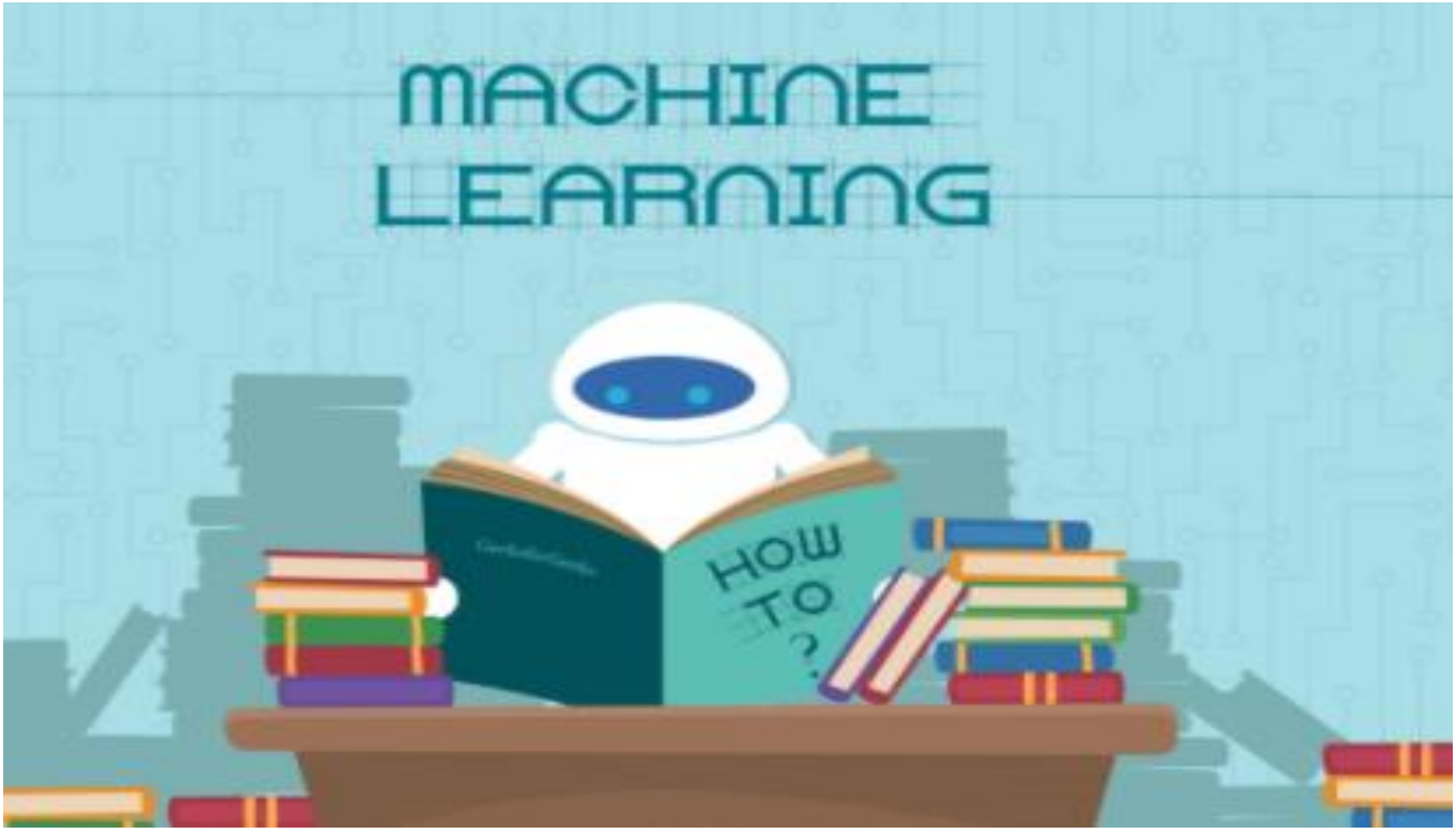
Machine Learning

Unit-1

Prepared by Dr.Y.UshaRani


MACHINE
LEARNING
HOW
TO
?

Before ,We start

- Sewing Machine
- Typewriting Machine
- Washing Machine
- Computer machine
- Like wise list me few things with machine…
- …….
- What’s difference .....?????

