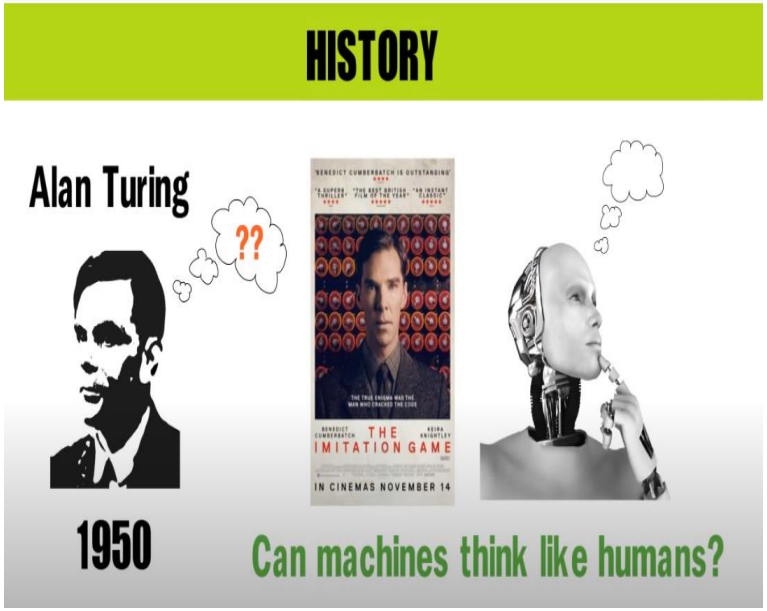
UNIT I: Introduction to Machine Learning,
History and Overview of machine learning,

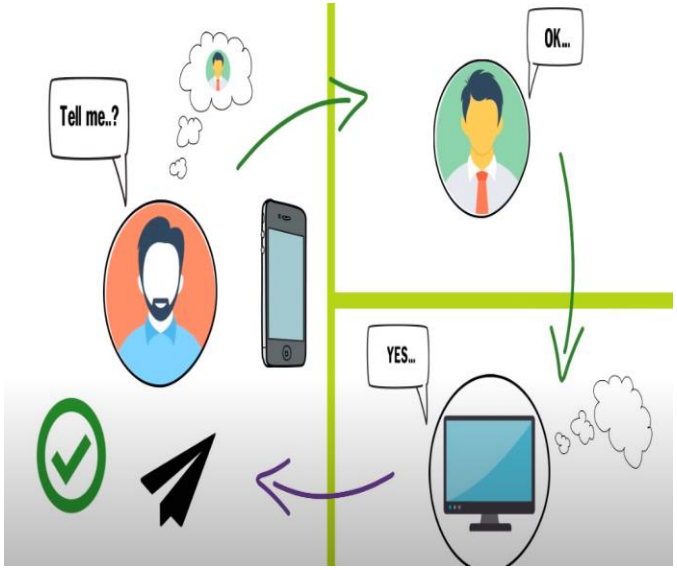
Applications,
Types of Machine Learning,
Basic Concepts. Concept Learning and candidate elimination learning Algorithm.

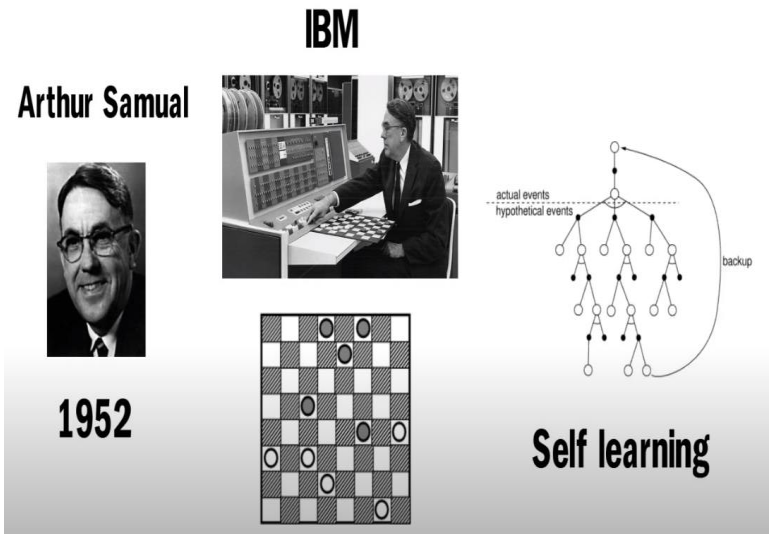
Machine learning (ML) is the study of computer algorithms that improve automatically through experience and by the use of data.

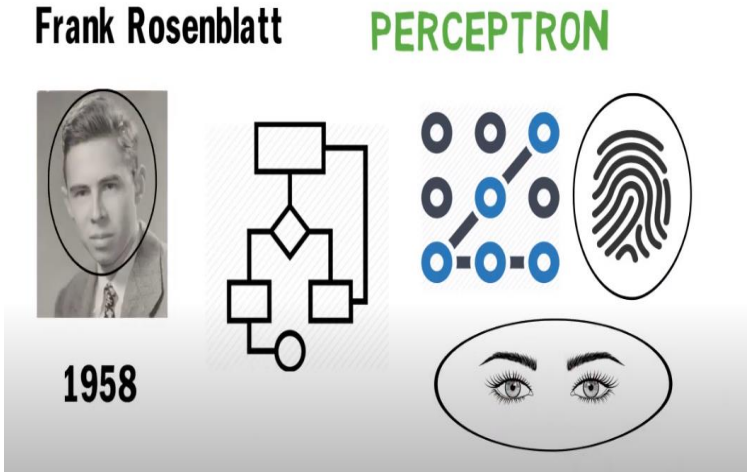
Its self learning process without explicitly involved any users. It learn from past data or experienced.

History


HISTORY
Alan Turing
??
THE IMITATION GAME
IN CINEMAS NOVEMBER 14
1950
Can machines think like humans?


Tell me..?
OK...
YES...


