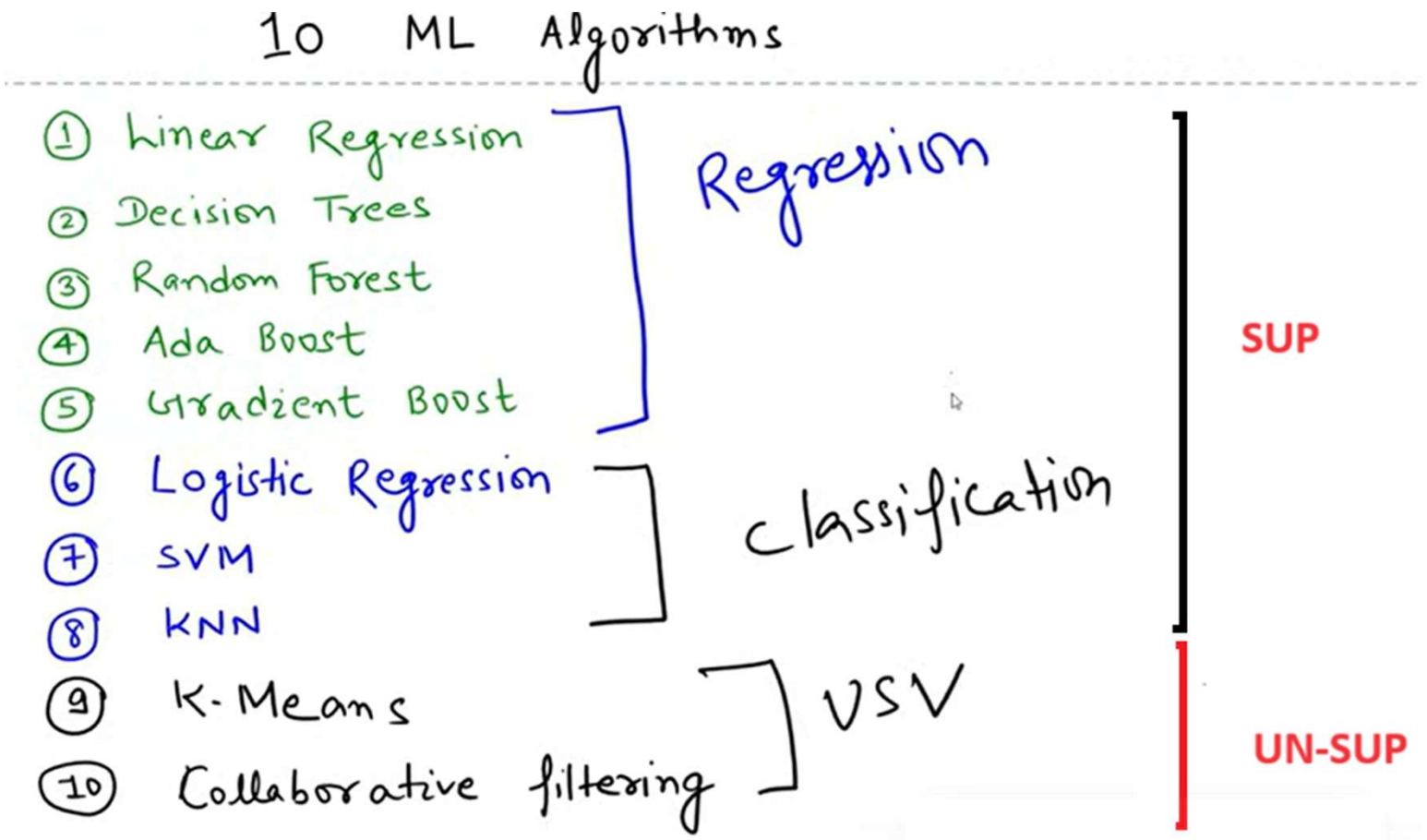
- K Means

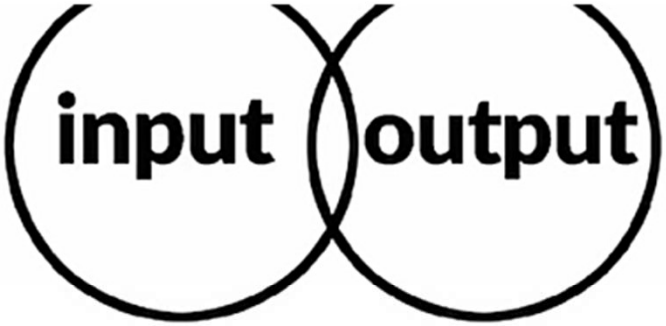
Regression

Used when a value needs to be predicted like the 'stock prices'

Algorithms used

- Linear Regression


10 ML Algorithms
1 Linear Regression
2 Decision Trees
3 Random Forest
4 Ada Boost
5 Gradient Boost
6 Logistic Regression