About the Course

- Course project is optional
- Homework can be submitted in pairs

Course Objectives

- To introduce students to the basic concepts and techniques of Machine Learning.
- To have a thorough understanding of the Supervised and Unsupervised learning techniques
- To study the various probability-based and Generalized learning techniques
- To understand ensemble models of machine learning algorithms

Course Outcomes

- Distinguish between, supervised, unsupervised and semi-supervised learning.
- Apply the appropriate machine learning strategy for any given problem.
- Ability to get the skill to apply machine learning techniques to address the real time problems in different areas.
- Modify existing machine learning algorithms to improve classification efficiency.

Syllabus at Abstract Level

- **UNIT-1:**Introduction: Introduction to Machine Learning
- **UNIT-2**: Decision tree learning
- **UNIT-3:** Instance based Learning: KNN,..
- **UNIT-4:**Probability and Baye's Learning
- **UNIT-5:**Artificial Neural Networks
- **UNIT-6**: Evaluating Machine Learning algorithms and Ensemble LearningModel Selection,

Unit -1 Syllabus :Introduction

Introduction to Machine Learning: Introduction

Different types of learning,

Perspectives and Issues in Machine Learning,

Parametric and Nonparametric Machine Learning Algorithms,

Hypothesis space and inductive bias,

Evaluation

Training and test sets,

cross validation,

Linear Regression: Introduction

Linear Models for Regression.

Pre-requisite

- Probability
  - distribution, random variable, expectation, conditional probability, variance, density
- Linear algebra
  - matrix multiplication
  - eigenvector
- Basic programming (in Python)

Do's

Form groups to discuss and work on Course based Projects

- Form a team for Home assignments and problems in groups
- write down the solutions independently
- write down the names of people with whom you've discussed the homework

Don'ts

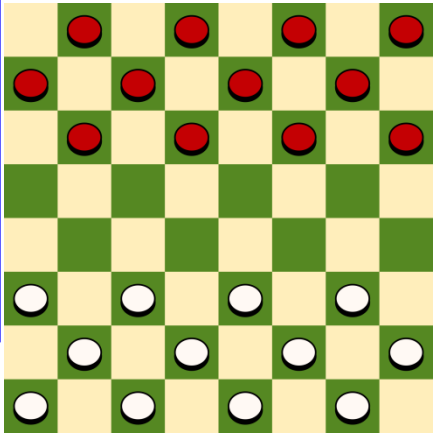
- Copy,
- But discuss
- Refer to, or look at any official or unofficial previous years' solutions in preparing the answers

Definition of Machine Learning

Arthur Samuel (1959): Machine Learning is the field of study that gives the computer the ability to learn without being explicitly programmed.

A. L. Samuel*

Some Studies in Machine Learning
Using the Game of Checkers. II—Recent Progress



Photos from Wikipedia

Definition of Machine Learning

Tom Mitchell (1998): 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 in T, as measured by P, improves with experience E.

Experience (data): games played by the program (with itself)