IBM
Arthur Samual
1952
actual events
hypothetical events
backup
Self learning


Frank Rosenblatt
PERCEPTRON
1958

History

1979

Stanford Cart

Stanford University

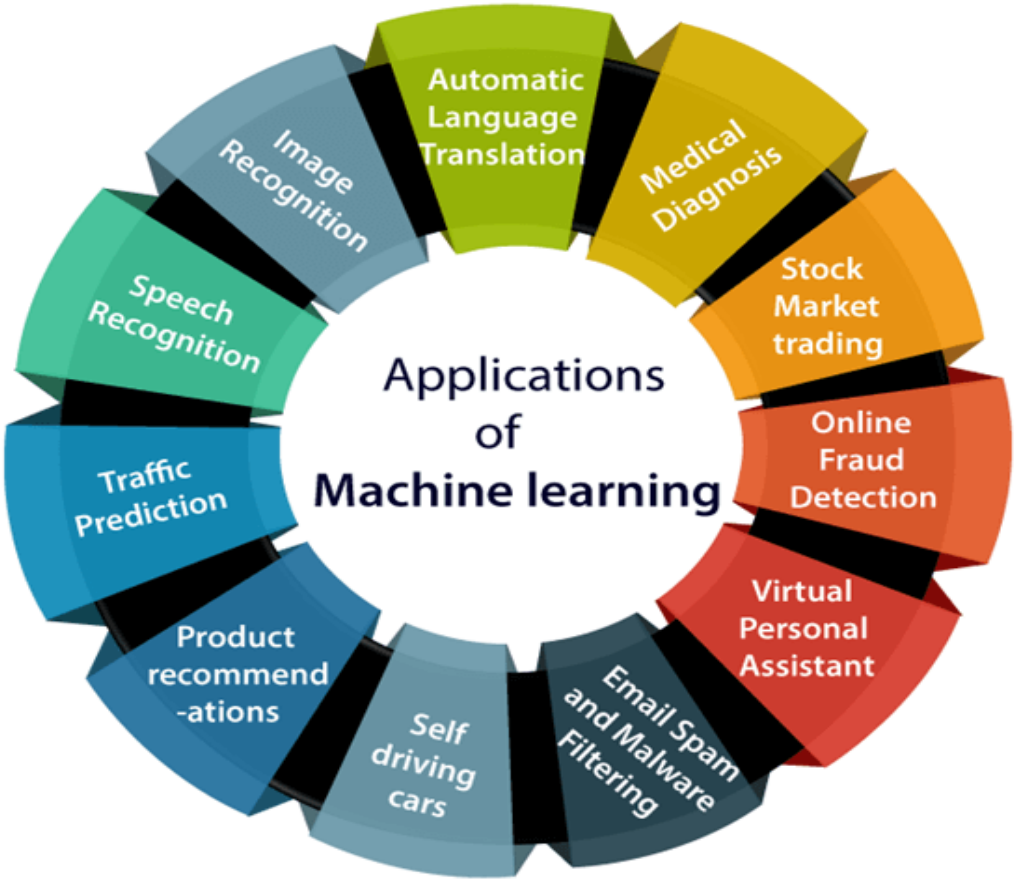
NETTALK

1985

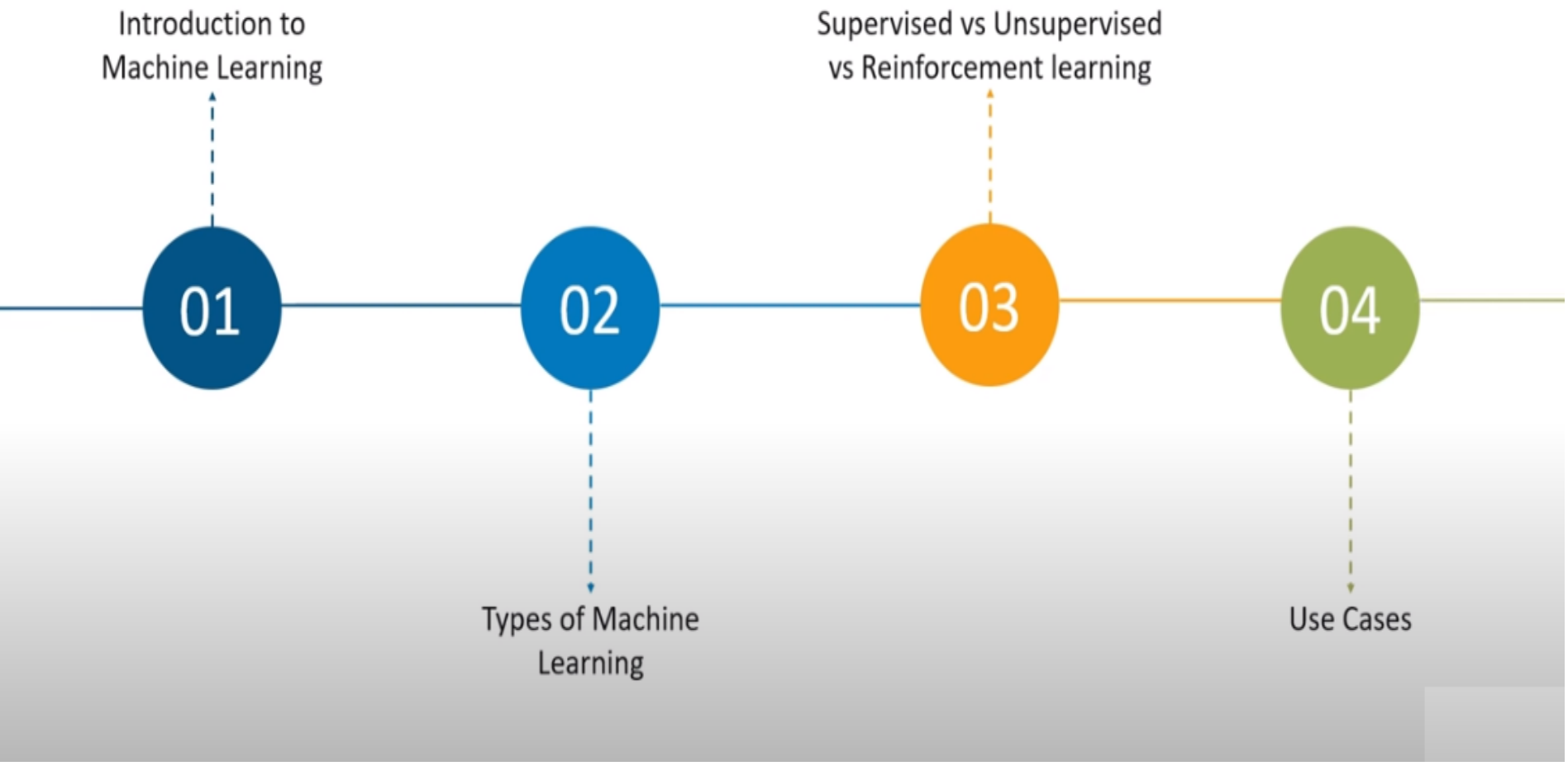
Terry Sejnowski

word pronounce learning

Application of M/C Learning

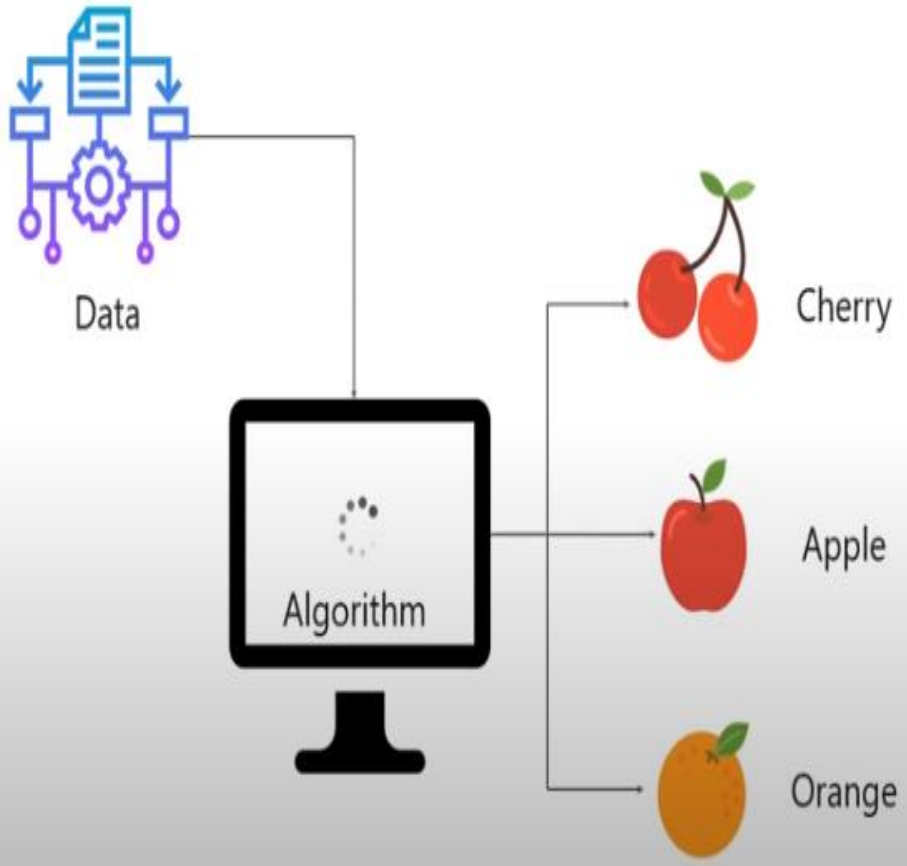


Agenda


Introduction to Machine Learning
01
02
Types of Machine Learning
Supervised vs Unsupervised vs Reinforcement learning
03
04
Use Cases

What Is Machine Learning?

Machine learning is a subset of artificial intelligence (AI) which provides machines the ability to learn automatically & improve from experience without being explicitly programmed.



