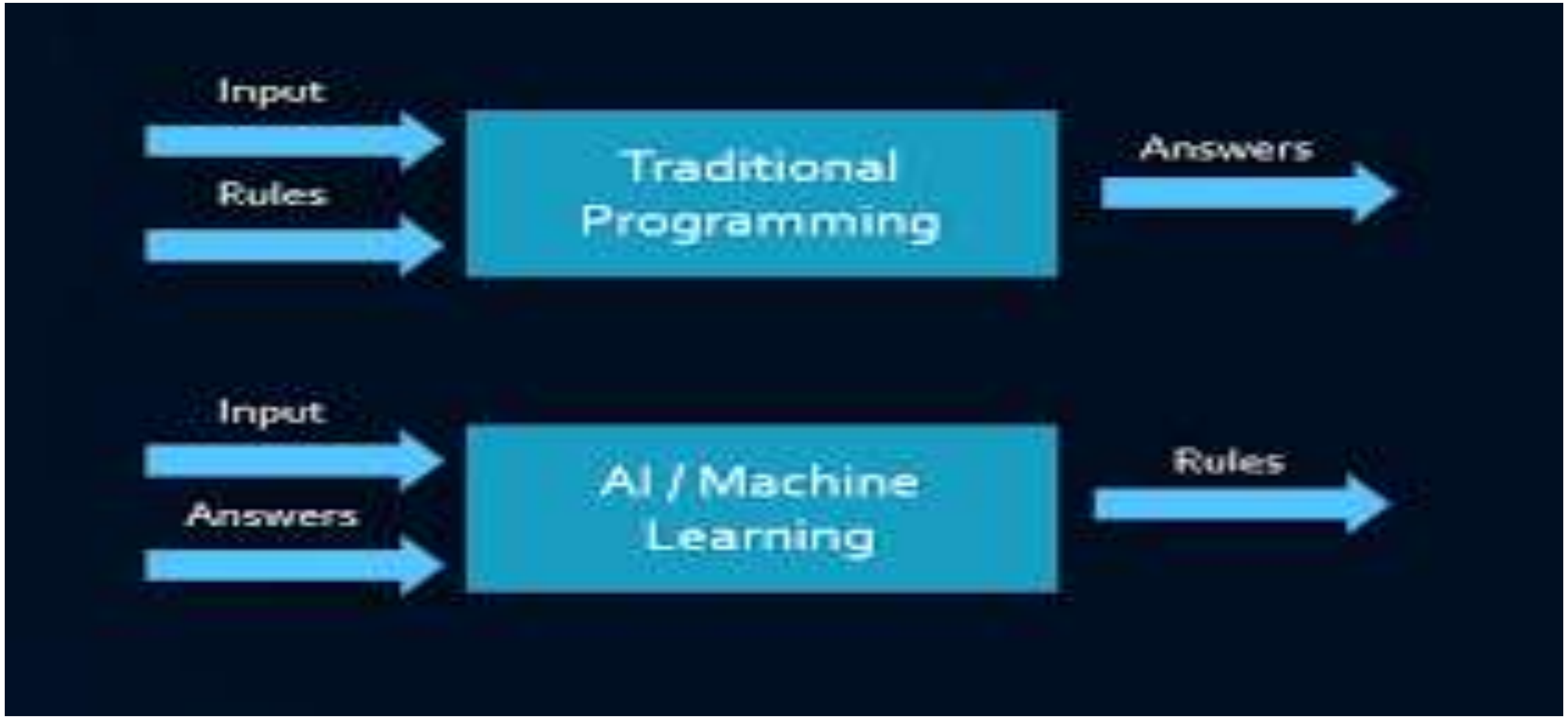
Machine Learning

Machine learning

- “Field of study that gives computers the capability to learn without being explicitly programmed”.
- Automating and improving the learning process of computers based on their experiences without being actually programmed i.e. without any human assistance.
- Main focus: Development of computer programs that can access data and use it to learn for themselves.
  - **Feed good quality data ->train the machine->build a model**
- Algorithm depends upon **type of data they input and output,** and **type of task or problem** that they are intended to solve.

Machine Learning

- **Traditional programming:** Feed input data to a program and then fetch output data
- **Machine learning:** Feed input data and expected output, run on the machine, then machine creates its own program(logic). This logic can be evaluated during testing.



Applications

- Video recommendations in YouTube, Netflix
- Predictions of words
- Image recognition, face prediction in Facebook
- Online Fraud detection
- Traffic prediction

Working of ML

- Machine learning algorithm works basically in 2 main phases:

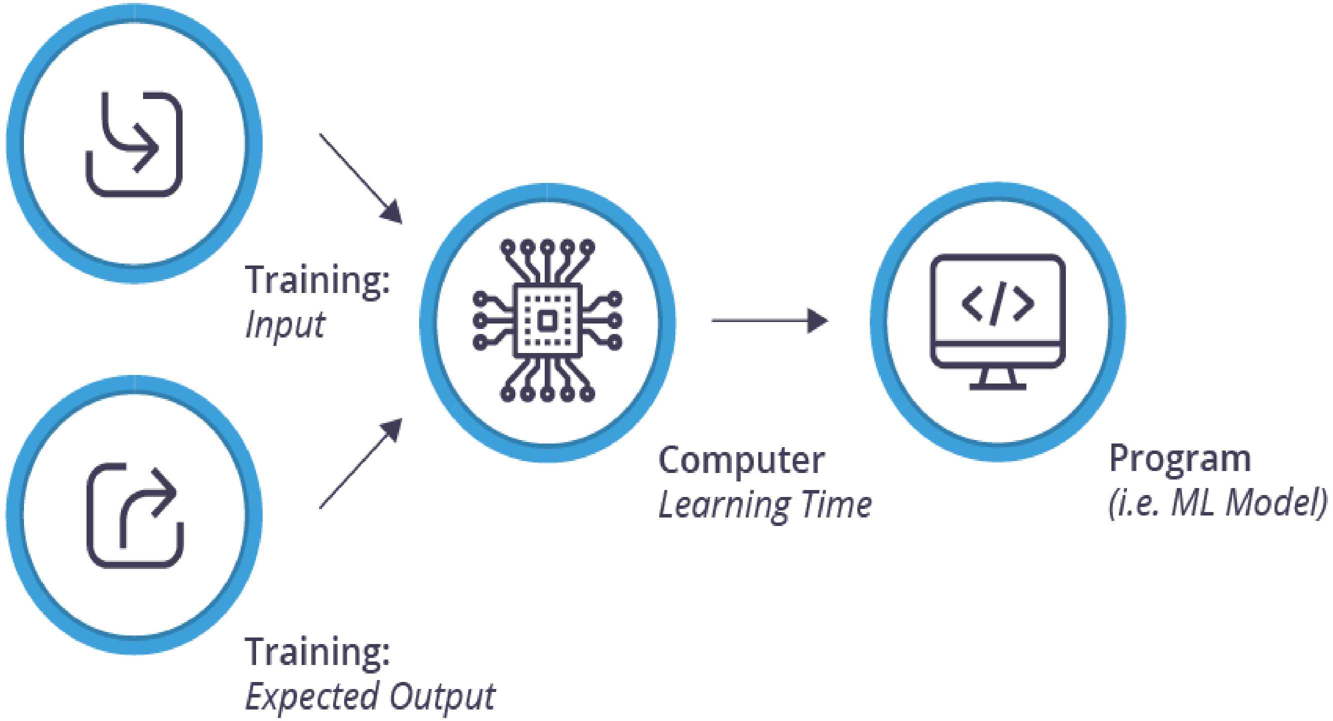
➢ Training phase

➢ Inference phase

Training phase

- Training: process of creating a machine learning algorithm
- Basically teach a machine to make a logic
- Also known as learning phase
- Provides a training set to the machine

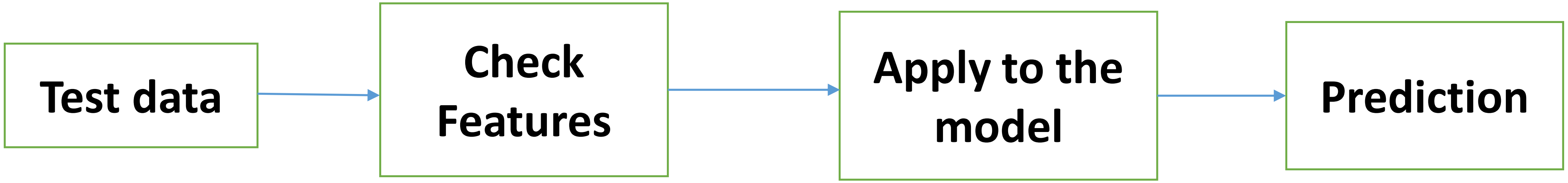
The Machine Learning Training Process



Inference phase

- Process of running **live data points into a machine learning algorithm to calculate an output** such as a single numerical score
- Process of using a machine learning algorithm to make a prediction

Inference from ML model



Test data

Check Features

Apply to the model

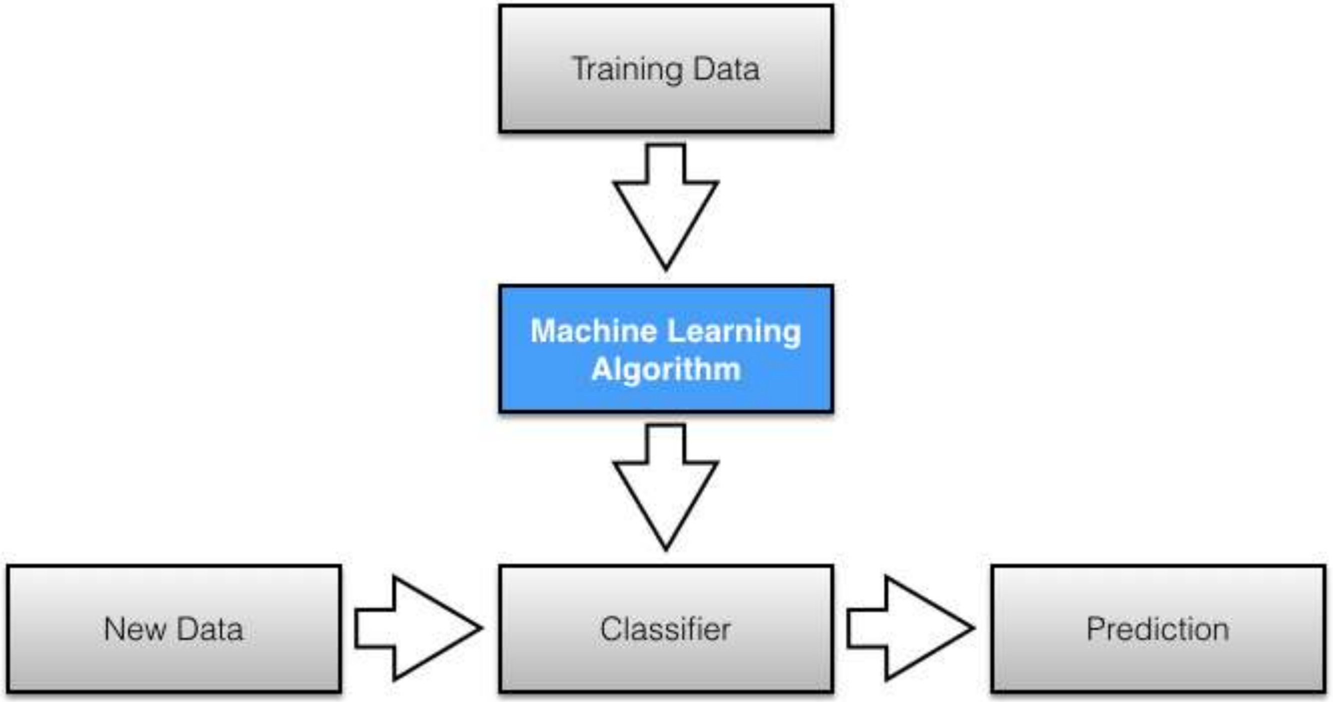
Prediction

Machine learning Cycle

- Steps of Machine learning with data preparation are as follows:

1. Data collection
2. Data preparation
3. Choose a model
4. Train the model
5. Evaluate/test the model
6. Parameter tuning
7. Make predictions

Machine learning working



Training Data

Machine Learning Algorithm

New Data

Classifier

Prediction

What is data and why?

- It can be any unprocessed fact, value, text, sound or picture that is not being interpreted or analyzed
- Its is the most important part of data analytics, machine learning and artificial intelligence
- Quality data is necessary for ML models to operate efficiently
- Clear understanding of how an ML works mandatorily needs understanding of the data which it operates

Data, Information and Knowledge

- **Information** is nothing but the refined form of data, which is helpful to understand the meaning.
- **Knowledge** is the relevant and objective information that helps in drawing conclusions.
- Data compiled in the meaningful context provides information.