7 SVM
8 KNN
9 K-Means
10 Collaborative filtering
Regression
classification
VSV
SUP
UN-SUP



Regression

Classification

input -> output

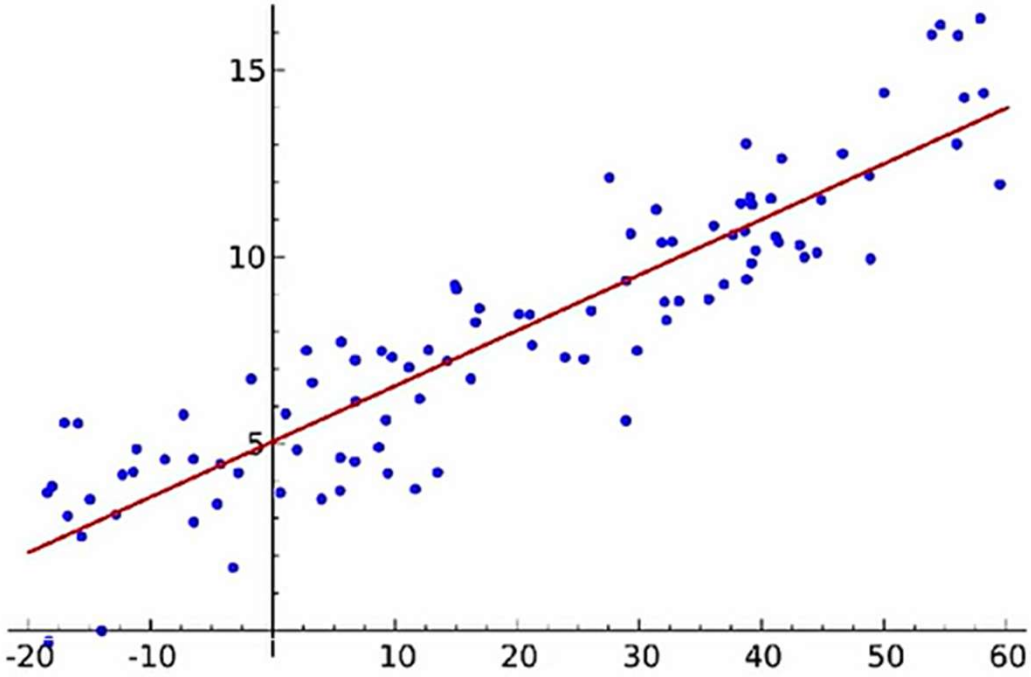
Age (input)	Shoe size (inch) (output)
5	7
8	8.25
11	9
17	9.75

→ Supervised Learning Model →

Eg, Predict shoe size of a 10 year old?