Performance measure: winning rate

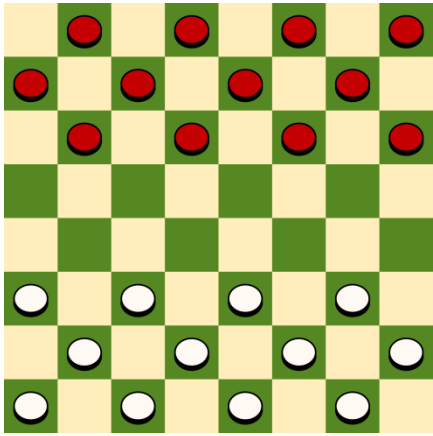
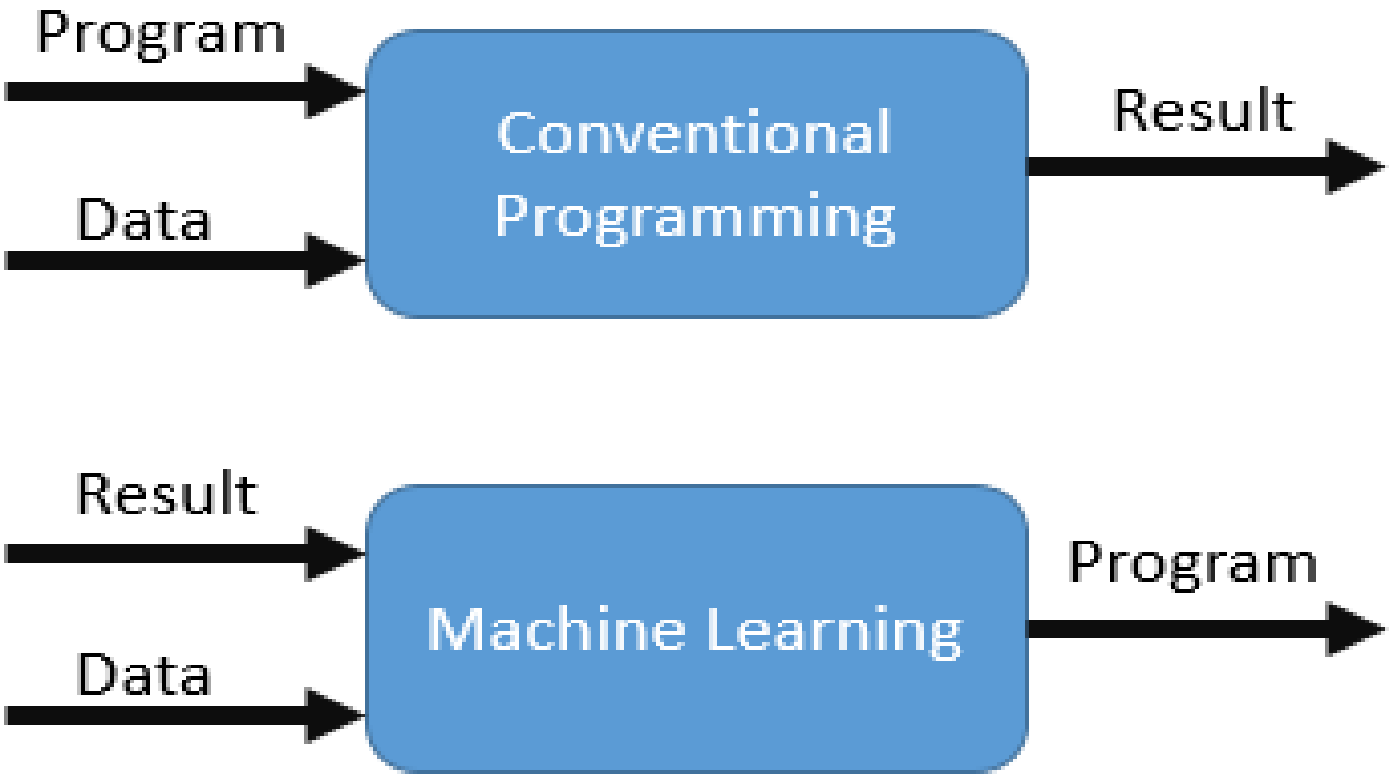