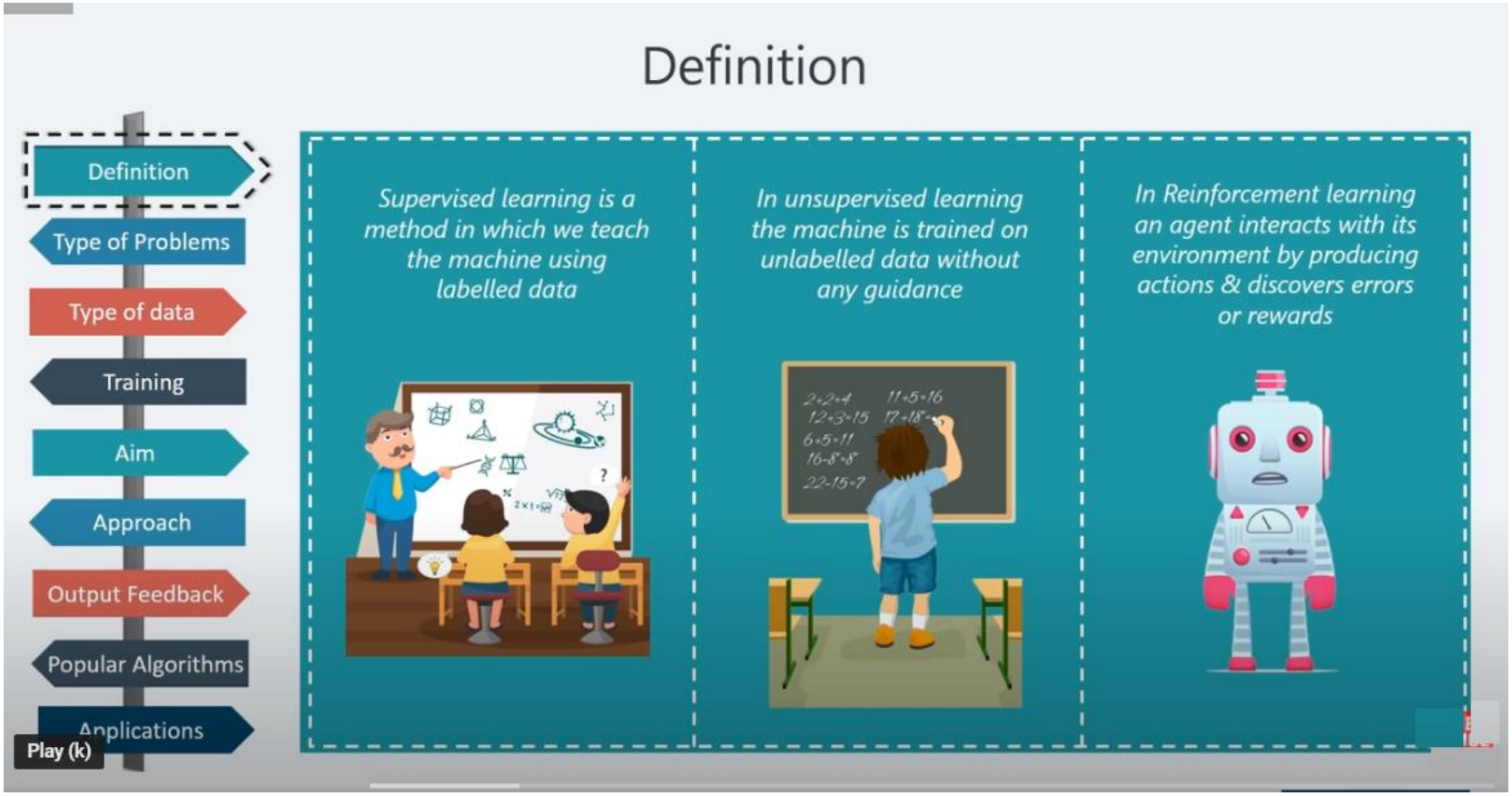
Types Of Machine Learning

Supervised Learning

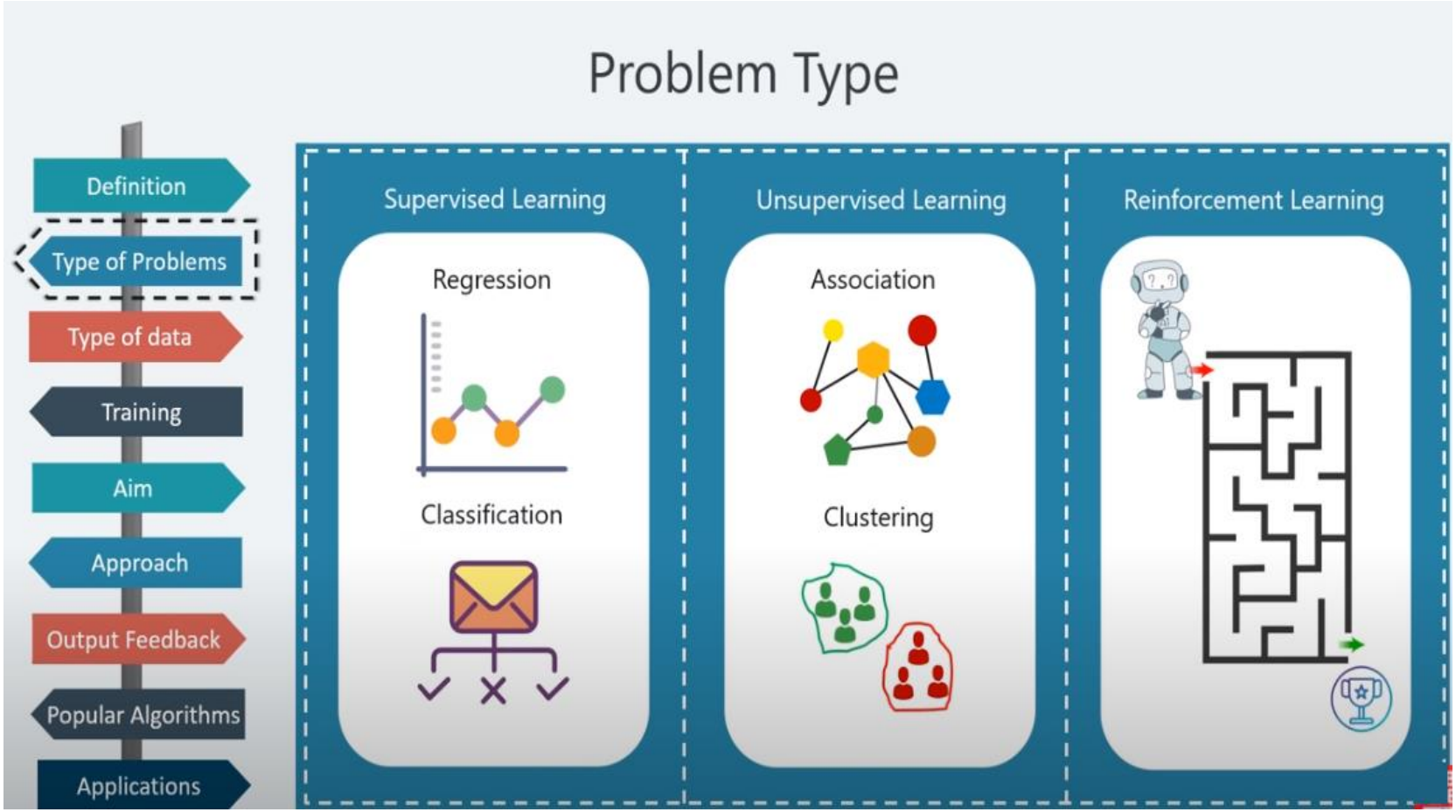
Unsupervised Learning

Reinforcement Learning

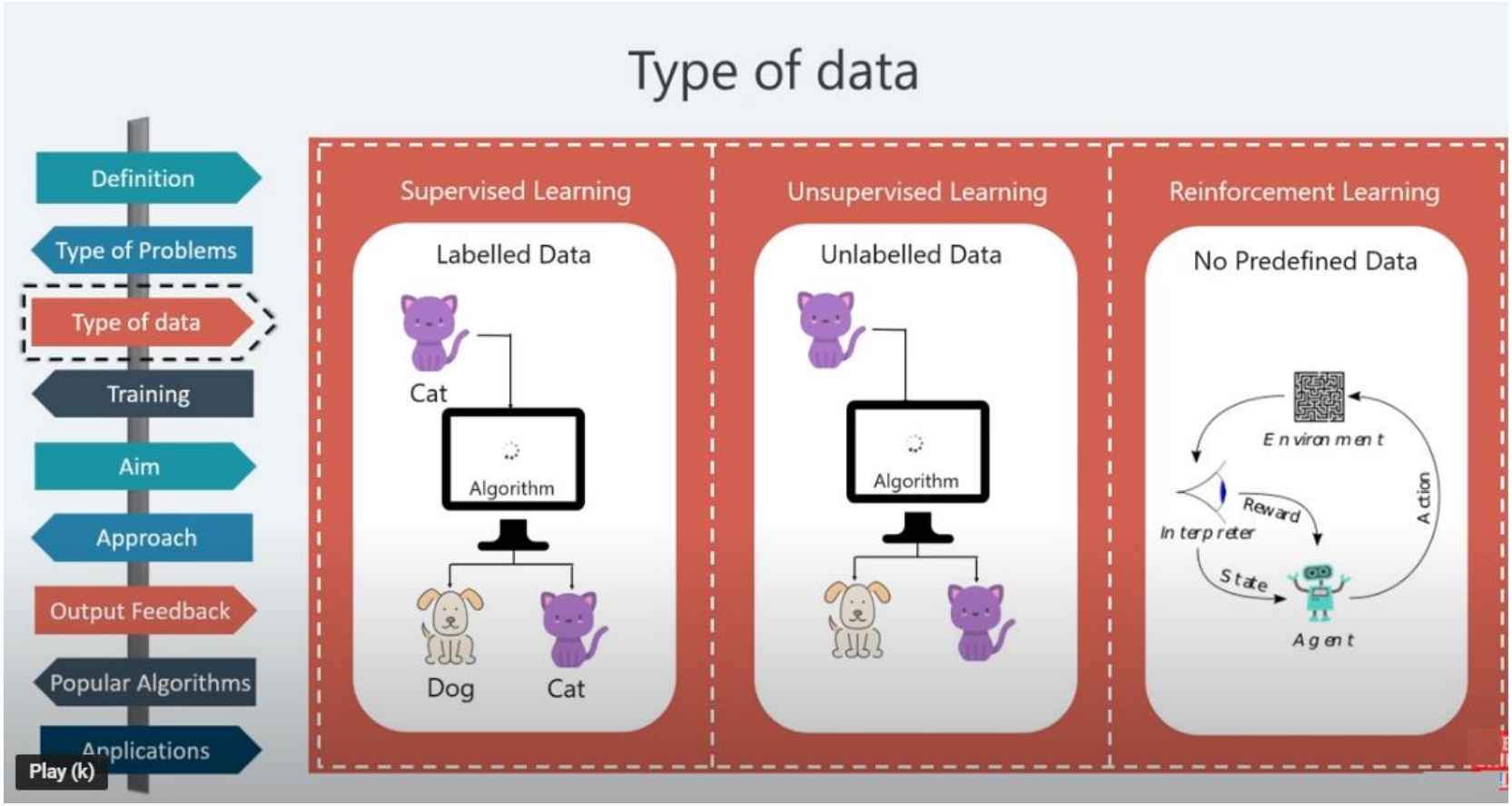
Difference Between Types of M/c Learning