Datasets in Machine learning

- The data used to build the final model usually comes from multiple datasets.
- Three datasets are commonly used in different stages of the creation of the model.

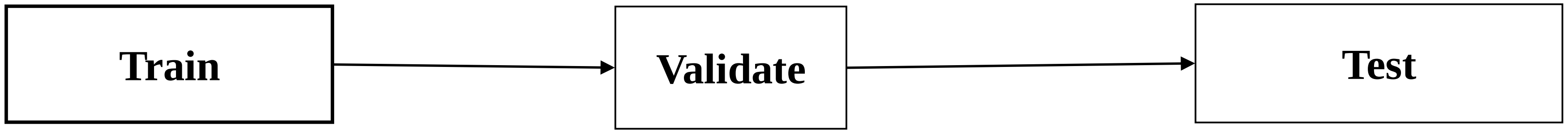
➢ Training data set

➢ Validate dataset

➢ Testing dataset

3 types of dataset

- Training dataset : dataset of examples used during the learning process and used to fit the parameters
- Validate dataset: dataset of examples used to tune the hyper parameters (i.e. the architecture) of a classifier.
- Testing dataset : dataset used to provide an unbiased evaluation of a final model fit on the training dataset



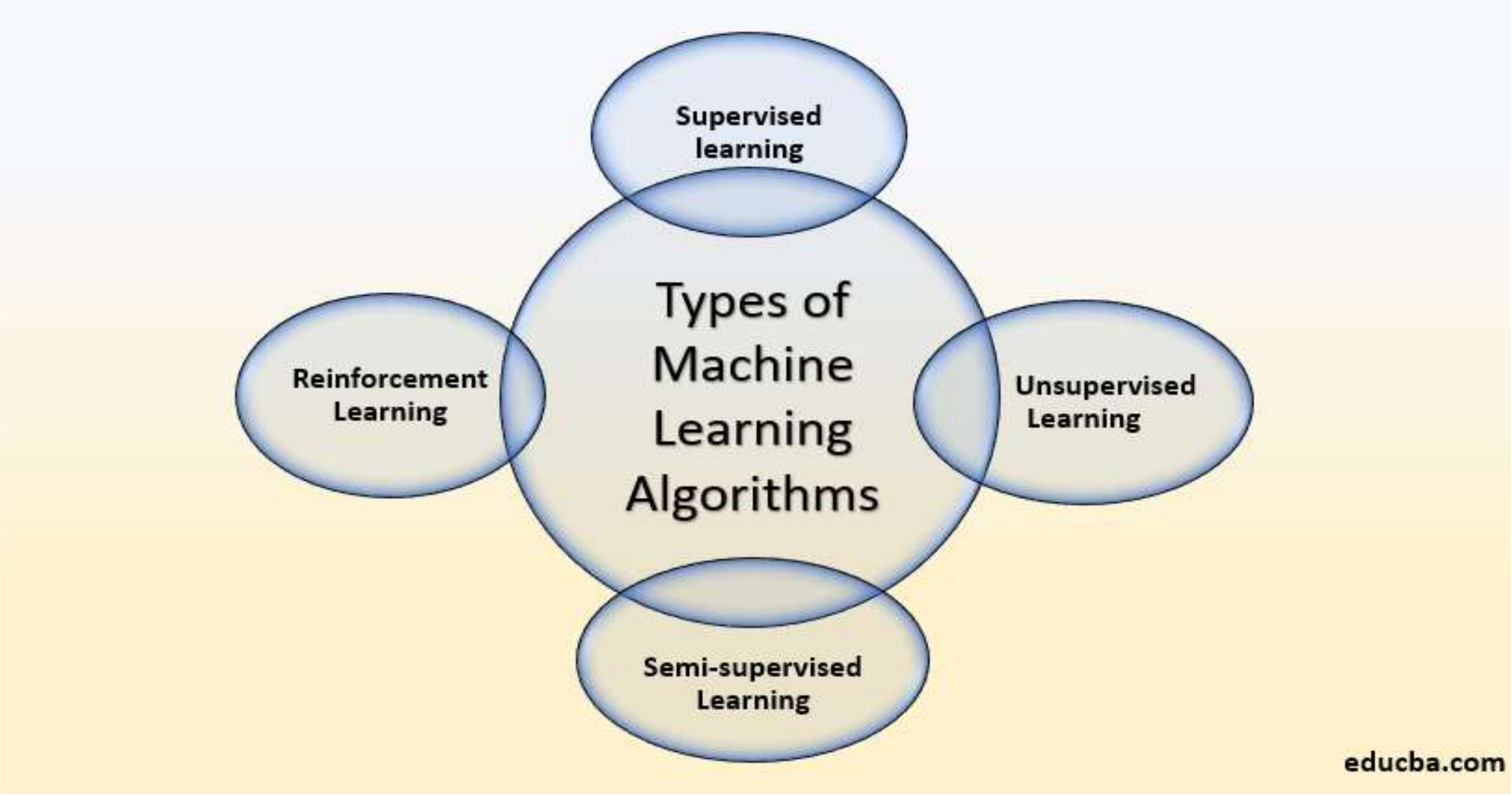
Train

Validate

Test

- The train-test-validation split ratio is also quite specific to your use case
- It gets easier to make judgement as you train and build more and more models.
- Initially split their dataset into 2 — *Train and Test.*
- *Keep aside the Test set, and randomly choose X% of their Train dataset to be the actual Train set and the remaining (100-X)% to be the Validation set, where X is a fixed number(say 80%), the model is then iteratively trained and validated on these different sets.*

Types of ML



Supervised learning

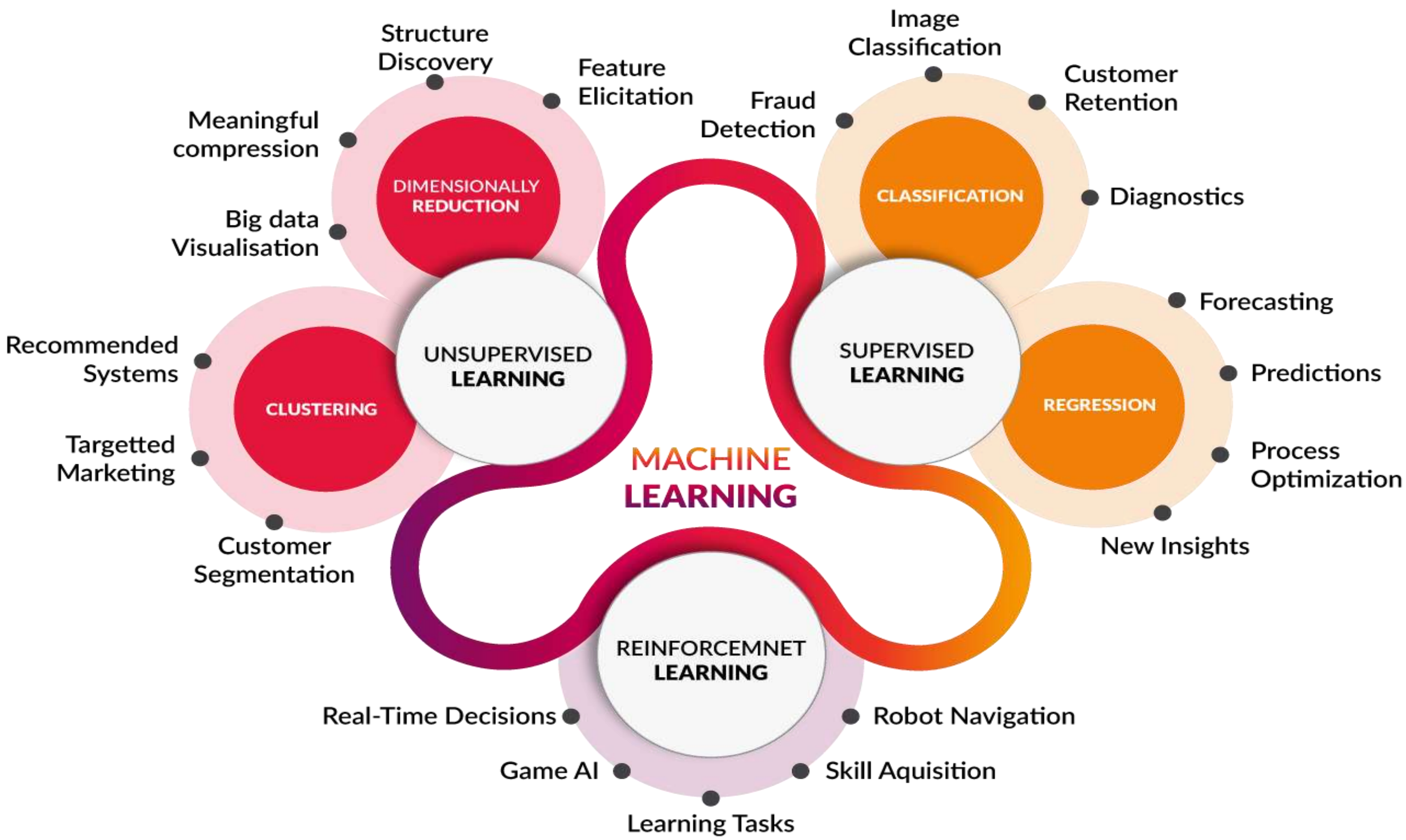
Reinforcement Learning

Types of Machine Learning Algorithms

Unsupervised Learning

Semi-supervised Learning

educba.com


MACHINE LEARNING
UNSUPERVISED LEARNING
DIMENSIONALLY REDUCTION
Structure Discovery
Feature Elicitation
Meaningful compression
Big data Visualisation
CLUSTERING
Recommended Systems
Targetted Marketing
Customer Segmentation
SUPERVISED LEARNING
CLASSIFICATION
Image Classification
Customer Retention
Fraud Detection
Diagnostics
REGRESSION
Forecasting
Predictions
Process Optimization
New Insights
REINFORCEMNET LEARNING
Real-Time Decisions
Robot Navigation
Game AI
Skill Aquisition
Learning Tasks

Supervised Learning

- It is the machine learning task of learning a function that maps an input to output based on example input-output pairs
- You train the machine using data which is well “labeled”
- Means some data is already tagged with correct answer
- It can be compared to learning which takes place in the presence of a supervisor or a teacher
- Basically learns from labeled trained data and helps to predict outcomes for unforeseen data

- Suppose you have a bunch of different kinds of flowers.

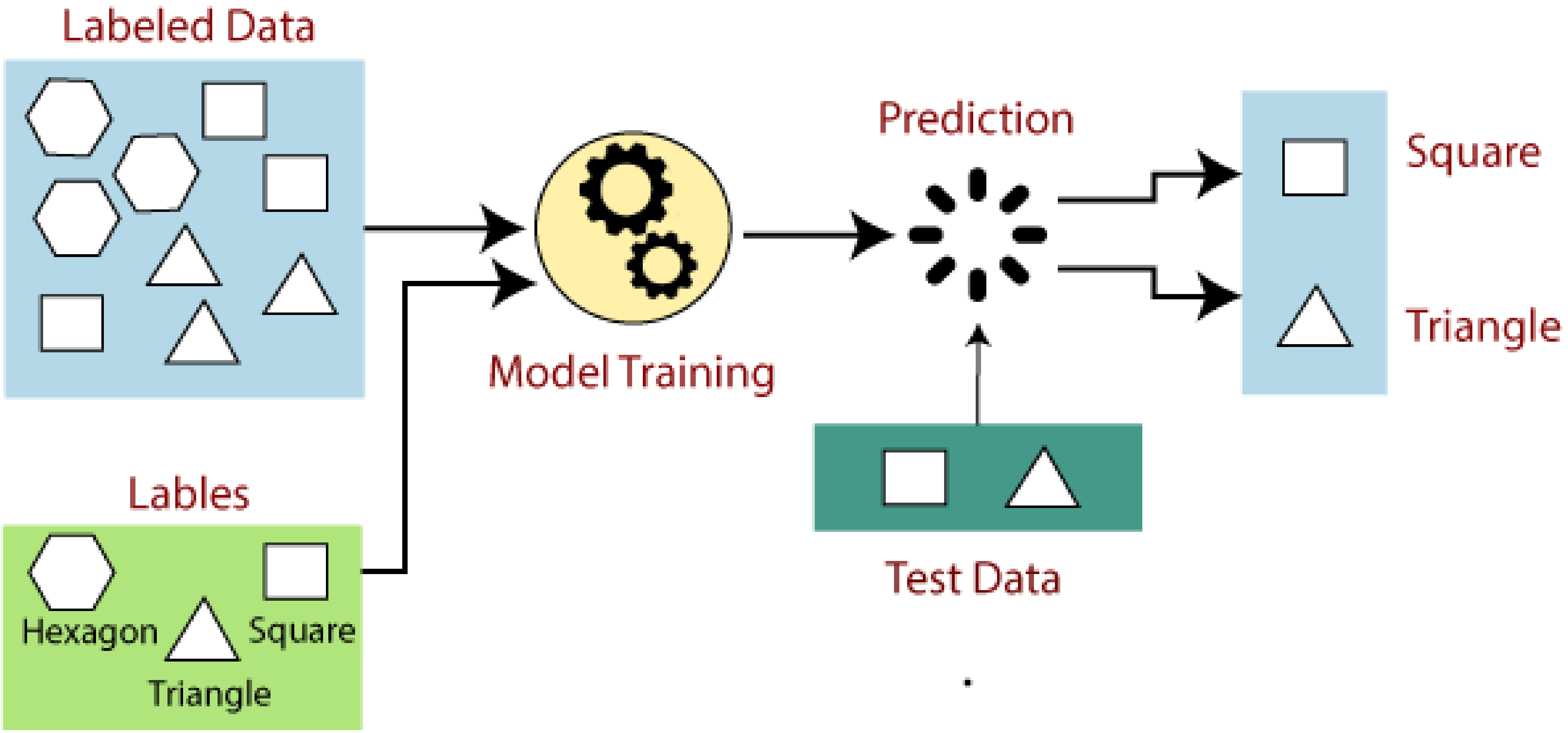
- First, you need to train the machine on how to classify all different flowers
- You can train it like this:
    - ➢ If there are thorns and the head has color Red then it will be labeled as Rose.
    - ➢ If there aren’t thorns and the head has color White then it will be labeled as Daisy.
- Above activities comes under which phase?