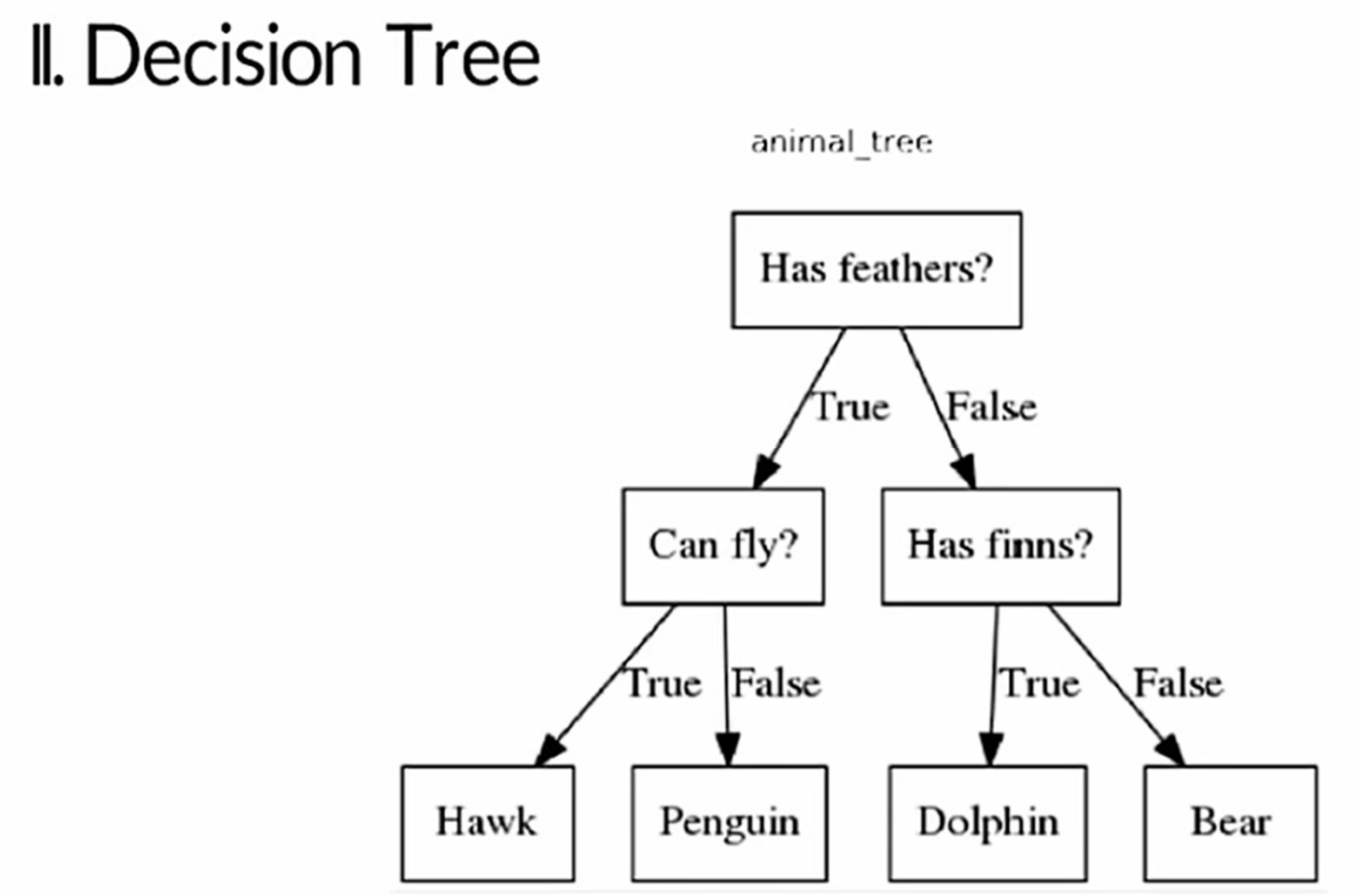
Types of Regression Models

I. Linear Regression



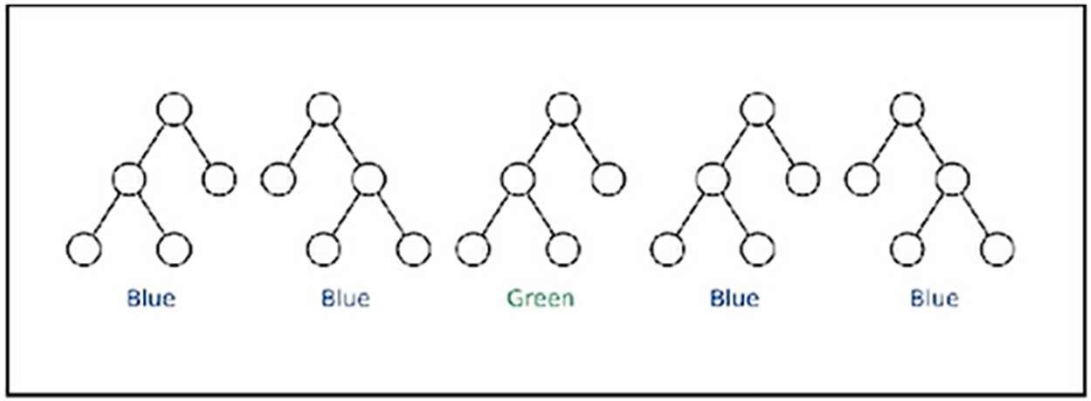
- Multiple Linear
- Polynomial

II. Decision Tree



III. Random Forests

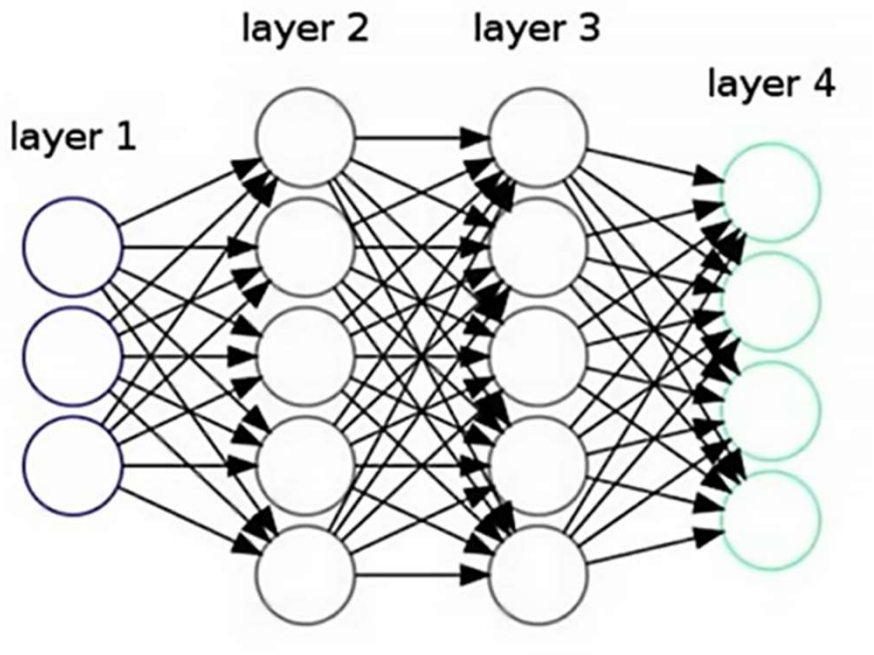