Problem Types



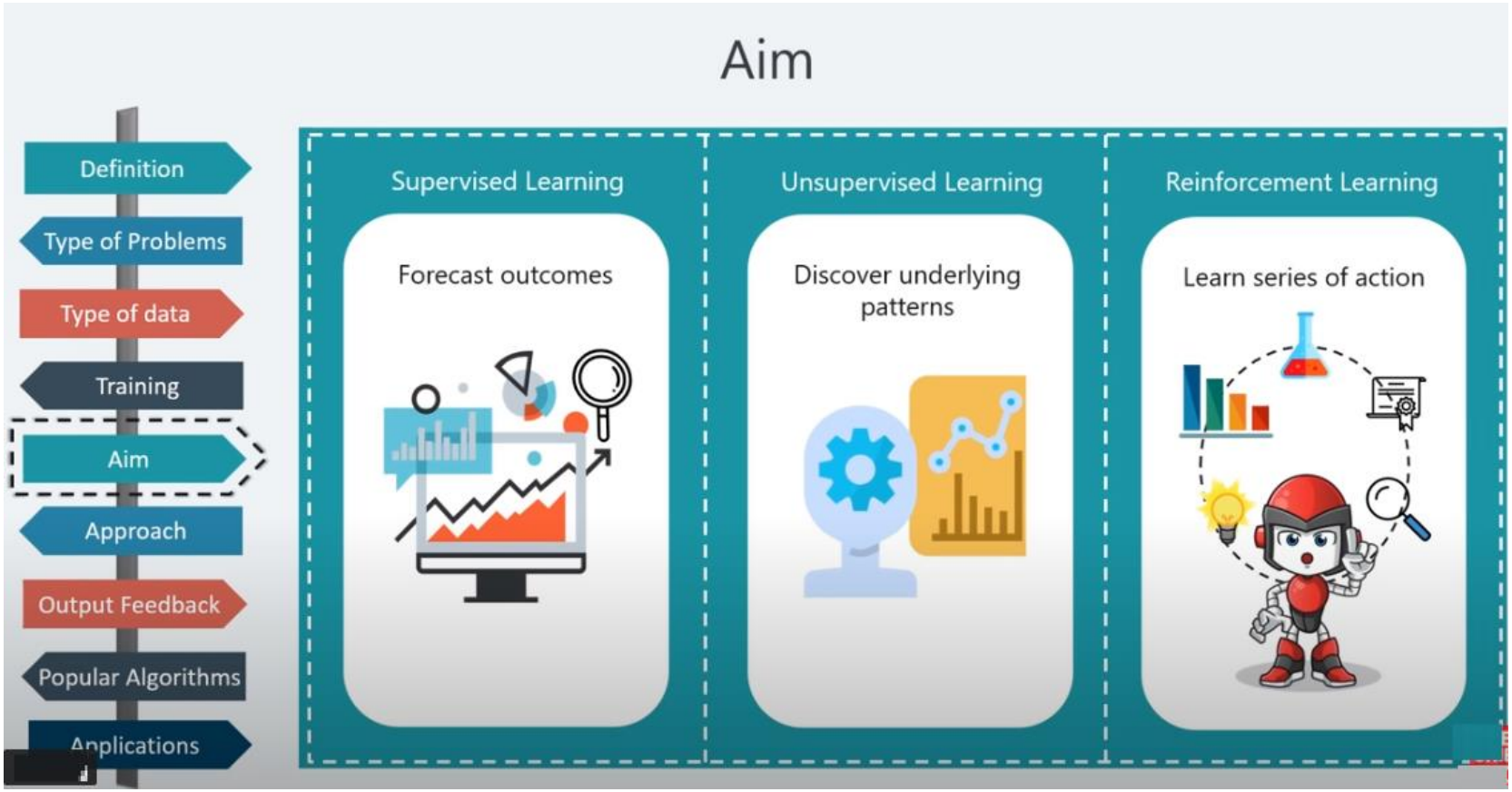
Types of Data



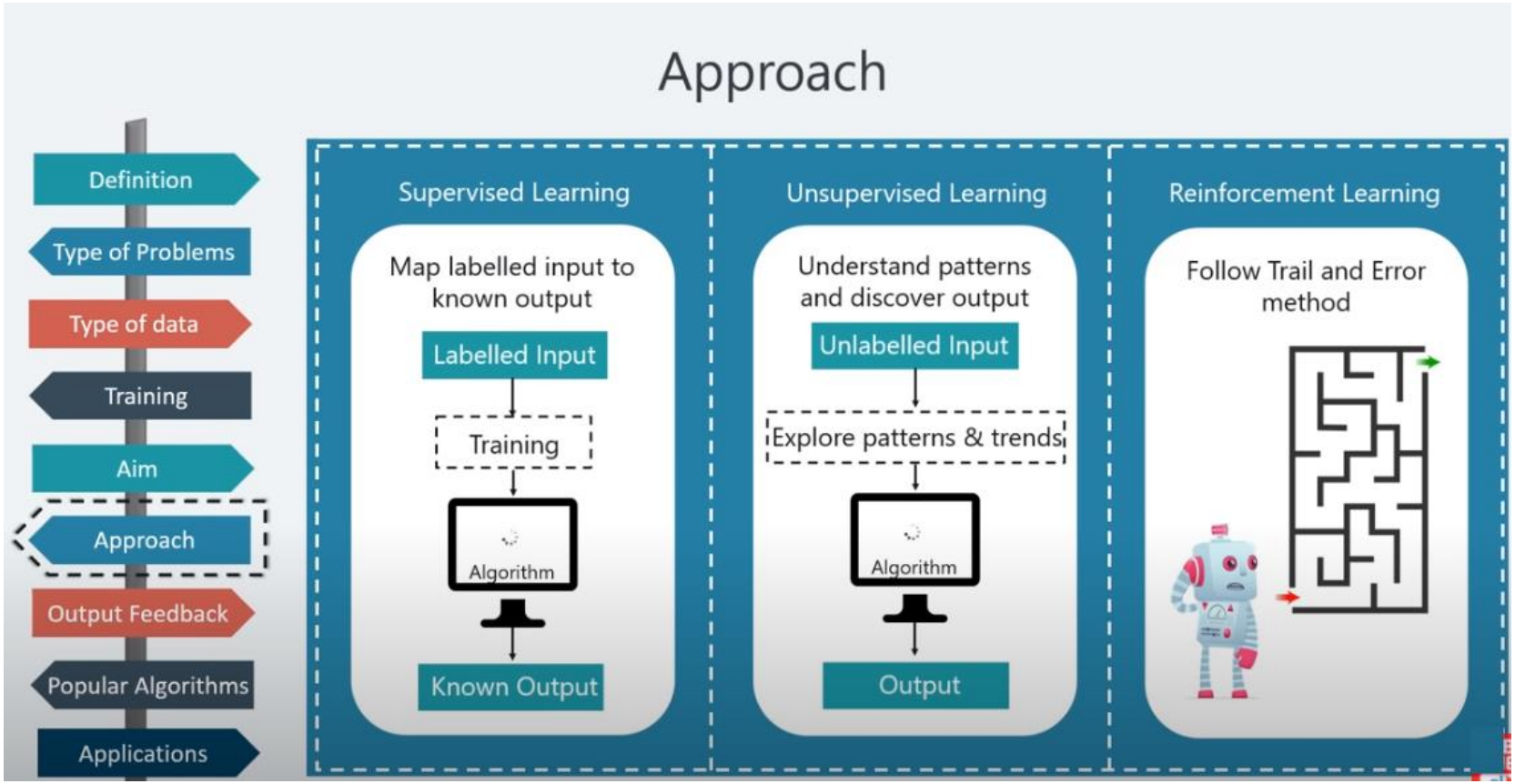
Training


Training
Definition
Type of Problems
Type of data
Training
Aim
Approach
Output Feedback
Popular Algorithms
Applications
Supervised Learning
External supervision
Unsupervised Learning
No supervision
Reinforcement Learning
No supervision

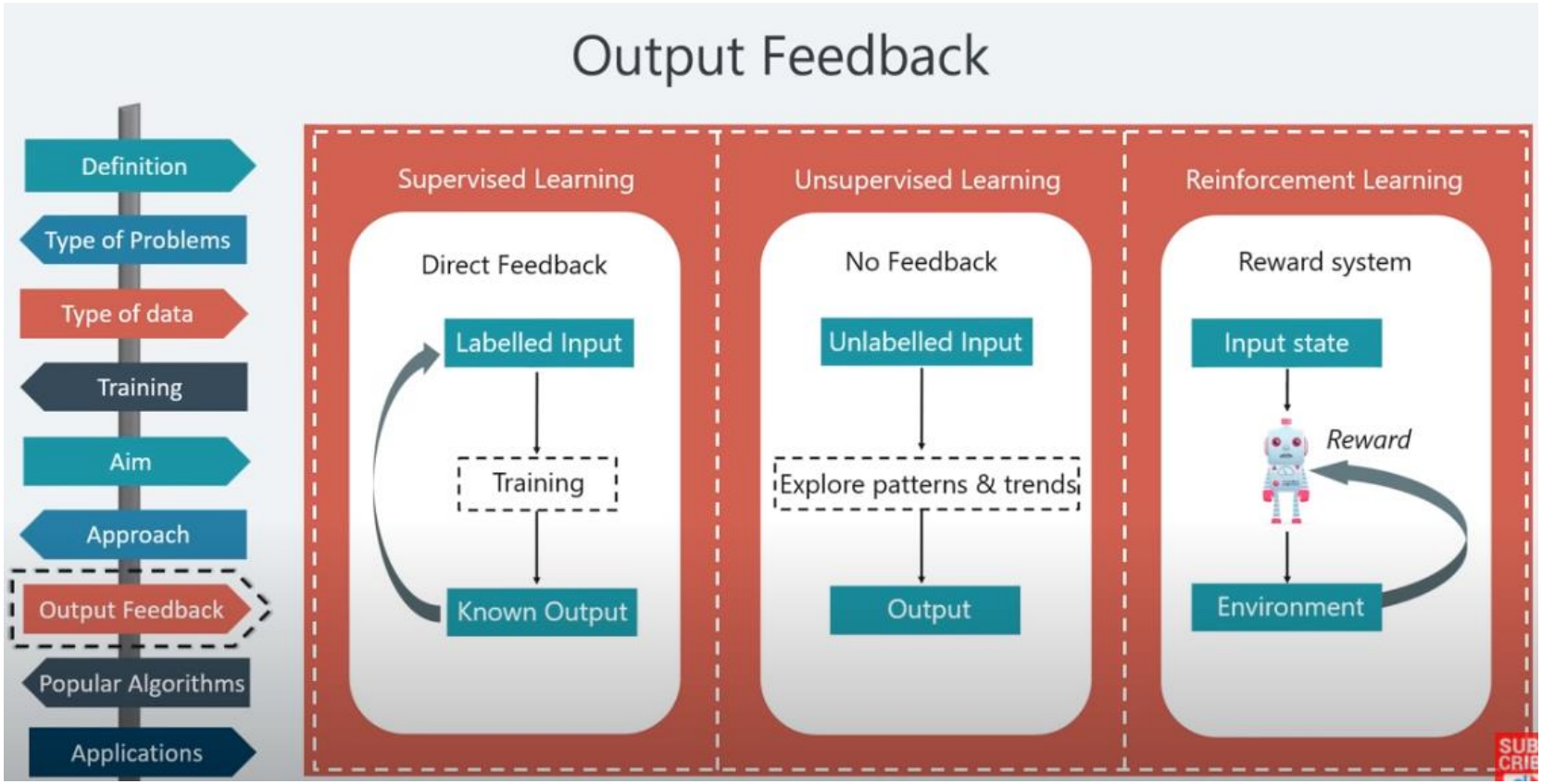
Aim


Aim
Definition
Type of Problems
Type of data
Training
Aim
Approach
Output Feedback
Popular Algorithms
Applications
Supervised Learning
Forecast outcomes
Unsupervised Learning
Discover underlying patterns
Reinforcement Learning
Learn series of action