Test in supervised learning

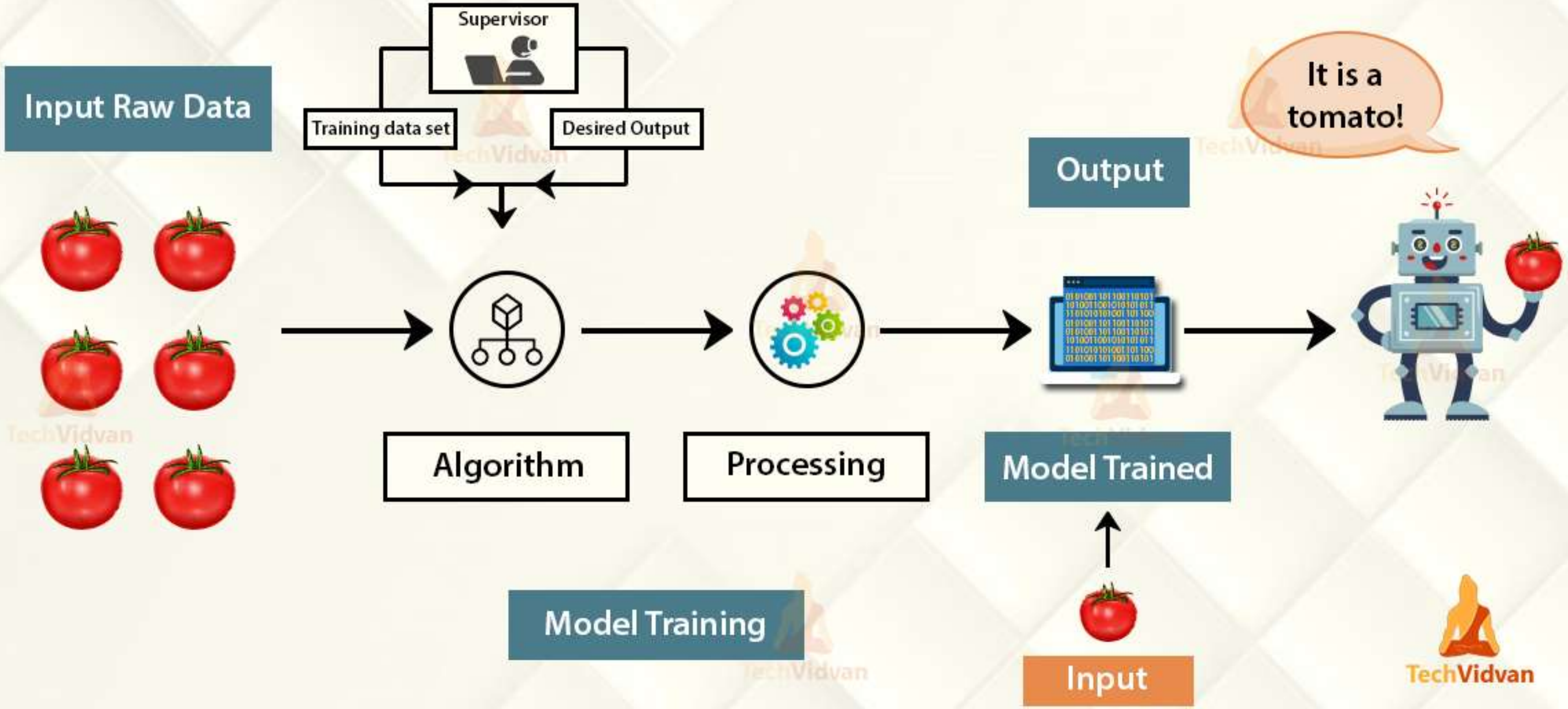
- Now, let’s say that after training the data, there is a new separate flower from the bunch and you need to ask the machine to identify it.

- Since your machine has already learned things, it needs to use that knowledge.
- The machine will classify the flower regarding the presence (or absence of thorns) and color and would label the flower name like Rose.
- This is how machines learn from training data (the bunch of flowers in our case) and then use the knowledge to label data.

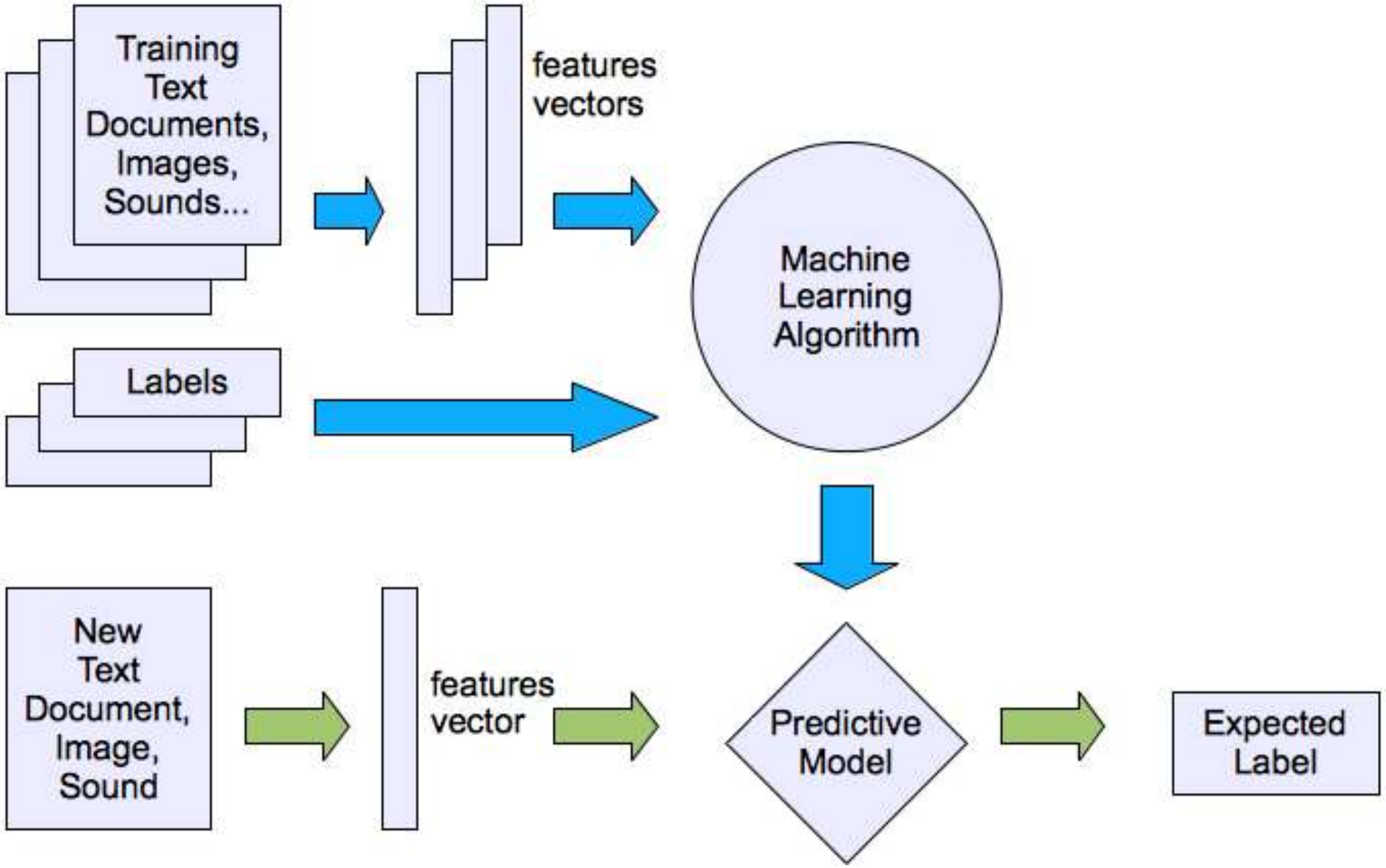
Supervised learning with shapes



Supervised Learning in ML



Supervised learning


Training Text Documents, Images, Sounds...
features vectors
Machine Learning Algorithm
Labels
New Text Document, Image, Sound
features vector
Predictive Model
Expected Label

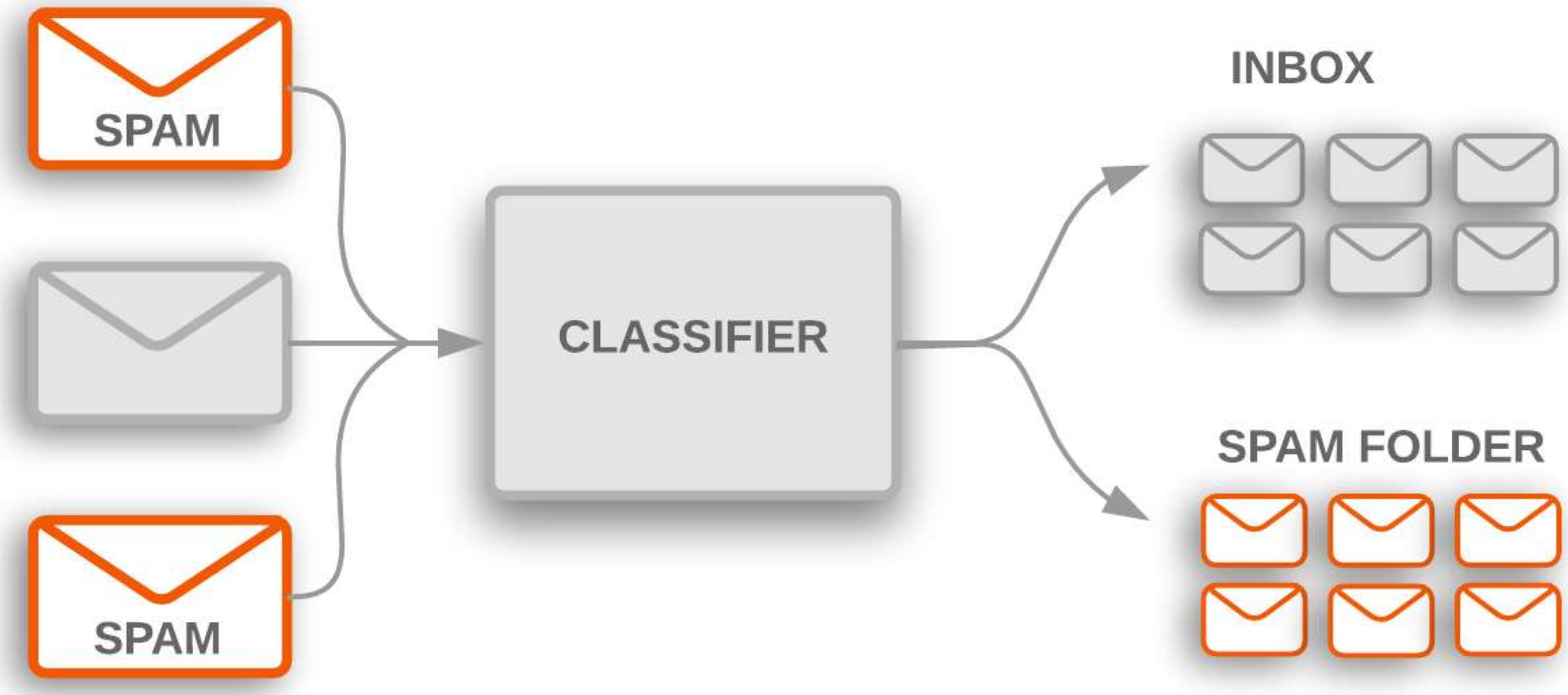
Classification vs Regression

- Supervised learning algorithm can be done in 2 ways:
- **Classification** :
- Classification is the task of predicting a discrete class label
- In classification problem, data is labeled into 1 or more classes
- **Regression**:
- Regression is the task of predicting a continuous quantity
- A regression problem requires the prediction of a quantity