Ensemble learning technique



Blue

"Majority Wins" Model

IV. Neural Network



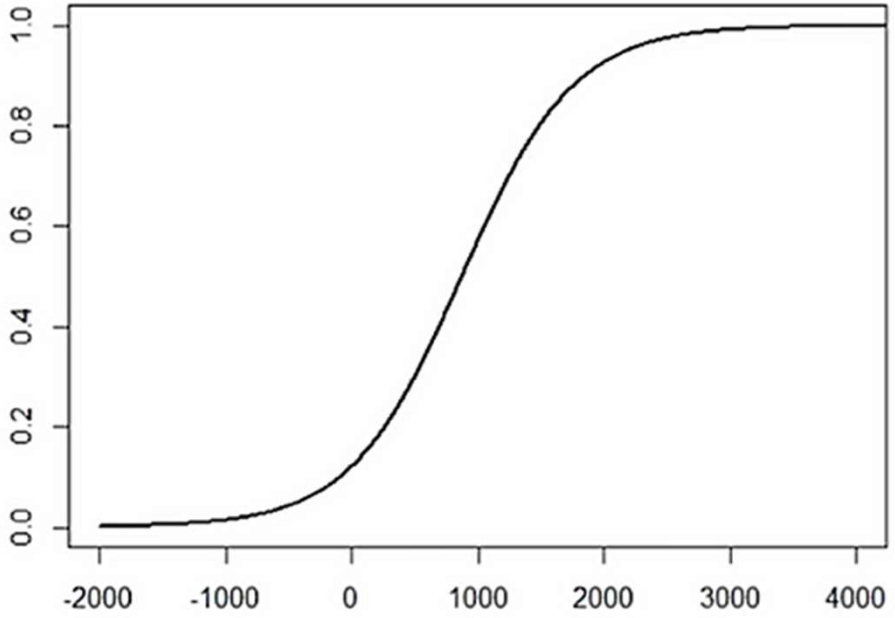
Layer 1 = Input Layer
Layer 2 = Hidden Layer
Layer 3 = Hidden Layer
Layer 4 = Output Layer