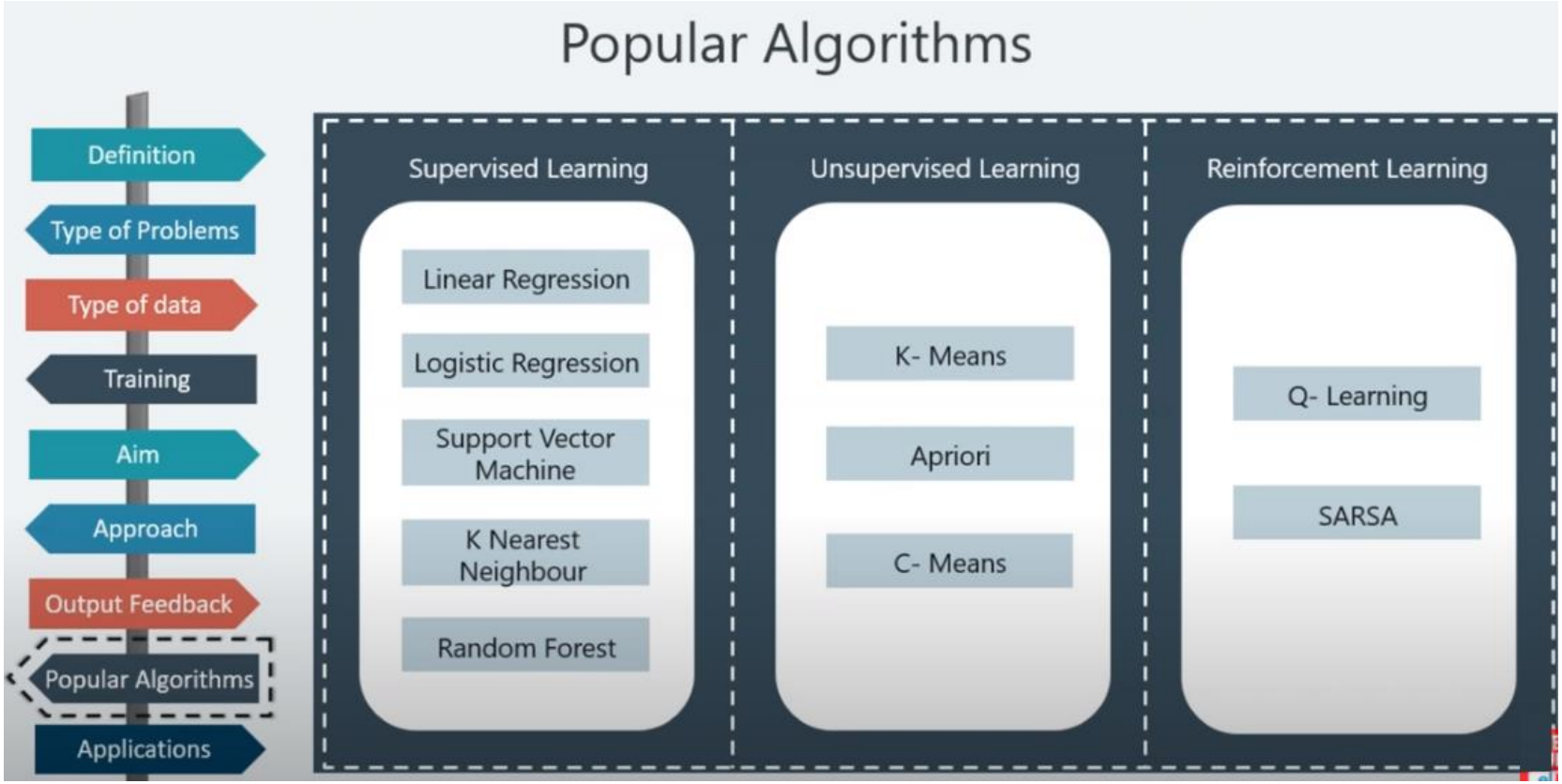
Approach


Approach
Definition
Type of Problems
Type of data
Training
Aim
Approach
Output Feedback
Popular Algorithms
Applications
Supervised Learning
Map labelled input to known output
Labelled Input
Training
Algorithm
Known Output
Unsupervised Learning
Understand patterns and discover output
Unlabelled Input
Explore patterns & trends
Algorithm
Output
Reinforcement Learning
Follow Trail and Error method

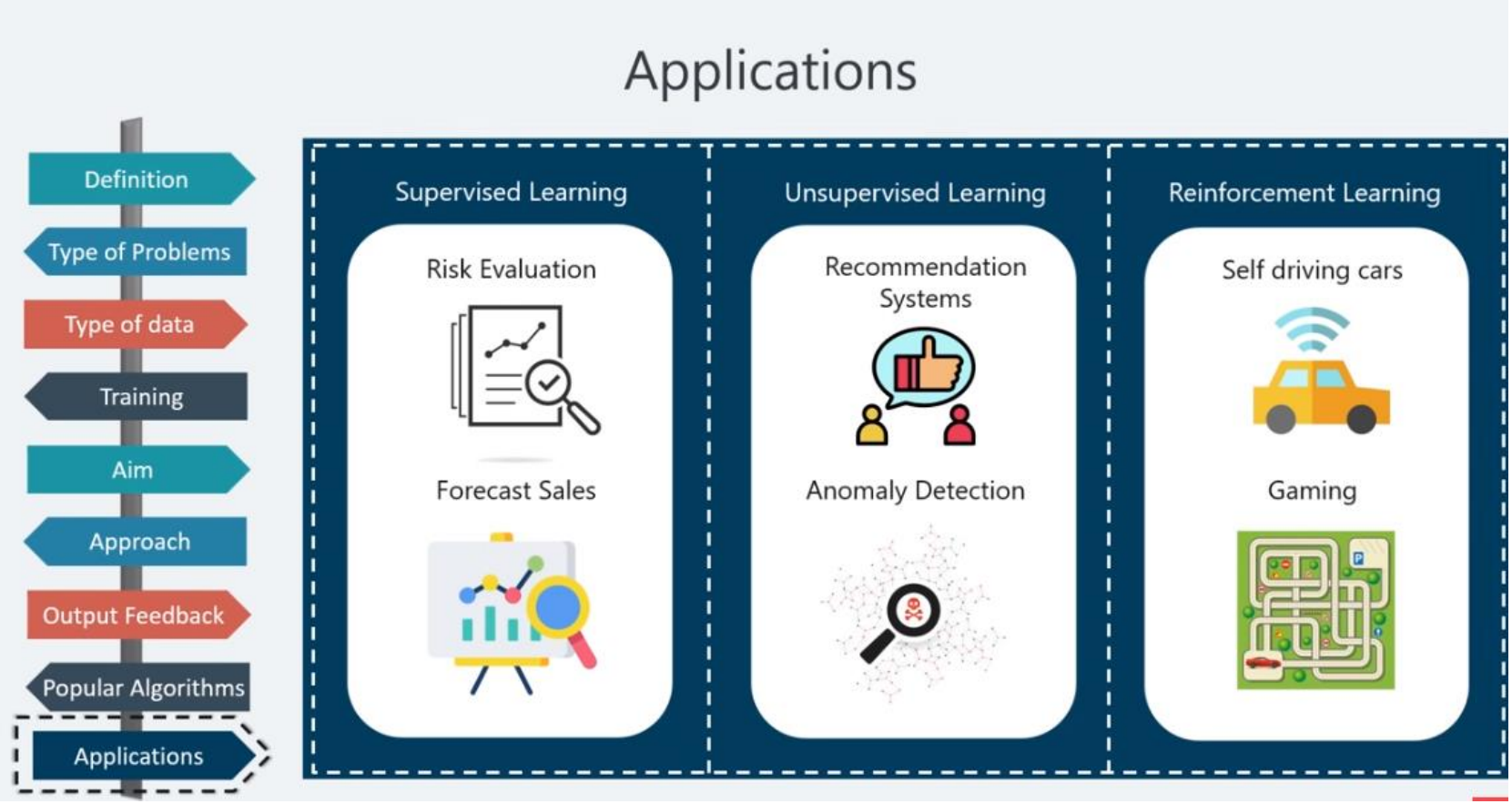
Approach


Output Feedback
Definition
Type of Problems
Type of data
Training
Aim
Approach
Output Feedback
Popular Algorithms
Applications
Supervised Learning
Direct Feedback
Labelled Input
Training
Known Output
Unsupervised Learning
No Feedback
Unlabelled Input
Explore patterns & trends
Output
Reinforcement Learning
Reward system
Input state
Reward
Environment

Algorithms


Popular Algorithms
Definition
Type of Problems
Type of data
Training
Aim
Approach
Output Feedback
Popular Algorithms
Applications
Supervised Learning
Linear Regression
Logistic Regression
Support Vector Machine
K Nearest Neighbour
Random Forest
Unsupervised Learning
K- Means
Apriori
C- Means
Reinforcement Learning
Q- Learning
SARSA

Applications

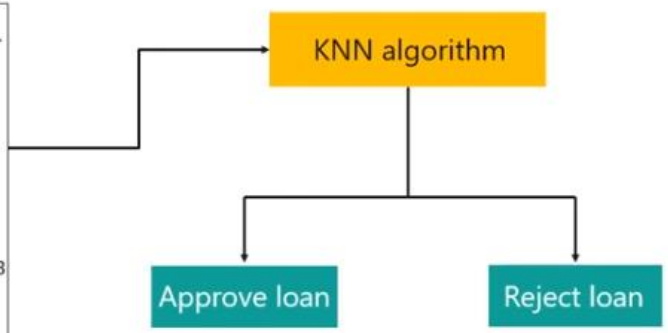


Use Cases

Use Case 1

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

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

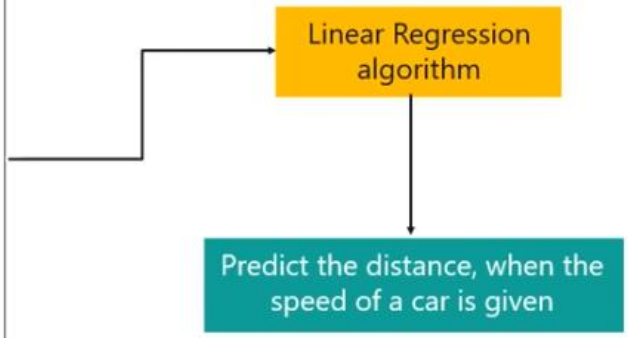


Use Cases

Use Case 2

Problem Statement: To establish a mathematical equation for **distance** as a function of **speed**, so you can use it to predict **distance** when only the **speed** of the car is known.

```
> cars
   speed dist
1      4    2
2      4   10
3      7    4
4      7   22
5      8   16
6      9   10
7     10   18
8     10   26
9     10   34
10    11   17
11    11   28
12    12   14
13    12   20
14    12   24
15    12   28
```