Classification

- Predicts the data into built in labels
- 2 types of classification methods: binary and multi class
- Binary classification : predicts data into 2 category
- Ex: Whether student is pass or fail, email is spam or not, given fruit is apple or not
- Multi class classification : predicts the data to more than 2 classes
- Ex: Categorizing emails to personal, official, spam

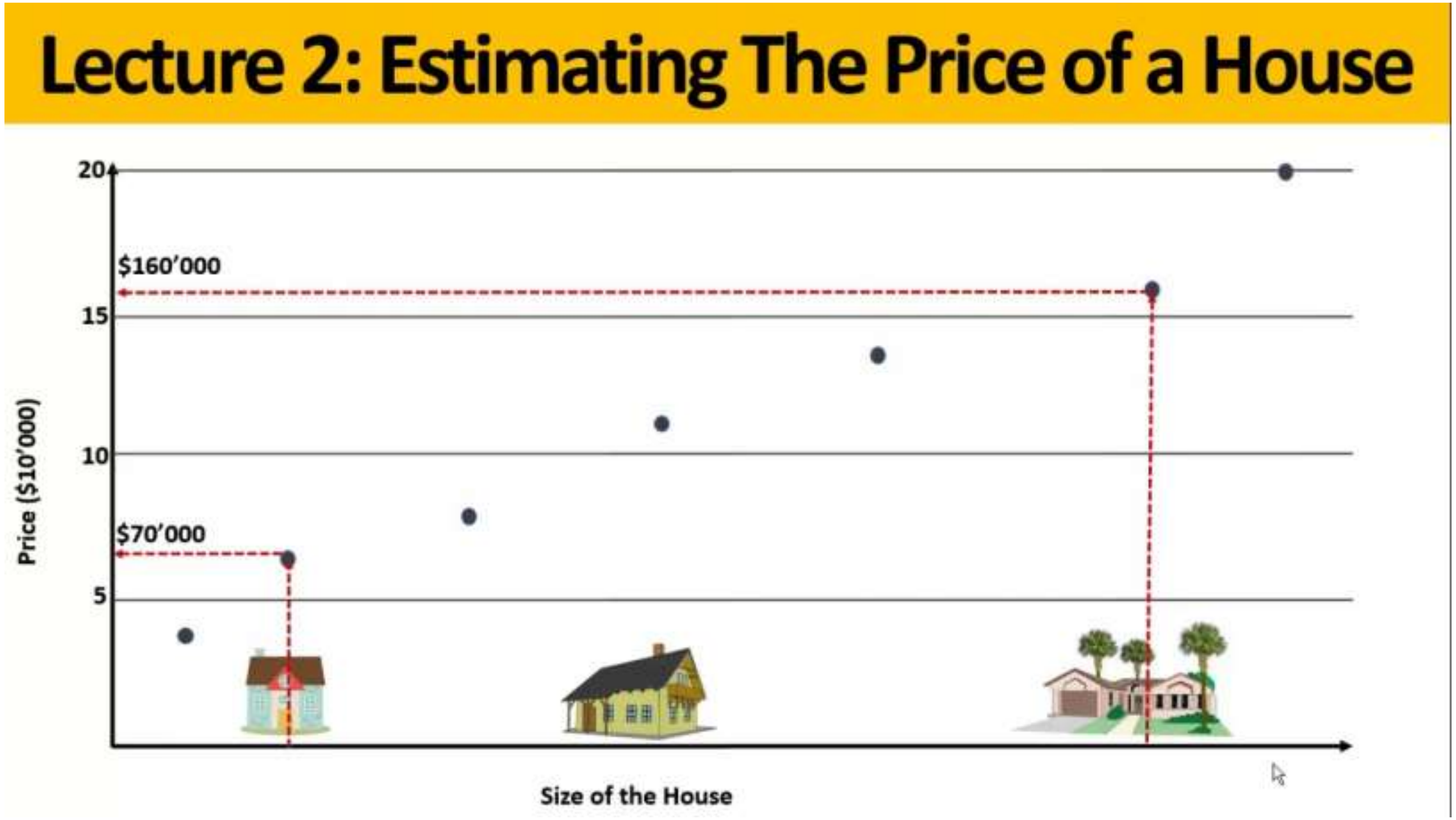
Classification



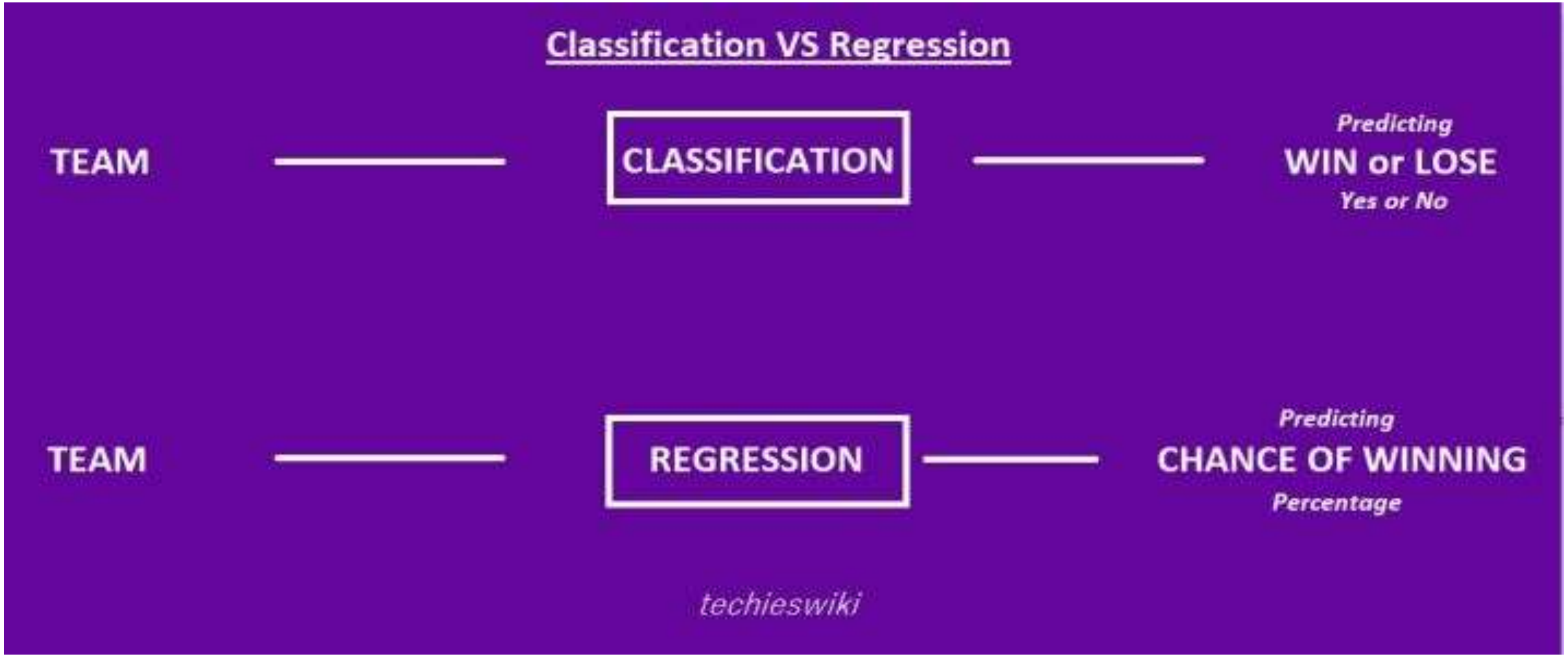
Regression

- Regression is a subfield of supervised learning
- Aims to model the relationship between a certain number of features and a continuous target variable
- Predicting prices of a house given the features of house like size, price etc is one of the common examples of Regression
- Predicting temperature of a region also comes under regression