1.2 Classification

discrete

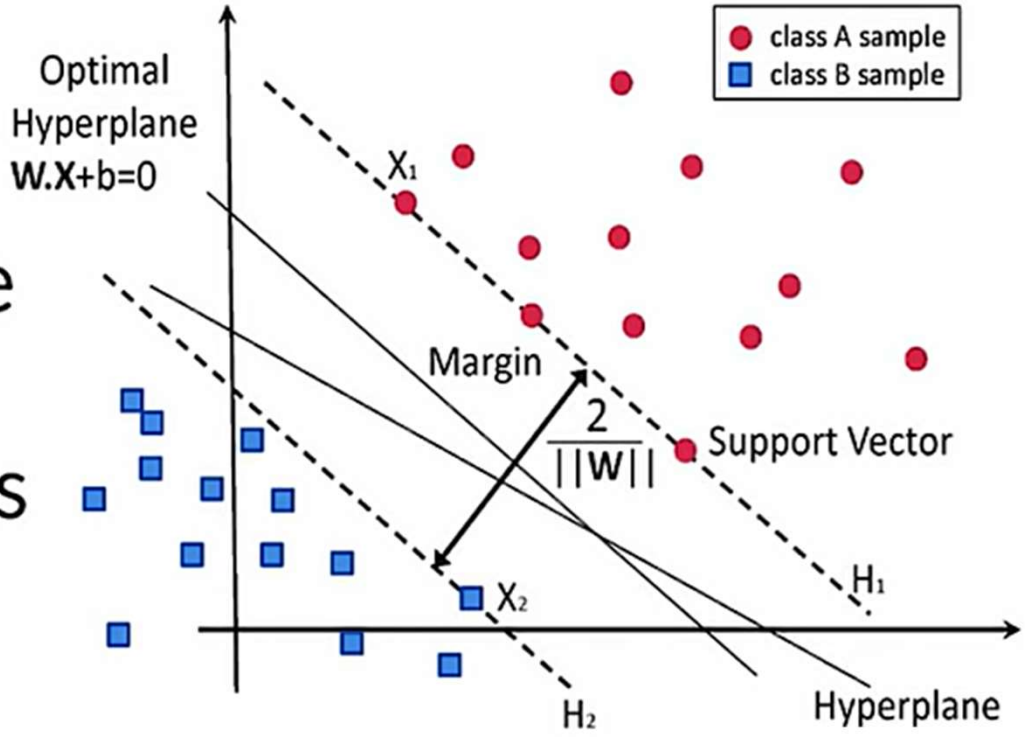
I. Logistic regression

the output values can only be between 0 and 1

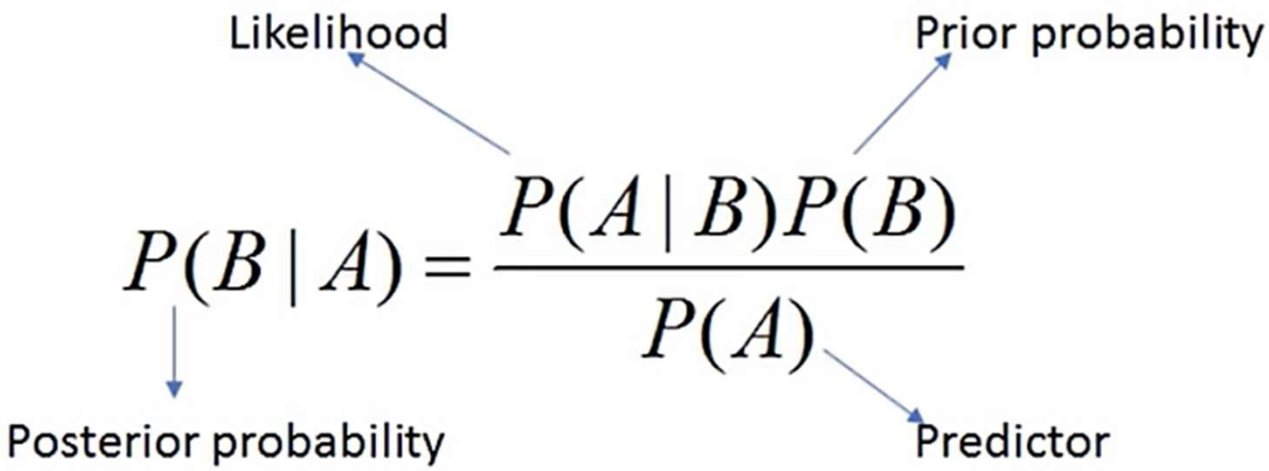


II. Support Vector Machine

N-dimensional space that distinctly classifies data points



III. Naive Bayes



IV. Decision Tree, Random Forest, Neural Network

These algoriyhms are same as Navie Bayes...Only difference is ....Output is Discrete rather than Continuos

Example

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$$P(PlayTennis = yes) = 9/14 = .64$$

$$P(PlayTennis = no) = 5/14 = .36$$

Outlook	Y	N
sunny	2/9	3/5
overcast	4/9	0
rain	3/9	2/5
Tempreature		
hot	2/9	2/5
mild	4/9	2/5
cool	3/9	1/5

Humidity	Y	N
high	3/9	4/5
normal	6/9	1/5
Windy		
Strong	3/9	3/5
Weak	6/9	2/5

Example

$$\langle Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong \rangle$$

$$v_{NB} = \underset{v_j \in \{yes,no\}}{\text{argmax}} P(v_j) \prod_i P(a_i|v_j)$$

$$= \underset{v_j \in \{yes,no\}}{\text{argmax}} P(v_j) \quad P(Outlook = sunny|v_j) P(Temperature = cool|v_j)$$

$$\cdot P(Humidity = high|v_j) P(Wind = strong|v_j)$$

$$v_{NB}(yes) = P(yes)\ P(sunny|yes)\ P(cool|yes)\ P(high|yes)\ P(strong|yes) = .0053$$

$$v_{NB}(no) = P(no)\ P(sunny|no)\ P(cool|no)\ P(high|no)\ P(strong|no) = .0206$$