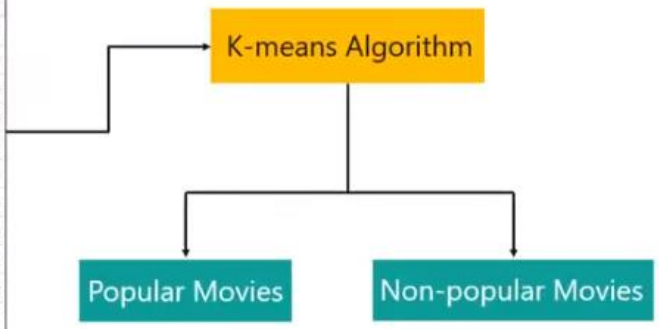


Use Cases

Use Case 3

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

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

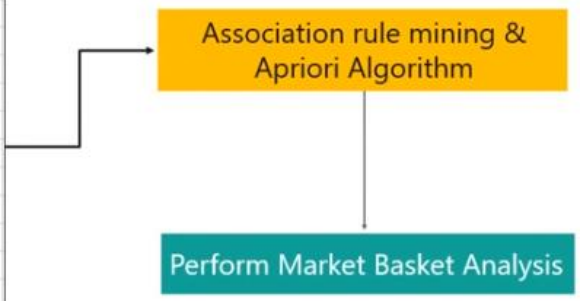


Use Cases

Use Case 4

Problem Statement: To perform Market Basket Analysis by finding association between items bought at the grocery store

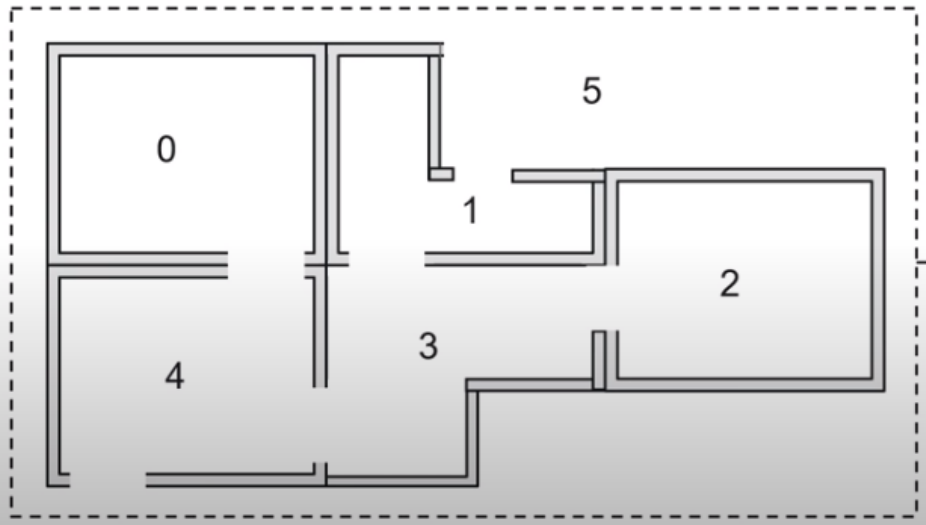
	A	B	C	D	E	F	G	H
1	citrus fruit	semi-finisl	margarine	ready soups				
2	tropical fr	yogurt	coffee					
3	whole milk							
4	pip fruit	yogurt	cream che	meat spreads				
5	other vege	whole mill	condensec	long life bakery product				
6	whole mill	butter	yogurt	rice	abrasive cleaner			
7	rolls/buns							
8	other vege	UHT-milk	rolls/buns	bottled be	liquor (appetizer)			
9	pot plants							
10	whole mill	cereals						
11	tropical fr	other vege	white brea	bottled wa	chocolate			
12	citrus fruit	tropical fr	whole mill	butter	curd	yogurt	flour	bottled wa
13	beef							
14	frankfurte	rolls/buns	soda					
15	chicken	tropical fruit						

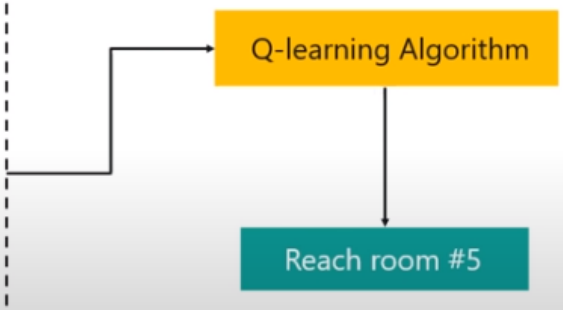


Use Cases

Use Case 5

Problem Statement: Place an agent in any one of the rooms (0,1,2,3,4) and the goal is to reach outside the building (room 5)





Concept Learning

- Concept learning also refers to a learning task in which a human or machine learner is trained to classify objects by being shown a set of example objects along with their class labels. The learner will simplify what has been observed in an example. This simplified version of what has been learned will then be applied to future examples

- Concept learning can be viewed as the task of searching through a large space of hypotheses implicitly defined by the hypothesis representation.

- The goal of this search is to find the hypothesis that best fits the training examples.

- By selecting a hypothesis representation, the designer of the learning algorithm implicitly defines the space of all hypotheses that the program can ever represent and therefore can ever learn.

Concept Learning

Consistent Hypothesis and Version Space

An hypothesis h is **consistent** with a set of training examples D iff $h(x) = c(x)$ for each example in D

$$Consistent(h, D) \equiv (\forall \langle x, c(x) \rangle \in D)\ h(x) = c(x))$$

Example	Citations	Size	InLibrary	Price	Editions	Buy
1	Some	Small	No	Affordable	One	No
2	Many	Big	No	Expensive	Many	Yes

h1 = (?, ?, No, ?, Many) **– Consistent**

h2 = (?, ?, No, ?, ?) **– Not Consistent**

Concept Learning

Consistent Hypothesis and Version Space

- The version space $VS_{H,D}$ is the subset of the hypothesis from H *consistent* with the training example in D

$$VS_{H,D} \equiv \{h \in H \mid Consistent(h, D)\}$$

Concept Learning

List-Then-Eliminate algorithm

Version space as list of hypotheses

1. *VersionSpace* ← a list containing every hypothesis in H
2. For each training example, $\langle x, c(x) \rangle$ Remove from *VersionSpace* any hypothesis h for which $h(x) \neq c(x)$
3. Output the list of hypotheses in *VersionSpace*

Concept Learning

Consistent Hypothesis and Version Space

- F1 - > A, B
- F2 - > X, Y
- **Instance Space:** (A, X), (A, Y), (B, X), (B, Y) – **4 Examples**
- **Hypothesis Space:** (A, X), (A, Y), (A, **ø**), (A, ?), (B, X), (B, Y), (B, **ø**), (B, ?), (**ø**, X), (**ø**, Y), (**ø**, **ø**), (**ø**, ?), (**?**, X), (**?**, Y), (**?**, **ø**), (**?**, ?) - **16 Hypothesis**
- **Semantically Distinct Hypothesis :** (A, X), (A, Y), (A, ?), (B, X), (B, Y), (B, ?), (**?**, X), (**?**, Y (**?**, ?), (**ø**, **ø**) – **10**

Concept Learning

Consistent Hypothesis and Version Space

- Version Space: (A, X), (A, Y), (A, ?), (B, X), (B, Y), (B, ?), (?, X), (?, Y) (?, ?), (ø, ø),
- Training Instances

F1	F2	Target
A	X	Yes
A	Y	Yes

- Consistent Hypothesis are: (A, ?), (?, ?)

Concept Learning

***List-Then-Eliminate* algorithm**

Problems

- The hypothesis space must be finite
- Enumeration of all the hypothesis, rather inefficient

Terminologies used in machine learning

- Here have explained what the terms mean and given examples of the same.
- We will we looking at following terms: -
- Label
- Features
- Examples
- Labelled data - Unlabelled Data
- Regression
- Classification

Terms……

Labels

Label is a value or thing we are trying to predict.

The label could be future price of a product, it can be whether the email needs to be routed to SPAM or INBOX

If we take example of following equation :

$$Y = Mx + C$$

So Y is the label in this case.

Terms……

Features

A feature is an input variable - the x variable in simple linear regression

$$Y = Mx + C$$

A simple machine learning project might have just one feature.

while a more complex machine learning project could use hundreds of features like:

x1,x2,...x100

In the modeling and prediction of what would be the future price of product, the features could include the following:

- Pack size of the product
- Month of the year
- Competitors price of similar product

Terminologies……

Examples

An "**example**" is a particular instance of data.

Examples are of two categories:

- Labeled

  This includes both features and Label

- Unlabeled

  This includes only features

Terminologies……

Labelled Examples

Here first 3 columns are features and 4th column is the label

Labelled examples are use to train and test the models to be used for predictions.

month of the year	pack size	competitor price	Products retail price
1	small	10	9.5
2	small	11	10.5
3	small	10.5	10
4	small	9.5	9

Terminologies……

Unlabelled Examples

It contains only features and no label

The model trained using labelled examples is then used to predict the labels on unlabelled examples

month of the year	pack size	competitor price	
1	small	10	
2	small	11	
3	small	10.5	
4	small	9.5	

Terminologies……

Model

It defines the relationship between features and label.

This relationship is derived by trying to fit various readily available algorithms or writing an custom algorithm.

Two key terms related to models are :

- Training – This is the process where we feed the labelled data to the model and make it learn the relationship between features and label
- Prediction – This is the process we feed unlabelled data to the trained model and obtain the values of labels

Terminologies……

Regression vs Classification

A regression model predicts continuous values.

For example, regression models make predictions that answer questions like the following

- What is the value of a house in California?
- What is the probability that a user will click on this ad?

A classification model predicts discrete values.

For example, classification models make predictions that answer questions like the following:

- Is a given email message spam or not spam?
- Is this an image of a dog, a cat, or a hamster?