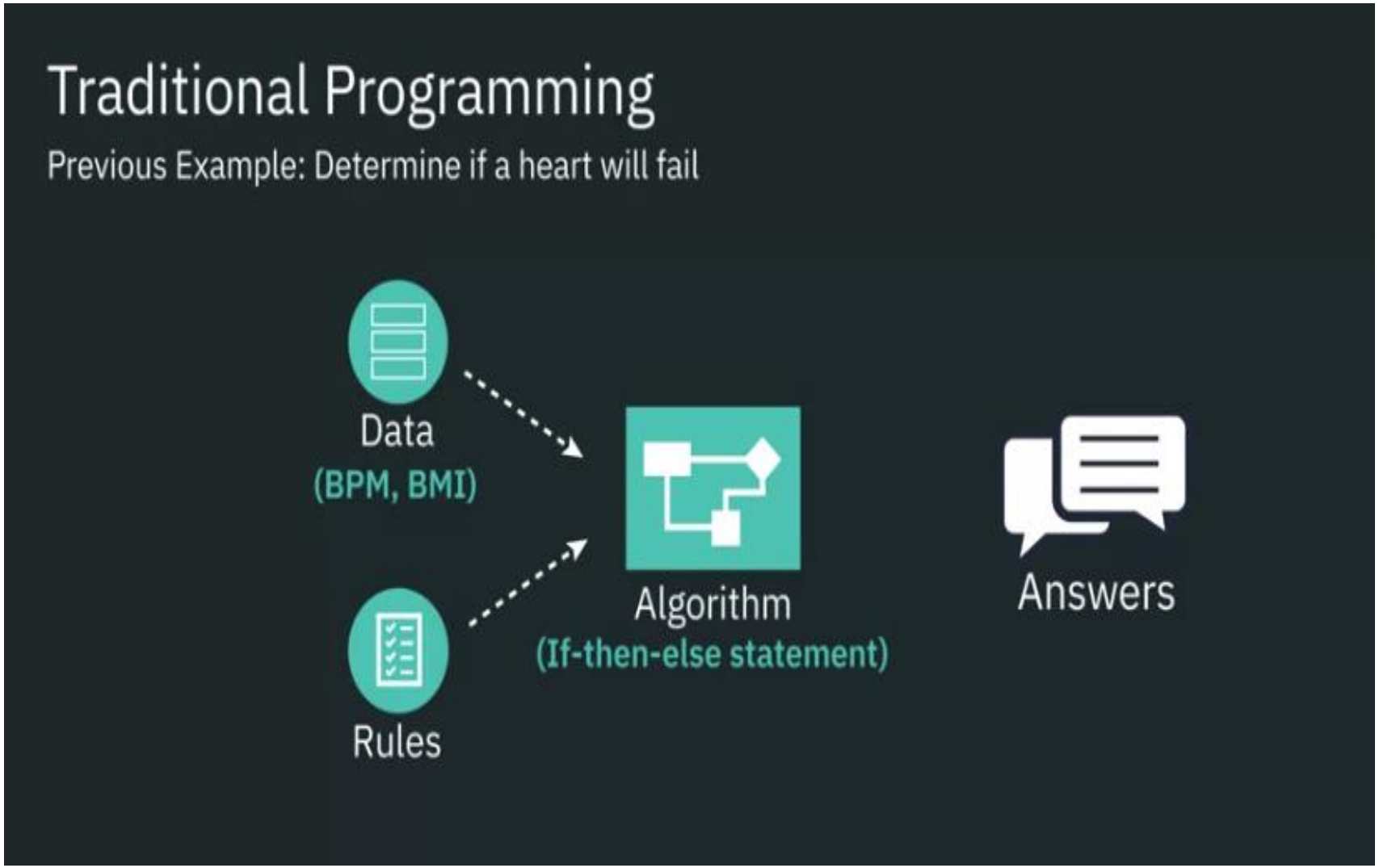
Image from Tom Mitchell's homepage

- **Machine Learning** is the field of study that gives computers the capability to learn without being explicitly programmed.
- ML is one of the most exciting technologies that one would have ever come across.
- As it is evident from the name, it gives the computer that makes it more similar to humans:
- ***The ability to learn***. Machine learning is actively being used today, perhaps in many more places than one would expect.