Regression in Machine learning



Lecture 2: Estimating The Price of a House



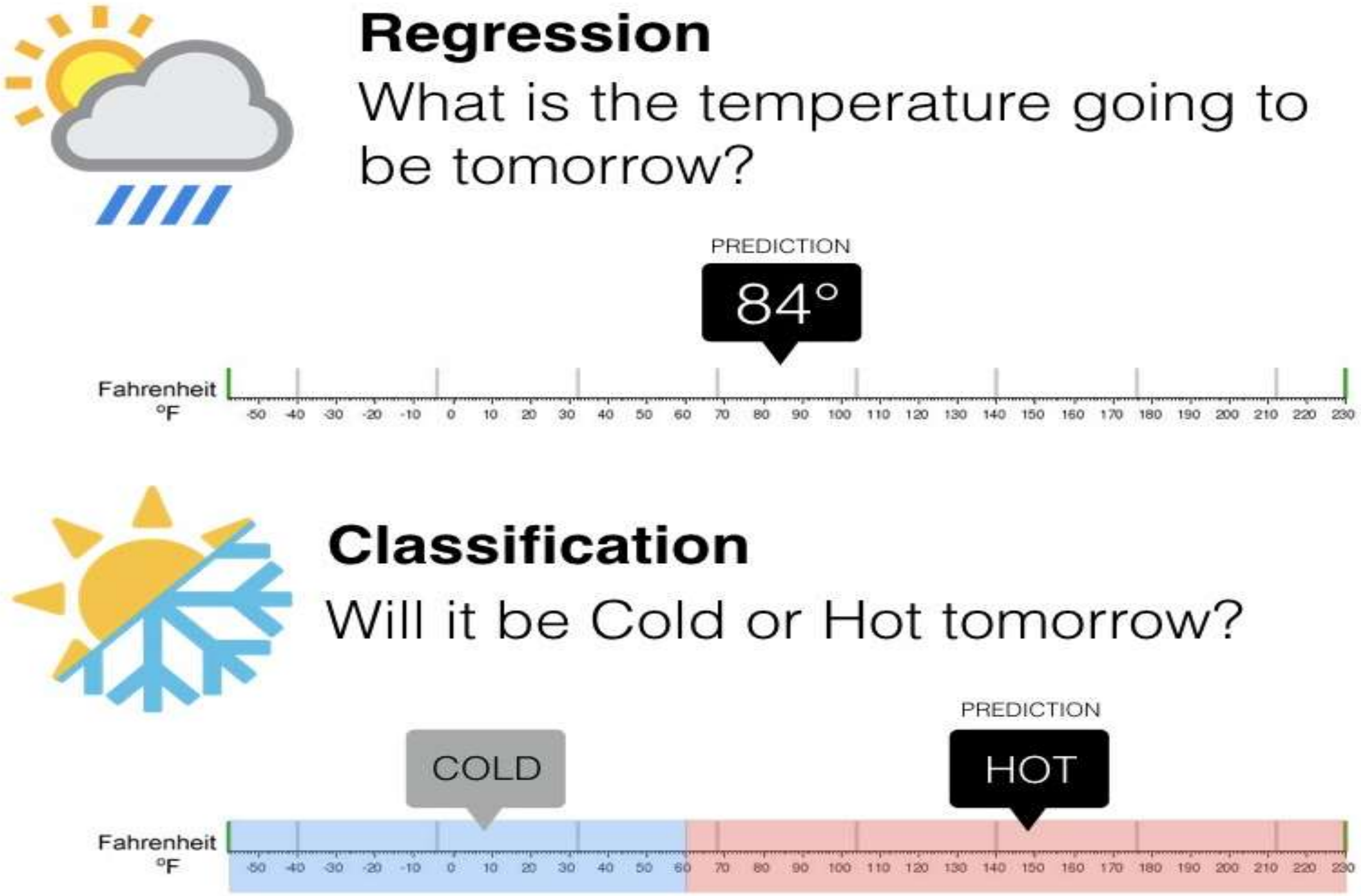
Classification VS Regression

TEAM — CLASSIFICATION — *Predicting* WIN or LOSE *Yes or No*

TEAM — REGRESSION — *Predicting* CHANCE OF WINNING *Percentage*

techieswiki

Classification Vs Regression



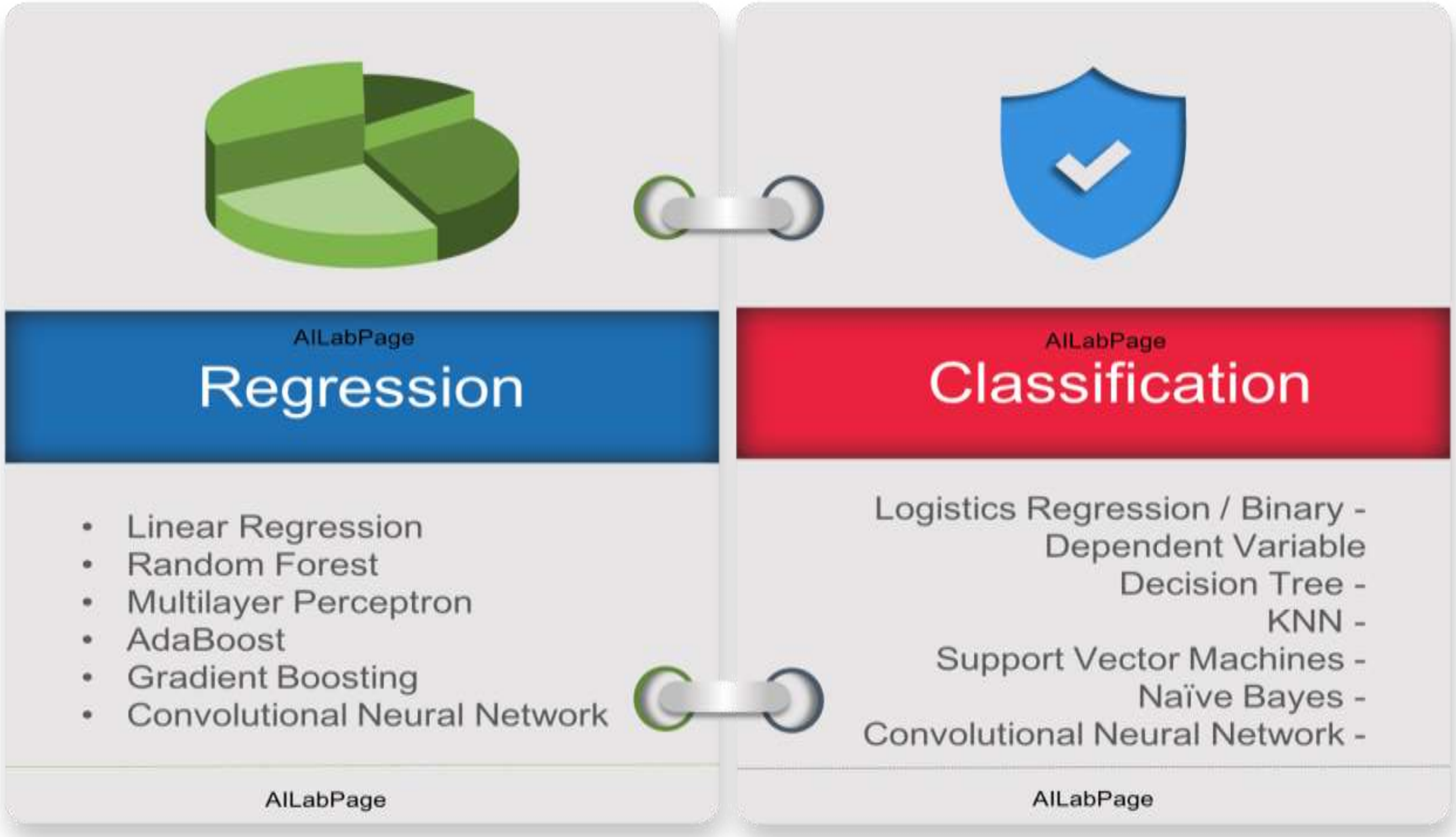
Regression

What is the temperature going to be tomorrow?

Classification

Will it be Cold or Hot tomorrow?

Major algorithms in supervised learning



Regression

- Linear Regression
- Random Forest
- Multilayer Perceptron
- AdaBoost
- Gradient Boosting
- Convolutional Neural Network

Classification

Logistics Regression / Binary - Dependent Variable

Decision Tree -

KNN -

Support Vector Machines -

Naïve Bayes -

Convolutional Neural Network -

Application of supervised learning

- Value modeling
- Object-Recognition
- Dynamic pricing
- Recommendation engines

Pros & Cons

- Pros of supervised learning:
  - ➢ Allows to collect data or produce a data output from the previous experience
  - ➢ Optimize performance criteria using experience
  - ➢ Solve various types of real-world computation problems.
- Disadvantages of Supervised Learning:
  - ➢ Decision boundary might be over trained if the training set which doesn't have examples that you want to have in a class
  - ➢ We need to select lots of good examples from each class while training the classifier.
  - ➢ Classifying big data can be a real challenge.
  - ➢ Training for supervised learning needs a lot of computation time.

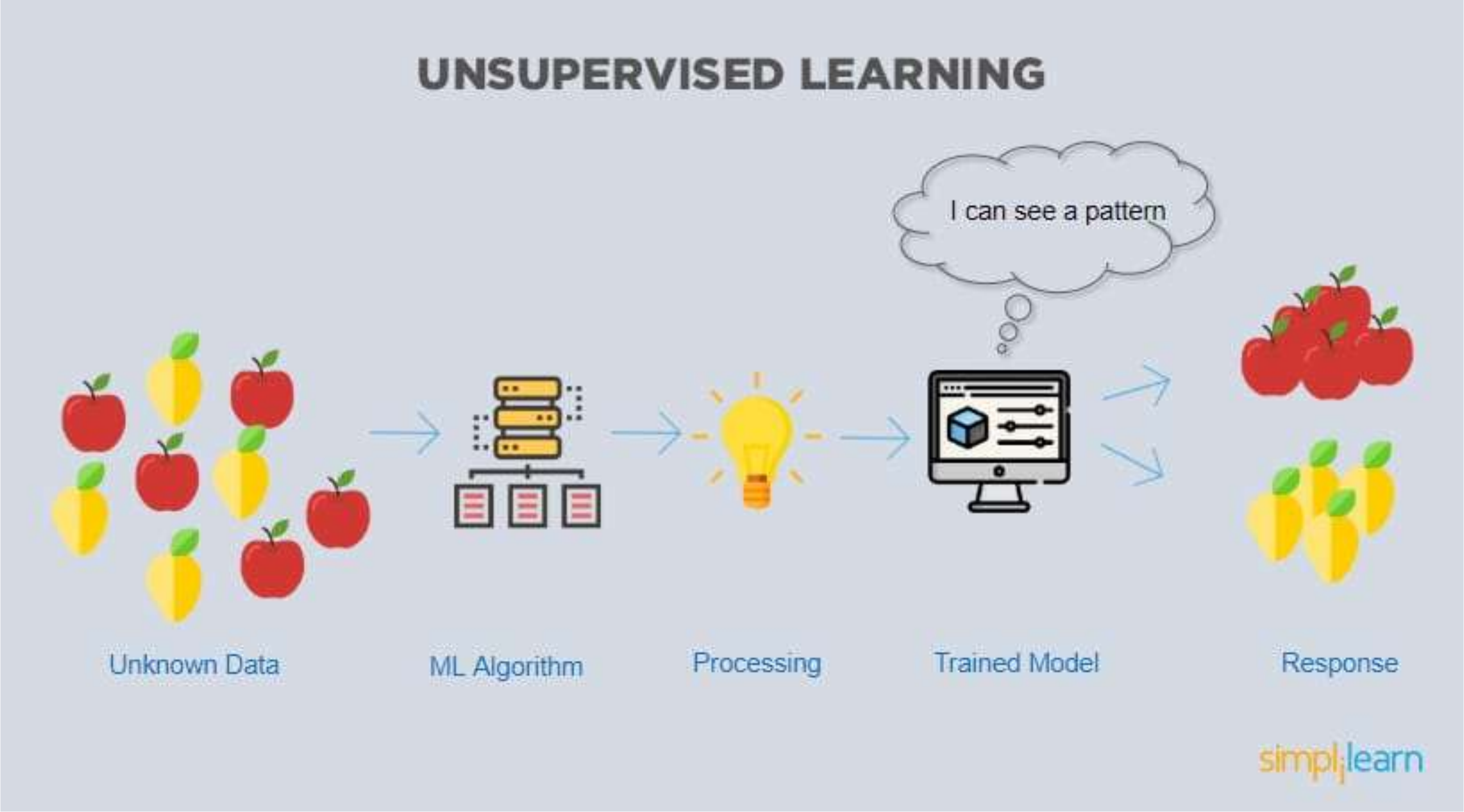
Unsupervised Learning

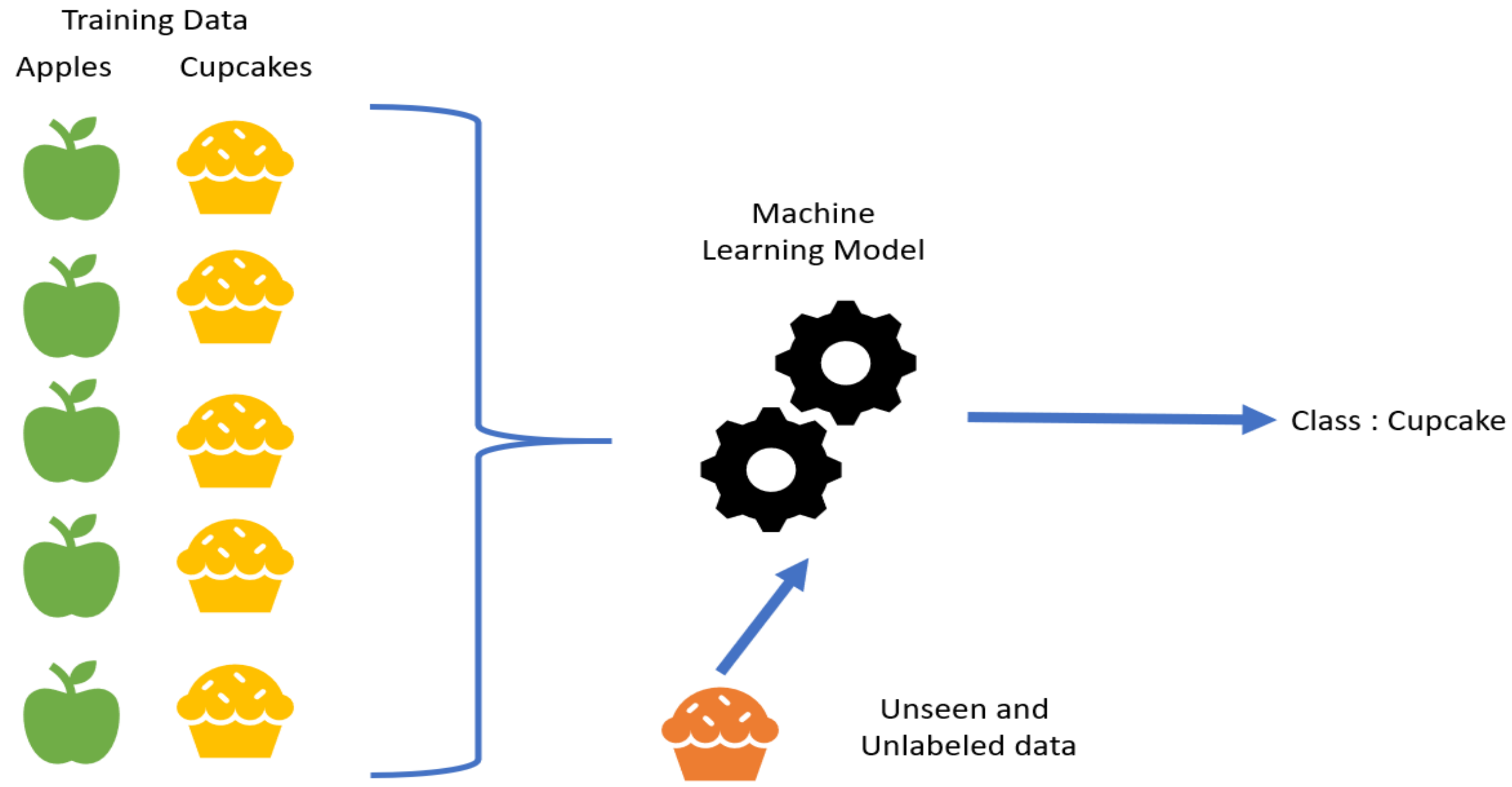
- Unsupervised learning is a type of machine learning that looks for previously undetected patterns in a data set with no pre-existing labels and with a minimum of human supervision
- Instead, it allows the model to work on its own to discover patterns and information that was previously undetected
- Here the task of the machine is to group the unsorted information according to similarities, patterns, and differences without any prior training of data.
- Unlike supervised learning, no teacher is provided that means no training will be given to the machine.
- Therefore machine is restricted to find the hidden structure in unlabeled data by our-self.