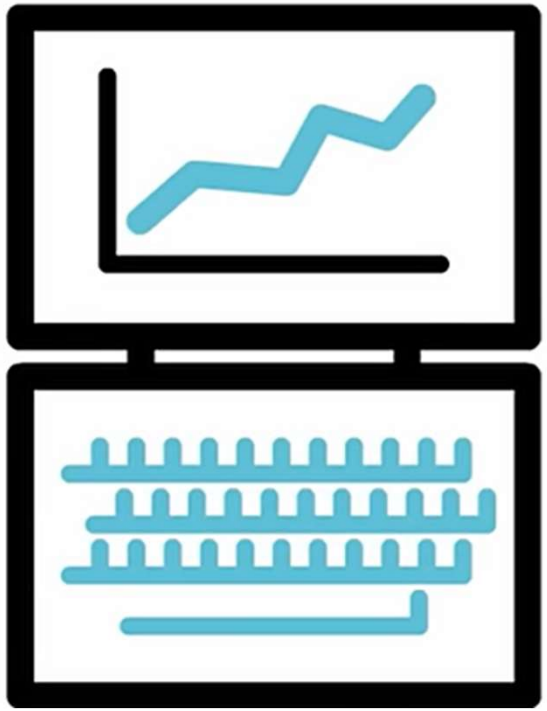
$$v_{NB}(yes) = \frac{v_{NB}(yes)}{v_{NB}(yes)+v_{NB}(no)} = 0.205$$

$$v_{NB}(no) = \frac{v_{NB}(no)}{v_{NB}(yes)+v_{NB}(no)} = 0.795$$

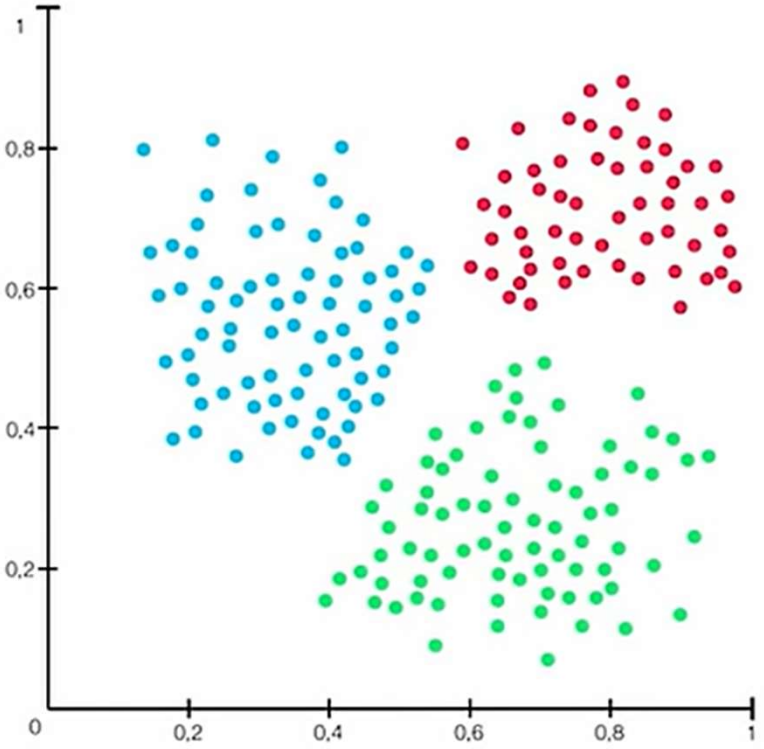
2. Unsupervised Learning

Patterns from input data without references to labeled outcomes

> Clustering
> Dimensionality Reduction



2.1 Clustering



> K-means
> Hierarchical
> Mean shift
> Density-based

2.2 Dimensionality Reduction

process of reducing the dimension
of your feature set

> feature elimination
> feature extraction

PRINCIPAL COMPONENT ANALYSIS (PCA)