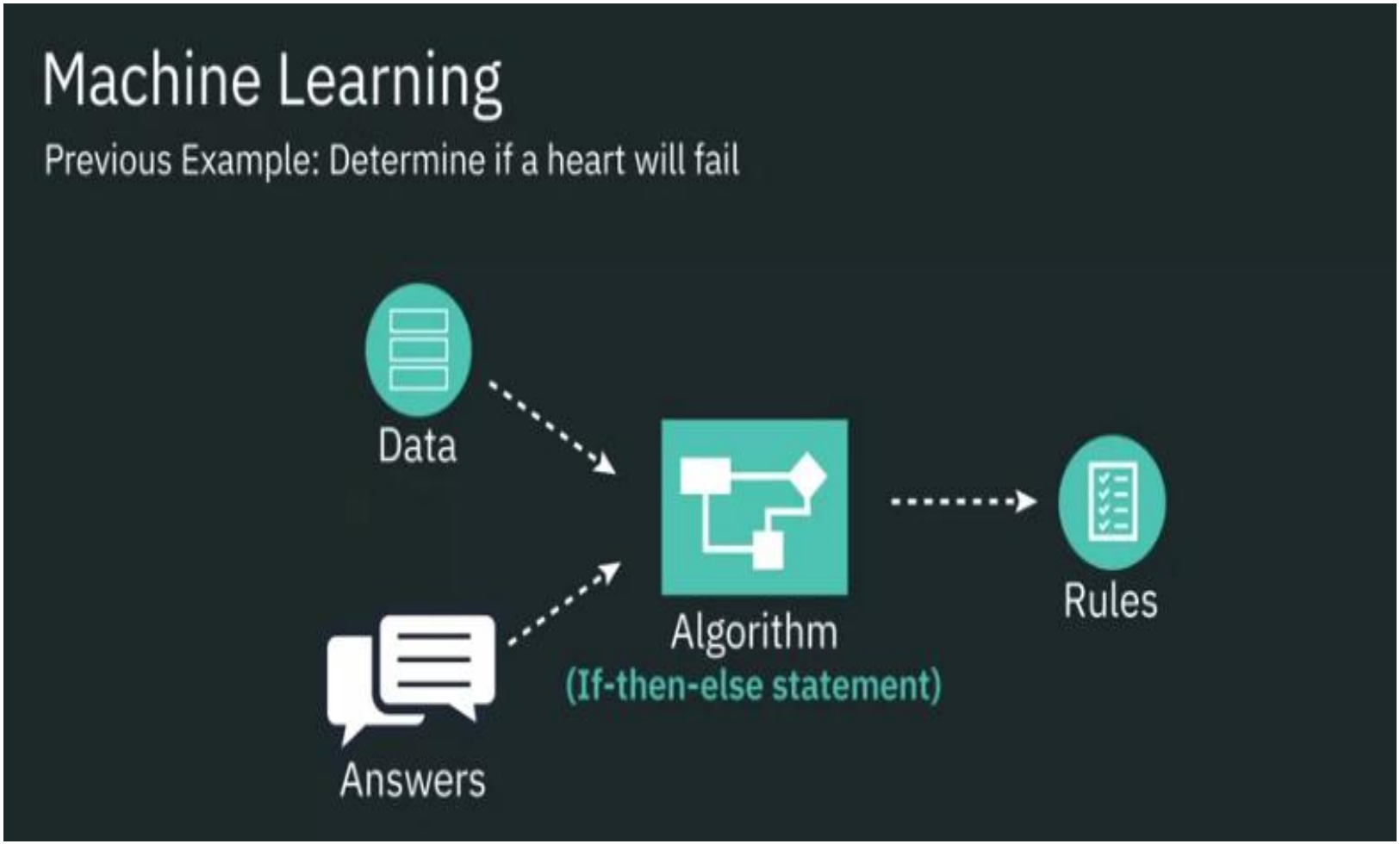
What's difference

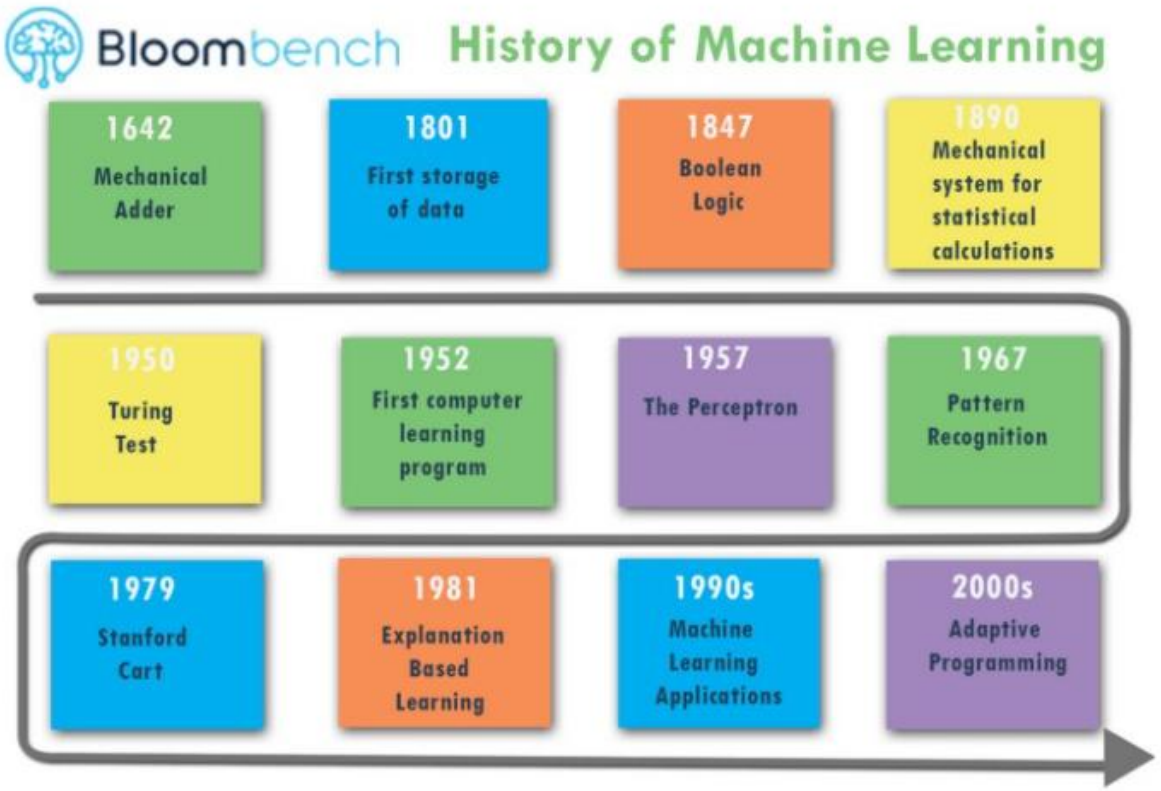


Machine Learning

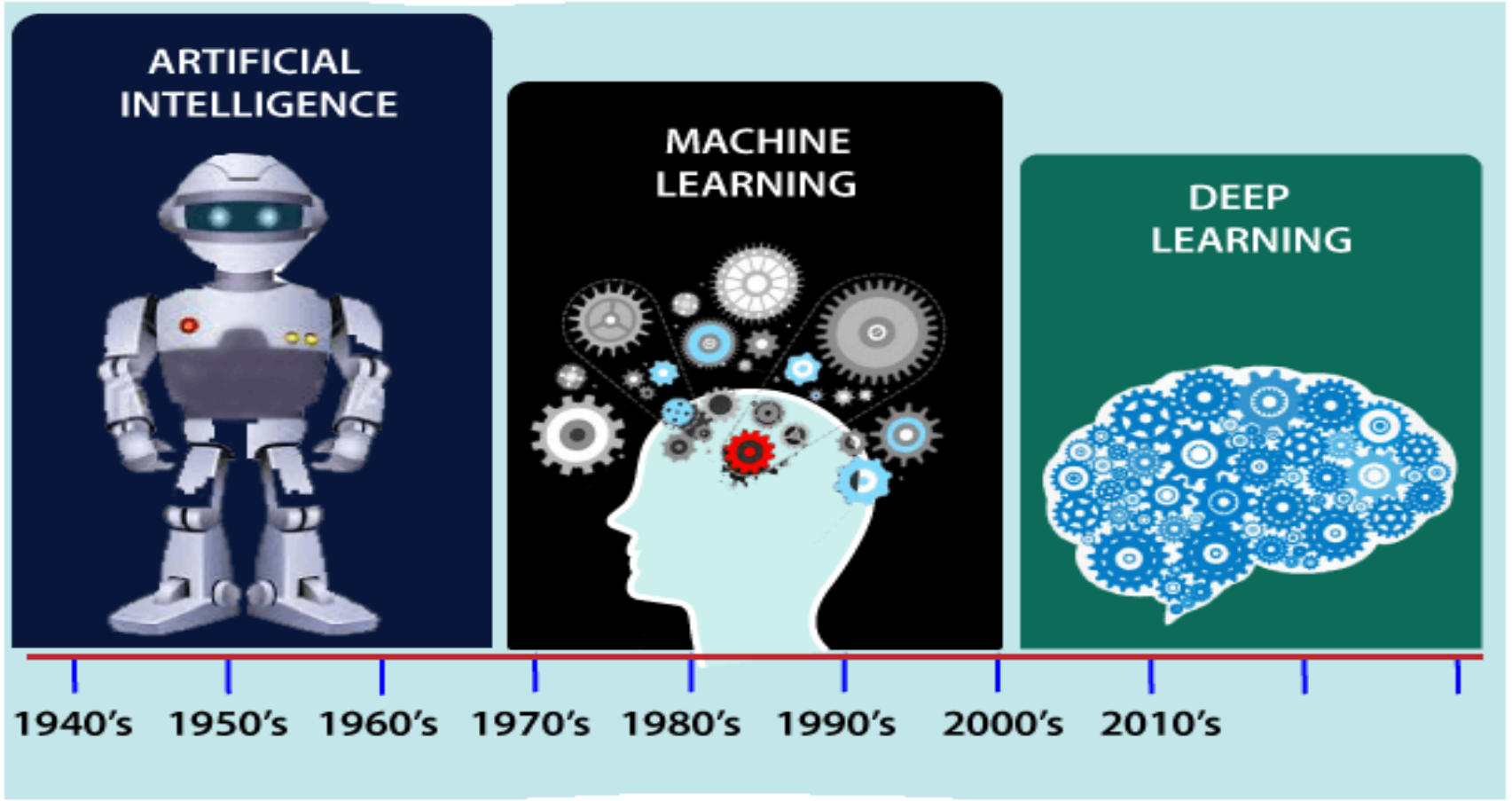


Machine Learning