- Suppose it is given an image having both cats and dogs which have not seen ever.

- This machine has no idea about the features of dogs and cats so we can't categorize it in cats and dogs.
- But it can categorize them according to their similarities, pattern and differences
- We can easily categorize the above picture into two parts.
- The first part may contain all pics having **DOGS** in it and the second part may contain all pics having **CATS** in it.
- Here you didn't learn anything before means no training data or examples.

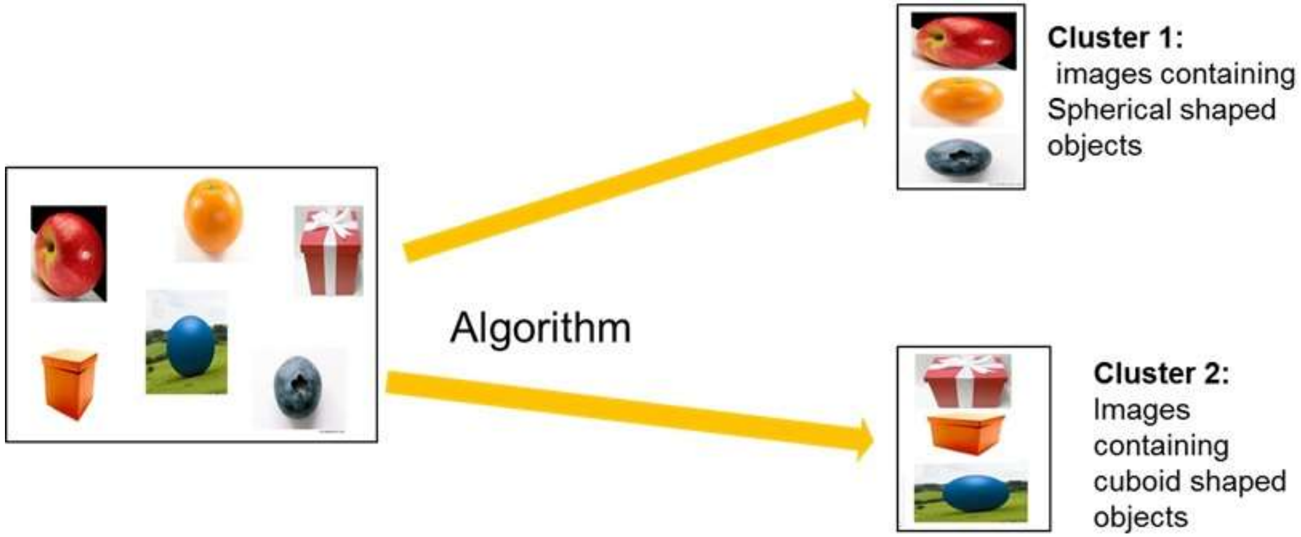
Unsupervised learning




Training Data
Apples
Cupcakes
Machine
Learning Model
Class : Cupcake
Unseen and
Unlabeled data

Clustering

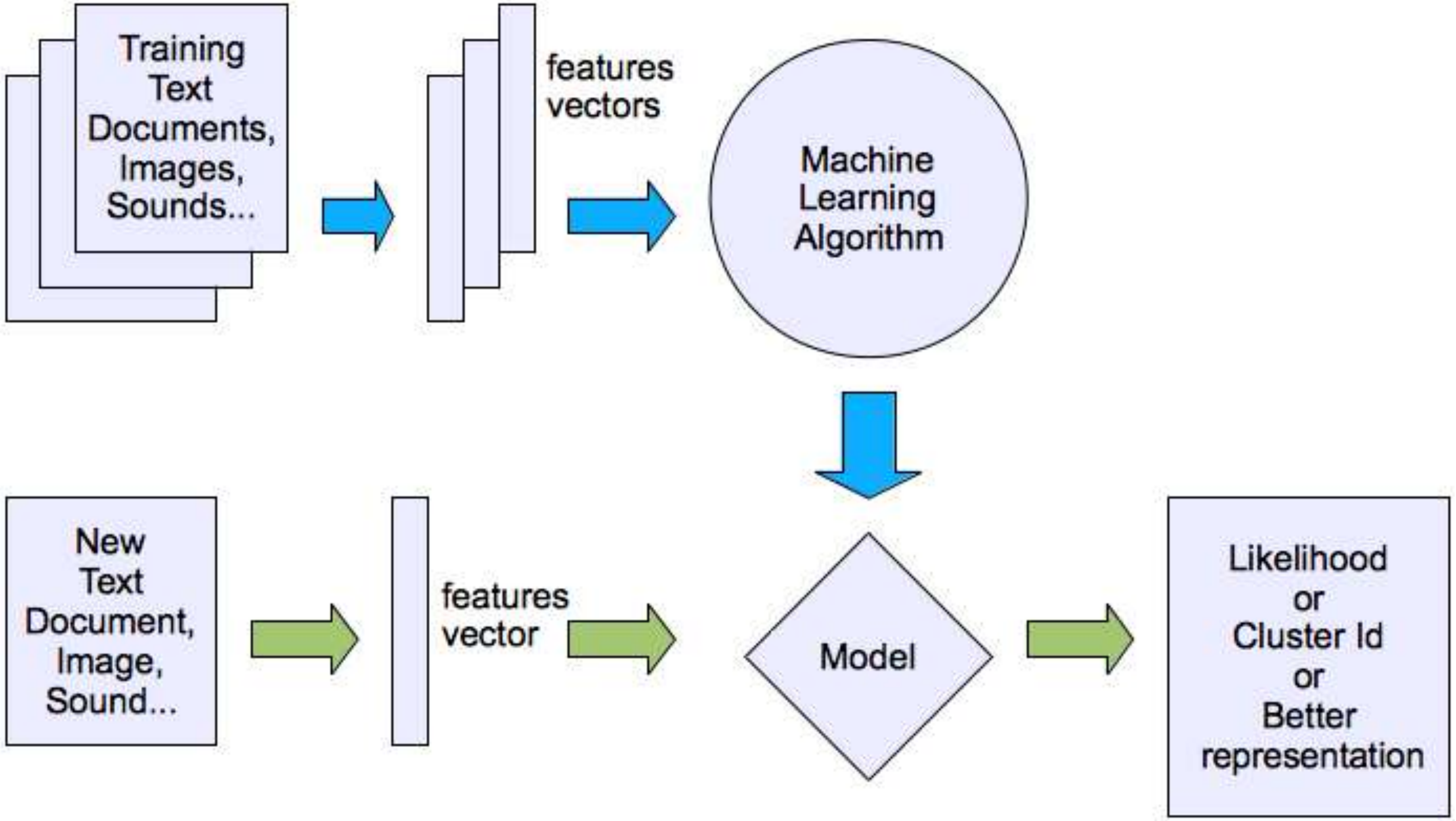
- **Clustering:** is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.
- It is basically a collection of objects on the basis of similarity and dissimilarity between them.

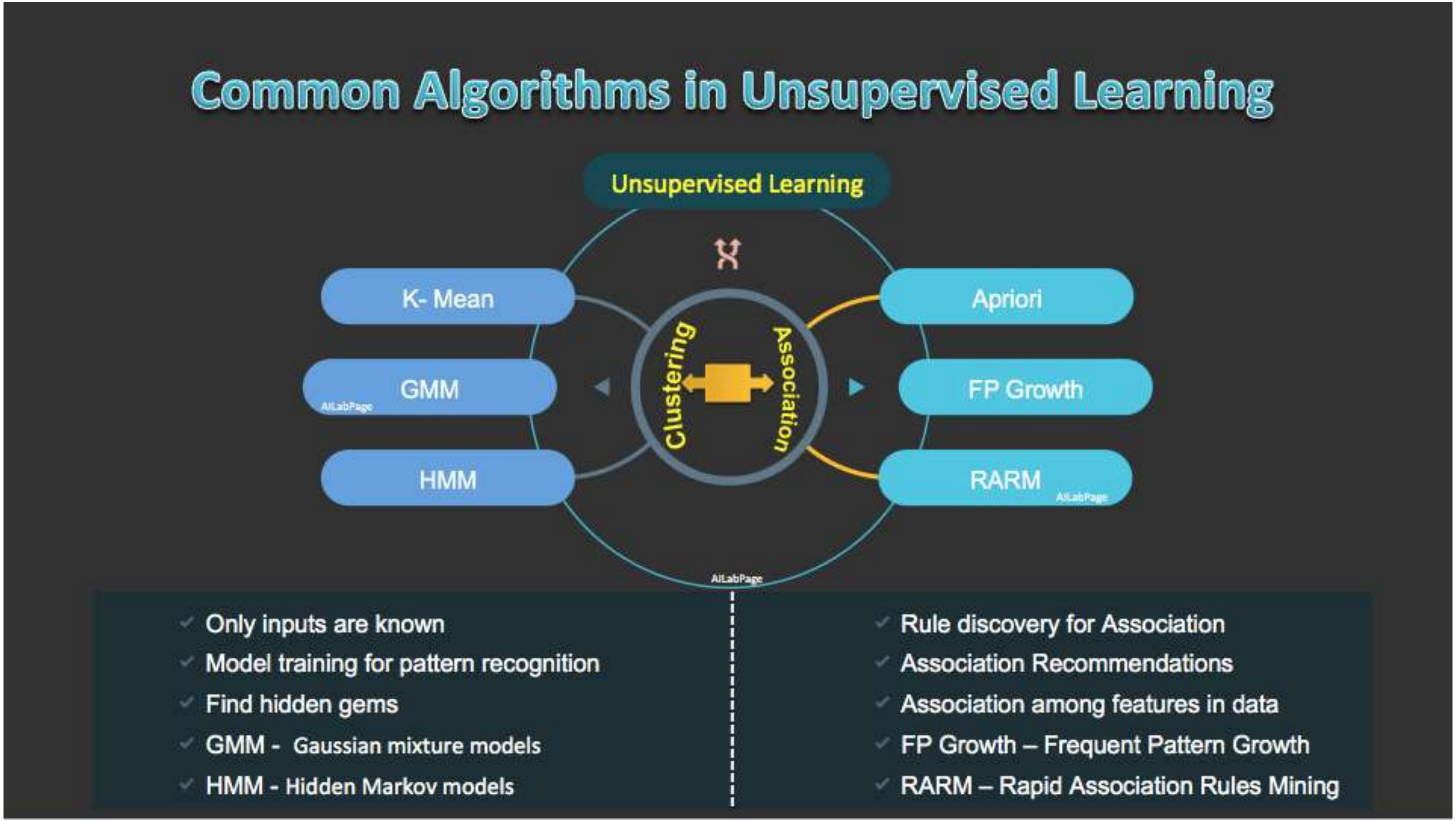


Association

- **Association rule learning** is a rule-based machine learning method for discovering interesting relations between variables in large databases.
- It is intended to identify strong rules discovered in databases using some measures of interestingness
- Ex: if a customer buys the bread and milk then he will buy the cheese. If a customer buys a mobile phone then he will buy a back case also.