Terminologies……

- Analytics : Descriptive, Predictive , Prescriptive.
- Visualization : Data in to Graphs

Terminologies……

- DataSet
- DataFrame
- Data
- RawData

Storage of data

Terminologies……

- Outliers
- Missing Values/Imputation
- Feature Selection/Dimensionality Reduction
- Imbalance Data: Oversampling and Undersampling
- Time Series Data
- Feature Engineering

Terminologies……

Outliers = ODDS

Detecting Outliers:

BoxPlot

Quartliles

Scatter Plot

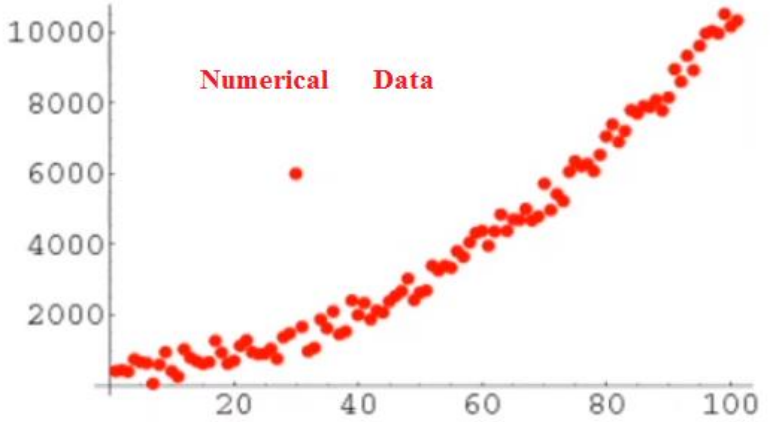
Z-score

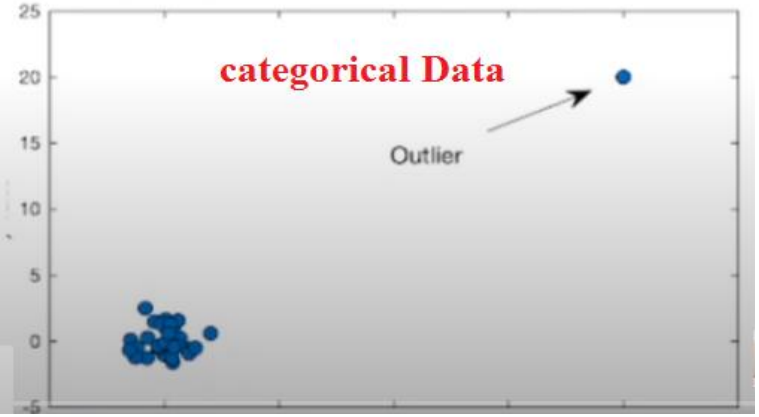
Treating Outliers

Capping

Deletion

Replacing by mean,median and mode





Terminologies……

Input and Output

- Input Variables | Features | Columns | Dimensions | Characteristics | Independent Variables | X | Multiple

- Output variables | Outcome | Result | Target | Y | Dependent Variables | Y Predicted | | Single

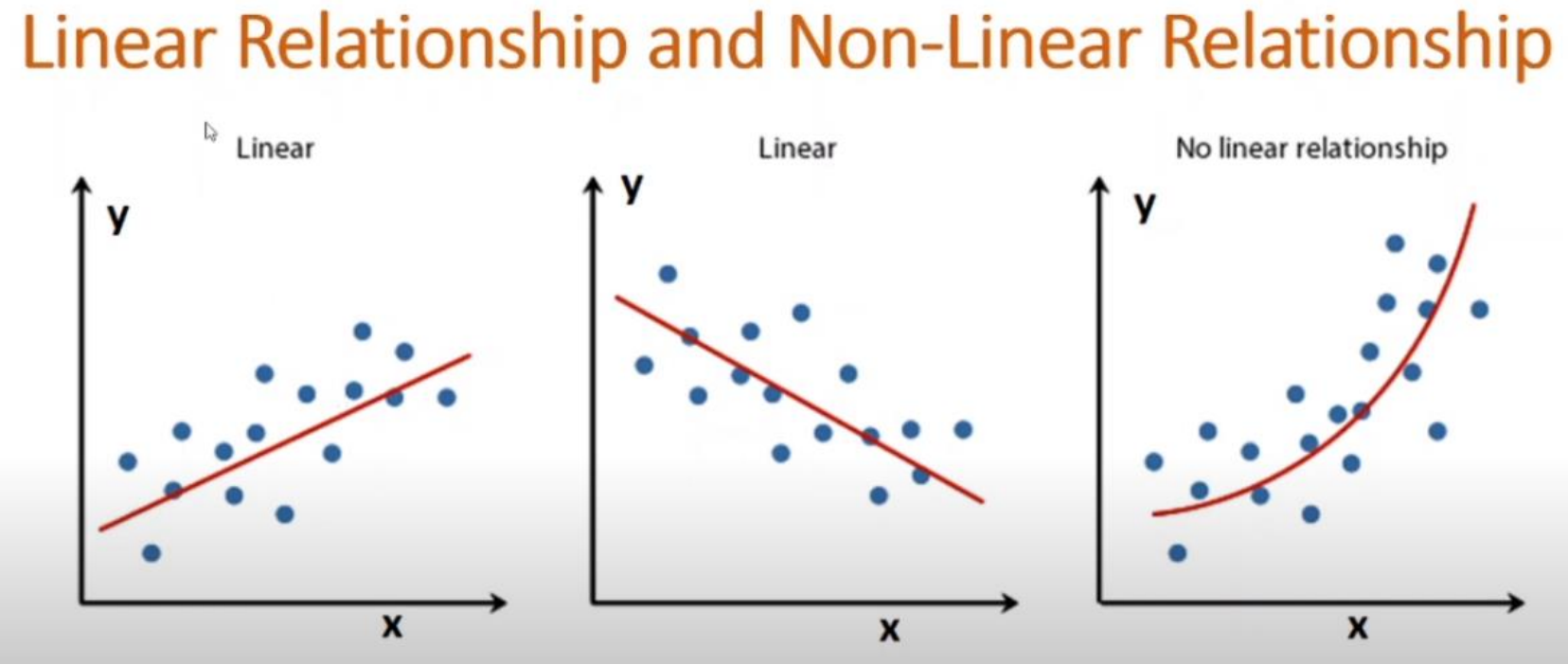
Terminologies……

Normalization

Normalization = Scaling

Normalization means to scale a variable to have a values between 0 and 1 . Goal of Normalization is to change the values of numerical columns in the dataset to a common scale, without distorting differences in the ranges of values.

Terminologies……



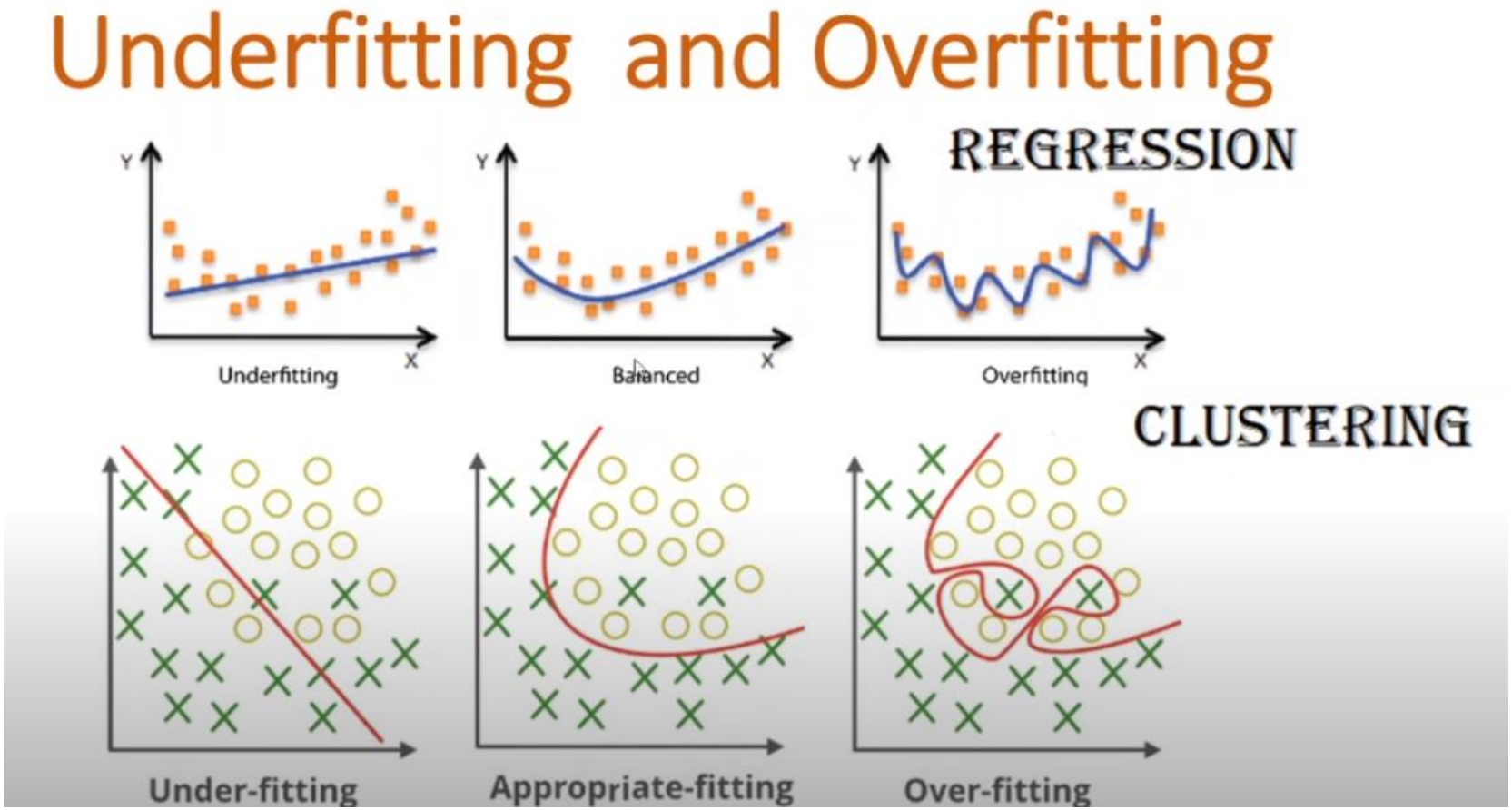
Terminologies……

Train and Test Splitting

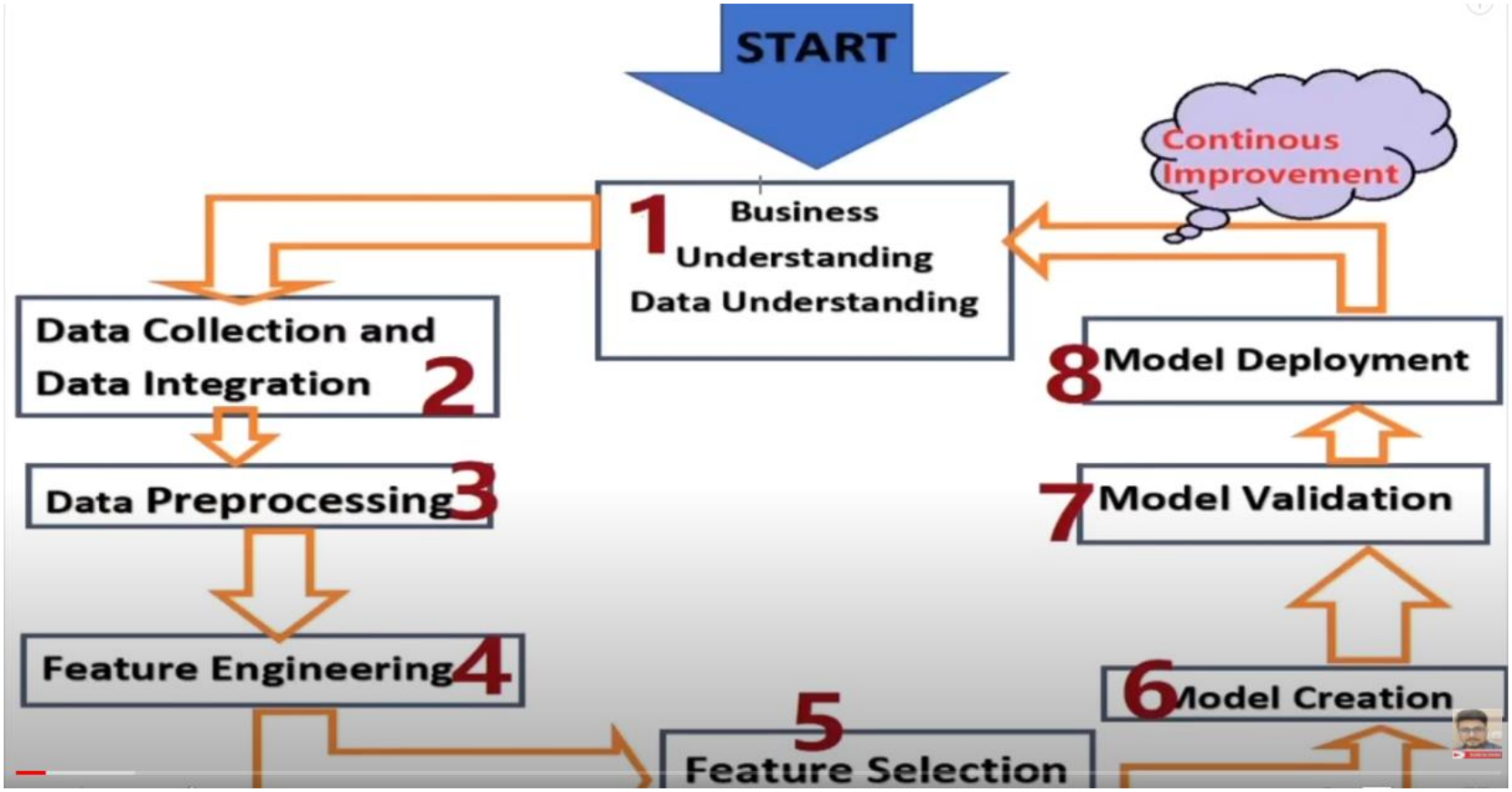
- X_train , Y_ train
- X_test , Y_test

Random State and Sample Size

Terminologies……



Life Cycle


START
Continous Improvement
1 Business Understanding Data Understanding
Data Collection and Data Integration 2
Data Preprocessing 3
Feature Engineering 4
5 Feature Selection
6 Model Creation
7 Model Validation
8 Model Deployment

Conti....

1) BUSINESS UNDERSTANDING AND DATA UNDERSTANDING

- Nature of business
- Requirements
- What we need ?
- Goal (Prediction , Descriptive and Prescriptive)
- Final Deployment (On-premises or Clodu) (Online or Offline Learning)
- UAT
- Timelines/Deadlines
- Read all Documents/SOPS

Conti……

2) DATA COLLECTION AND INTEGRATION

- Type of Data (structured/unstructured or offline/online)
- Source of Data (on premises or Cloud)
- Size of data
- Format of data
- Data transport
- Data Ingestion/Data Collection from all sources
- Data Integration (BODS ETL)

Conti....

3) DATA PREPROCESSING (cd mouz)

- Unstructured in to Structured Format of data.
- **Cleaning** : (Noise,Special Chars,Lower),
- Removing **Duplicate**,
- Removing **Missing**,
- Removing **Outliers**,
- **Unidecode**,
- Removing **Zeroes** values(imputation,
- **Rounding**,
- **Formatting** and **Repairing**(MCORDF)
- Organize the data : version conrol of data , auditing , maintaining data, Court Reviews conferences.

Conti……

4) FEATURE ENGINEERING (td soieng)

- Data **Transformation** :Normalization (skewing (z-score)), Logarithm log(x) (heteroscedasticity), 1/x, sqrt(x), exp(x)
- **Discretization** : Binning , Equal frequency discretization, Equal length discretization,
- **Scaling** : Standardization , Min-Max Scaling, Mean Scaling, Max Absolute Scaling, Unit norm-Scaling
- **One hot encoding** , dummy variables , Rare variables .
- **Imbalance dataset** : SMOTE, SMOTEtomek , SMOTEEN
- **Extracting** features from text: Bag of words, Tfidf, n-grams,Word2vec, topic extraction
- create some **new** features by using domain knowledge from domain expertise or by using internet and google.
- Performing Statistical and **Graphical** Data Analysis (EDA) : story telling

Conti....

5) Feature Selection

- Convert many features in to important features
- Correlation, Heat Matrix (covariance)
- Multi colinearity (RIDGE, Combining)
- Backward Elimination
- PCA
- NMF
- ICA and FastICA
- SVD

Sklearn library and scikit library

Conti.....

6) Model Creation

- Basis of EDA , we consider Algorithms.
- Free Lunch Theorem **no model is work well for every problem meaning of Free lunch theorem**
- We consider at least 5 to 6 models.
- **Regression** : Linear, Lasso, SVR, Random Forest Classfier, Adaboost Regressor , XG Boost Regressor
- **Classification**: Logistics, SVC, NB, Random Forest Classifier, Adaboost Classifier , XG Boost Regressor, Light GBM and CatGBM
- **Anomaly detection** : Isolation Forest , logistics and Local outlier Factor.

Conti.....

7) Model Validation and Model Selection

- We will apply cross validation on all models
- Time and effort and resources if need to consider.
- Tradeoff between Bais/variance and True Positive/True Negative.
- Best Model will be selected for further tuning. (confusion matrix in case of Imbalance Dataset)
- Hyper Parameter optimisation is always important (Grid Search and Random Search)
- Automated code should be written that which model out of these 4-5-6 models.

Conti.....

8) Model Deployment

- Will generate the pickle file for ML Models
- Develop Front end API using Flask or Django framework
- On premises or Cloud
- Storing the prediction in the Storage.
- Setting up the logging and monitoring frameworks to generate reports and dashboards based on the client requirements and to do continuous monitoring the output to find out whether the model is preforming well or not.

Conti……

Continuous Improvement

- Internal Evaluation
- External Evaluation
- UAT done by customer to check how model is responding as per their expectations

Scikit-learn

- Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.
-

Features

Rather than focusing on loading, manipulating and summarising data, Scikit-learn library is focused on modeling the data. Some of the most popular groups of models provided by Sklearn are as follows −

- **Supervised Learning algorithms** − Almost all the popular supervised learning algorithms, like Linear Regression, Support Vector Machine (SVM), Decision Tree etc., are the part of scikit-learn.
- **Unsupervised Learning algorithms** − On the other hand, it also has all the popular unsupervised learning algorithms from clustering, factor analysis, PCA (Principal Component Analysis) to unsupervised neural networks.
- **Clustering** − This model is used for grouping unlabeled data.
- **Cross Validation** − It is used to check the accuracy of supervised models on unseen data.

Continue

- **Dimensionality Reduction** − It is used for reducing the number of attributes in data which can be further used for summarisation, visualisation and feature selection.
- **Ensemble methods** − As name suggest, it is used for combining the predictions of multiple supervised models.
- **Feature extraction** − It is used to extract the features from data to define the attributes in image and text data.
- **Feature selection** − It is used to identify useful attributes to create supervised models.

Dataset Loading

- A collection of data is called dataset. It is having the following two components −
- **Features** − The variables of data are called its features. They are also known as predictors, inputs or attributes.
- **Feature matrix** − It is the collection of features, in case there are more than one.
- **Feature Names** − It is the list of all the names of the features.
- **Response** − It is the output variable that basically depends upon the feature variables. They are also known as target, label or output.
- **Response Vector** − It is used to represent response column. Generally, we have just one response column.
- **Target Names** − It represent the possible values taken by a response vector.
- Scikit-learn have few example datasets like **iris** and **digits** for classification and the **Boston house prices** for regression.