Unsupervised learning


Training Text Documents, Images, Sounds...
features vectors
Machine Learning Algorithm
New Text Document, Image, Sound...
features vector
Model
Likelihood or Cluster Id or Better representation



Common Algorithms in Unsupervised Learning

Unsupervised Learning

Clustering

- K- Mean
- GMM
- HMM

Association

- Apriori
- FP Growth
- RARM

AILabPage

- ✓ Only inputs are known
- ✓ Model training for pattern recognition
- ✓ Find hidden gems
- ✓ GMM - Gaussian mixture models
- ✓ HMM - Hidden Markov models

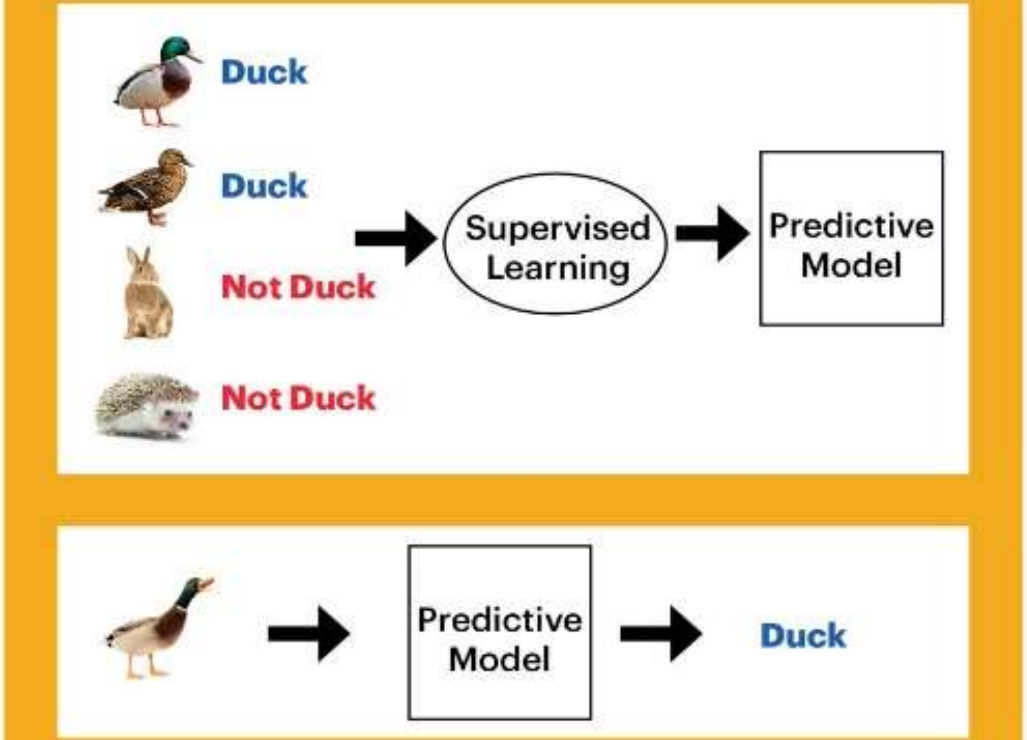
- ✓ Rule discovery for Association
- ✓ Association Recommendations
- ✓ Association among features in data
- ✓ FP Growth – Frequent Pattern Growth
- ✓ RARM – Rapid Association Rules Mining

Pros & Cons

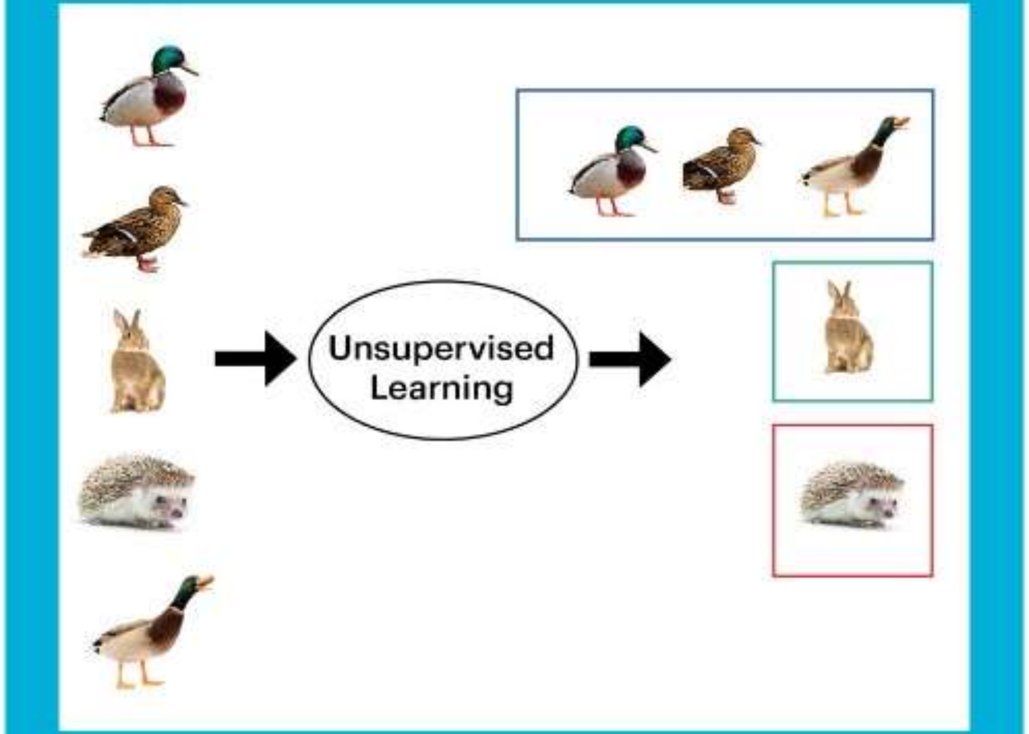
- Pros
  - It can detect what human eyes can not understand
  - The potential of hidden patterns can be very powerful for the business or even detect extremely amazing facts, fraud detection etc.
  - Output can determine the un explored territories and new ventures for businesses.

- Cons
  - Harder than supervised learning.
  - It can be a costly affair, as we might need external expert look at the results for some time.
  - Usefulness of the results; are of any value or not is difficult to confirm since no answer labels are available.

Supervised Learning (Classification Algorithm)



Unsupervised Learning (Clustering Algorithm)



Western Digital.

Applications : Unsupervised Learning

- Bioinformatics: sequence analysis & genetic clustering
- Data mining : sequence & pattern mining
- Medical imaging : image segmentation

Semi-supervised Learning

- **Semi-supervised machine learning** is a combination of **supervised** and unsupervised **machine** learning methods.
- In semi-supervised learning, an algorithm learns from a dataset that includes both labeled and unlabeled data, usually mostly unlabeled.
- When you don’t have enough labeled data to produce an accurate model and you don’t have the ability or resources to get more data, you can use semi-supervised techniques to increase the size of your training data
- First the programmer will cluster similar data using an unsupervised learning algorithm and then use the existing labeled data to label the rest of the unlabeled data.

What is semi-supervised clustering?

- Cluster analysis is a method that seeks to partition a dataset into homogenous subgroups, meaning grouping similar data together with the data in each group being different from the other groups.
- Clustering is conventionally done using unsupervised methods.
- Since the goal is to identify similarities and differences between data points, it doesn’t require any given information about the relationships within the data.
- However, there are situations where some of the cluster labels, outcome variables, or information about relationships within the data are known.
- This is where semi-supervised clustering comes in.
- Semi supervised clustering uses some known cluster information in order to classify other unlabeled data, meaning it uses both labeled and unlabeled data

Pseudo Labeling

- Pseudo Labeling is a simple and an efficient method to do semi-supervised learning.
- It can combine almost all neural network models and training methods (Pseudo-Label).
- Here is an example of the steps to follow if you want to learn from your unlabeled data too:
  - Take the same model that you used with your training set and that gave you good results.
  - Use it now with your unlabeled test set to predict the outputs (or pseudo-labels).
  - We don’t know if these predictions are correct, but we do now have quite accurate labels and that’s what we aim in this step.
  - Concatenate the training labels with the test set pseudo labels.
  - Concatenate the features of the training set with the features of the test set.
  - Finally, train the model in the same way you did before with the training set.
- This method will make the error decreases and it will improve the model by better learning the general structure.

Pseudo Labeling

- **But, how can I know the proportion of true labels and pseudo-labels in each batch**?
- In other words, how much do I make it a mix of training vs pseudo?
- The general rule of thumb is to have 1/4–1/3 of your batches be pseudo-labeled.

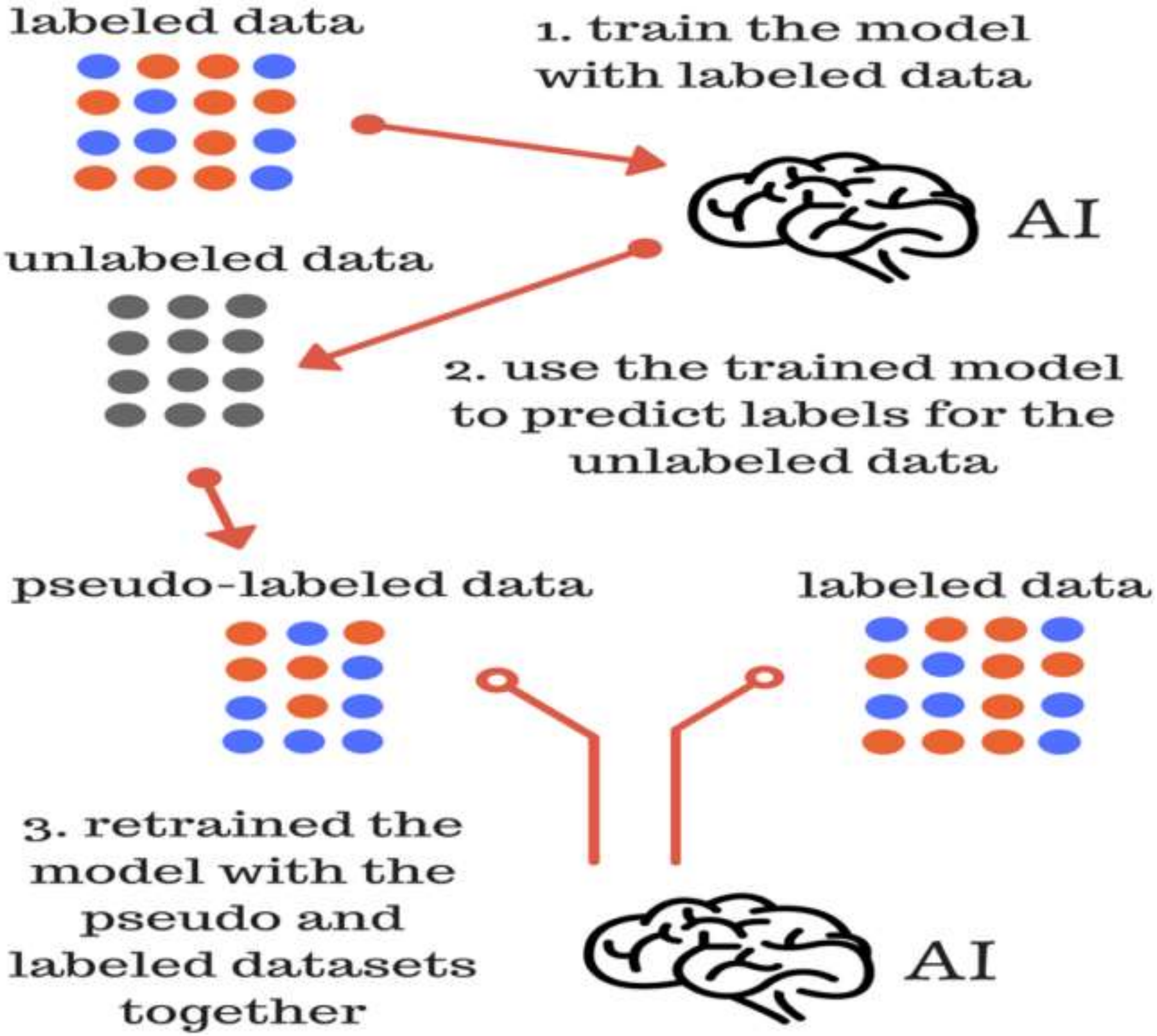
1.Train the model with small amount of labeled training data like supervised learning until it gives good results

2.Then use it with the unlabeled training dataset to predict the outputs, which are pseudo labels since they may not be quite accurate

3.Link the labels from the labeled training data with the pseudo labels created in the previous step

4.Link the data inputs in the labeled training data with the inputs in the unlabeled data

5.Then,train the model the same way as we did with the labeled set in the beginning in order to decrease the error and improve the model’s accuracy



labeled data

1. train the model with labeled data

AI

unlabeled data

2. use the trained model to predict labels for the unlabeled data

pseudo-labeled data

labeled data

3. retrained the model with the pseudo and labeled datasets together

AI

Example application of semi-supervised learning

- A common example of an application of semi-supervised learning is a text document classifier.
- This is the type of situation where semi-supervised learning is ideal because it would be nearly impossible to find a large amount of labeled text documents.
- This is simply because it is not time efficient to have a person read through entire text documents just to assign it a simple classification.
- So, semi-supervised learning allows for the algorithm to learn from a small amount of labeled text documents while still classifying a large amount of unlabeled text documents in the training data.

Applications

- Speech analysis
- Internet content classification(Even the Google search algorithm uses a variant of Semi-Supervised learning to rank the relevance of a webpage for a given query.)
- Protein sequence classification

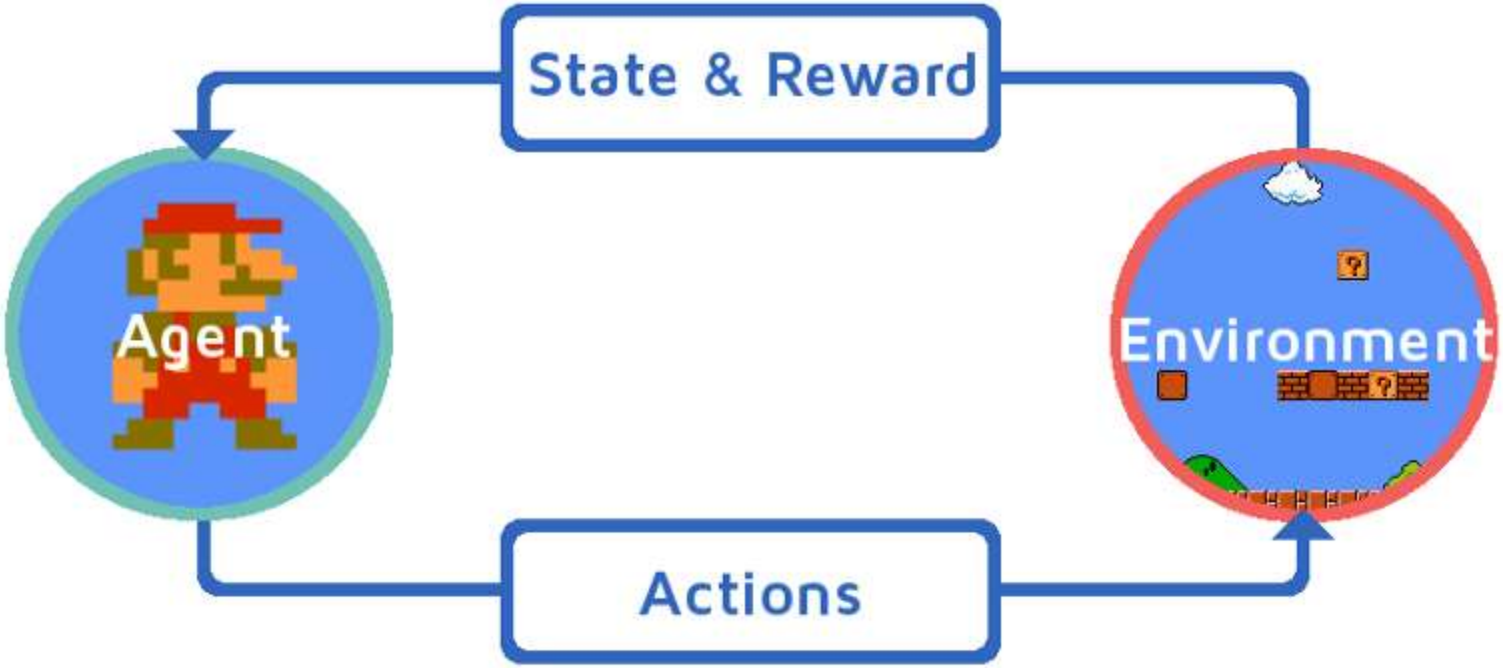
Reinforcement Learning

- Reinforcement learning is the training of machine learning models to make a sequence of decisions.
- Reinforcement learning is all about making decisions sequentially.
- Reinforcement Learning(RL) is a type of machine learning technique that enables an agent to learn in an interactive environment by trial and error using feedback from its own actions and experiences.
- Output depends on the state of the current input and the next input depends on the output of the previous input
- And decision is dependent, so we give labels to sequences of dependent decisions
- Ex: Chess game

Reinforcement, Supervised & Unsupervised

- Reinforcement & supervised:
  - ➢ Both use mapping between input and output
  - ➢ In supervised learning feedback provided to the agent is correct set of actions for performing a task
  - ➢ Whereas in reinforcement learning, it uses rewards and punishment as signals for positive and negative behavior.
- Reinforcement & Unsupervised:
  - ➢ Goal of unsupervised learning is to find similarities and differences between data points
  - ➢ Whereas goal of reinforcement is to find a suitable action model that would maximize the total cumulative reward of the agent

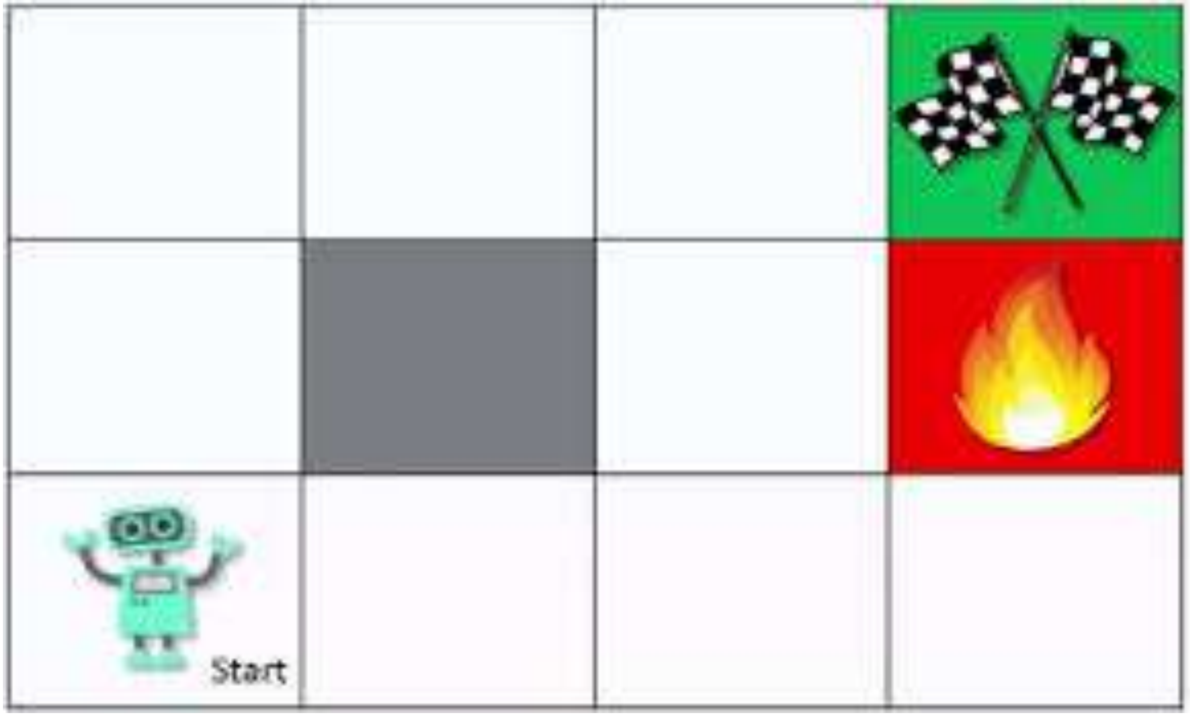
Agent, Actions, Reward



Example for reinforcement learning

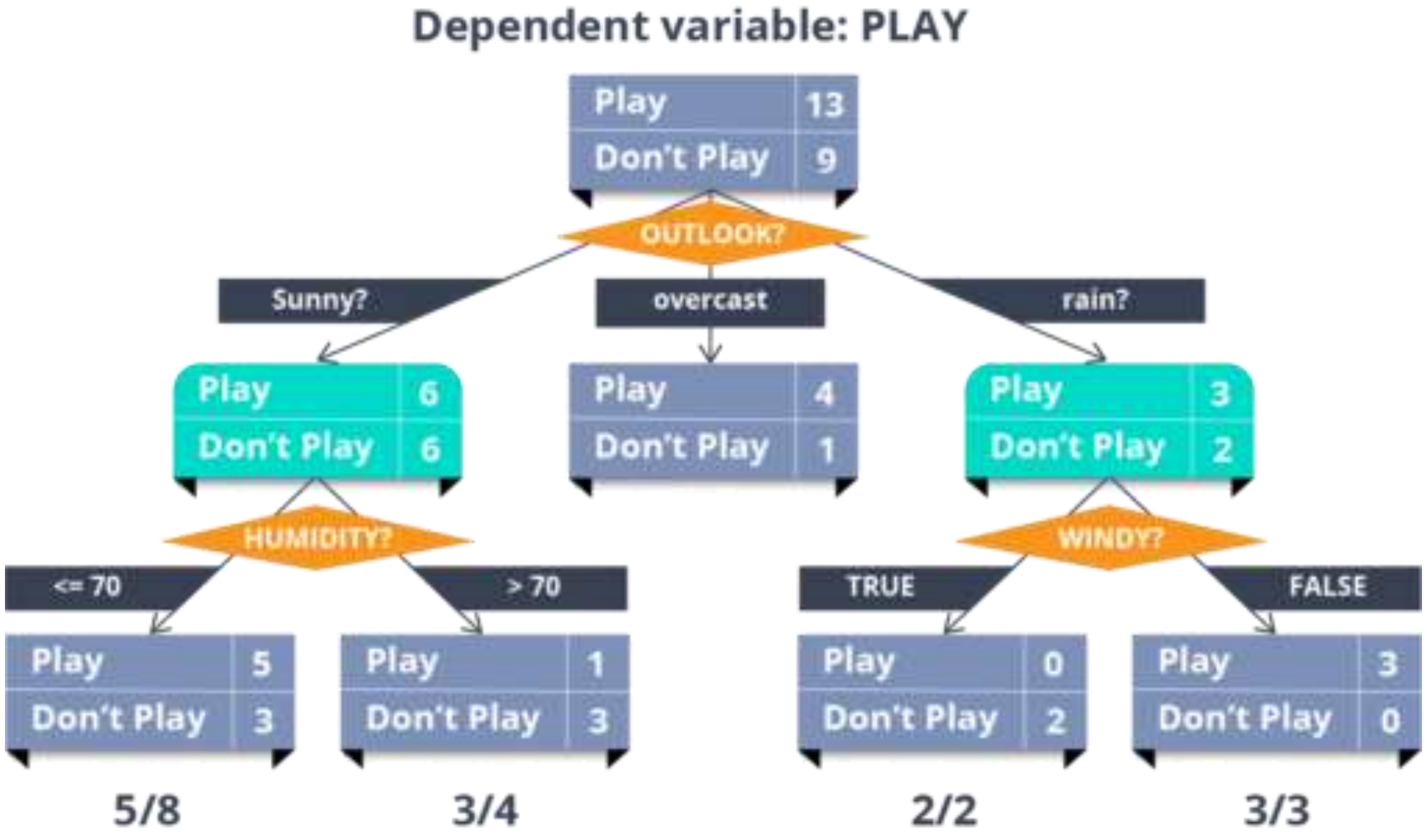
- In usual circumstances we would require an autonomous vehicle to put safety first, minimize ride time, reduce pollution, offer passengers comfort and obey the rules of law.
- With an autonomous race car, on the other hand, we would emphasize speed much more than the driver’s comfort.
- The programmer cannot predict everything that could happen on the road.
- Instead of building lengthy “if-then” instructions, the programmer prepares the reinforcement learning agent to be capable of learning from the system of rewards and penalties.
- The agent (another name for reinforcement learning algorithms performing the task) gets rewards for reaching specific goals.

- Scenario is :We have an agent and a reward, with some obstacles in between. The agent goal is to find the best possible path to get the reward. Let’s explain the scenario with an example.



- The above image shows robot, flag and fire.
- Here agent is robot, environment is the play area(including fire) and flag is the reward
- The goal of the robot is to get flag and avoid the obstacle, fire.
- The robot learns by trying all the possible paths and then choosing the path which gives him the reward with the least hurdles.
- Each right step will give the robot a reward and each wrong step will subtract the reward of the robot.
- The total reward will be calculated when it reaches the final reward that is the flag.

Machine learning algorithms



- In the image above, you can see that population is classified into four different groups based on multiple attributes to identify ‘if they will play or not’.

References

- https://www.geeksforgeeks.org/ml-machine-learning/
- https://en.wikipedia.org/wiki/Machine_learning
- https://www.kdnuggets.com/2018/05/general-approaches-machine-learning-process.html
- https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7
- https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/
- https://www.guru99.com/supervised-machine-learning.html
- https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292
- https://towardsdatascience.com/simple-explanation-of-semi-supervised-learning-and-pseudo-labeling-c2218e8c769b