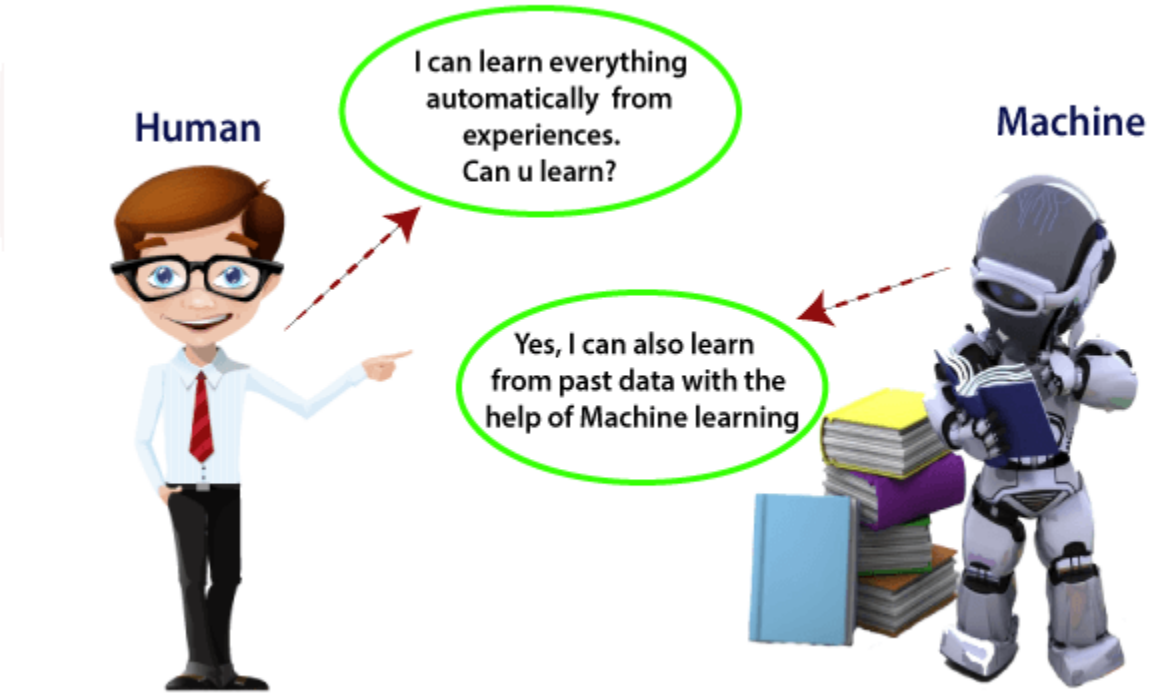
Down the Machine Learning lane

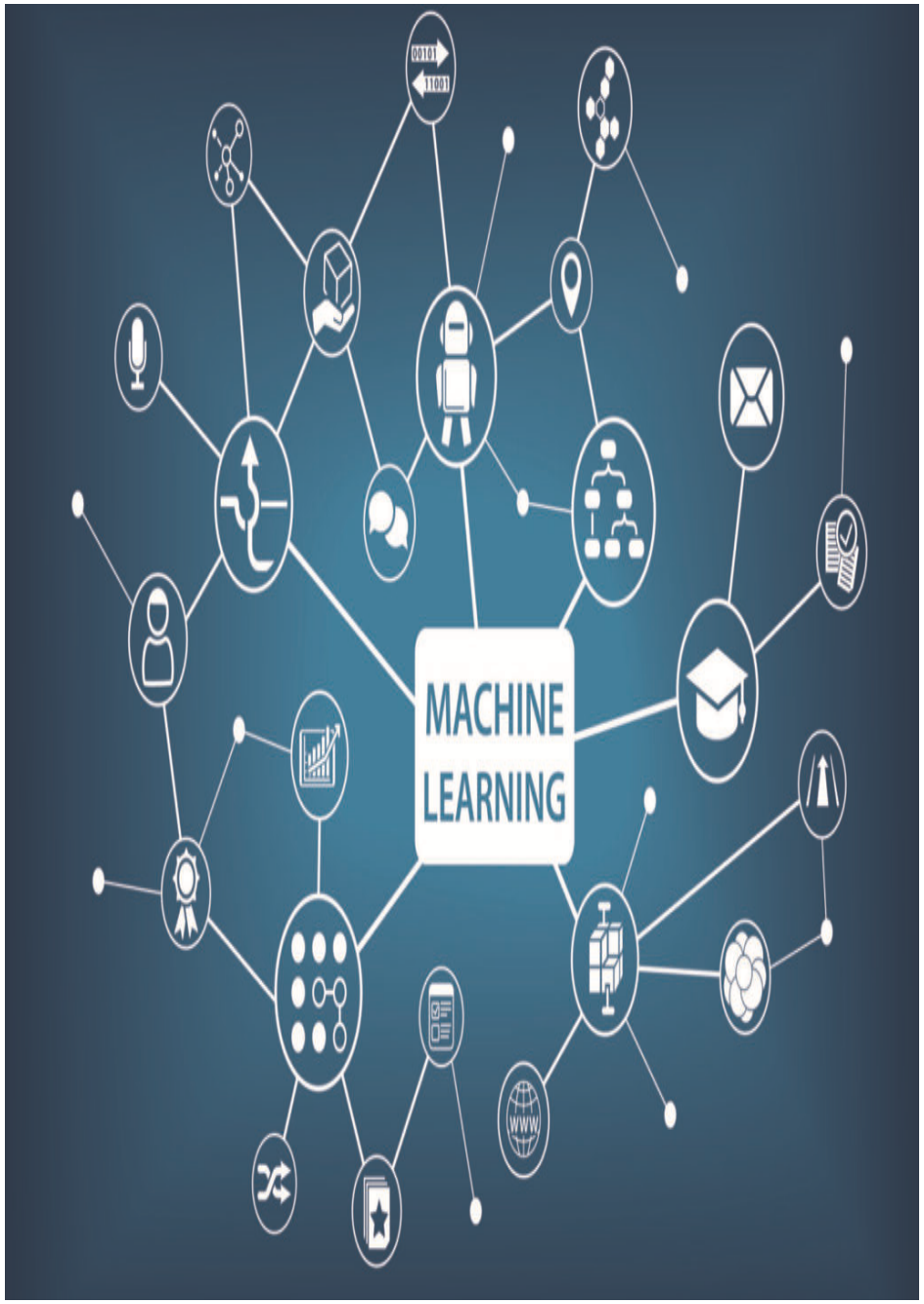

ARTIFICIAL INTELLIGENCE
MACHINE LEARNING
DEEP LEARNING
1940's
1950's
1960's
1970's
1980's
1990's
2000's
2010's

Lets have a glance and ....

- https://medium.com/bloombench/history-of-machine-learning-7c9dc67857a5
- https://cloud.withgoogle.com/build/data-analytics/explore-history-machine-learning/

Why ML is a boom now!!!!!




MACHINE
LEARNING


Humans are HOOKED
Machines are LEARNING

- https://tinyurl.com/54exnds9

For Mini Project Team formation link..

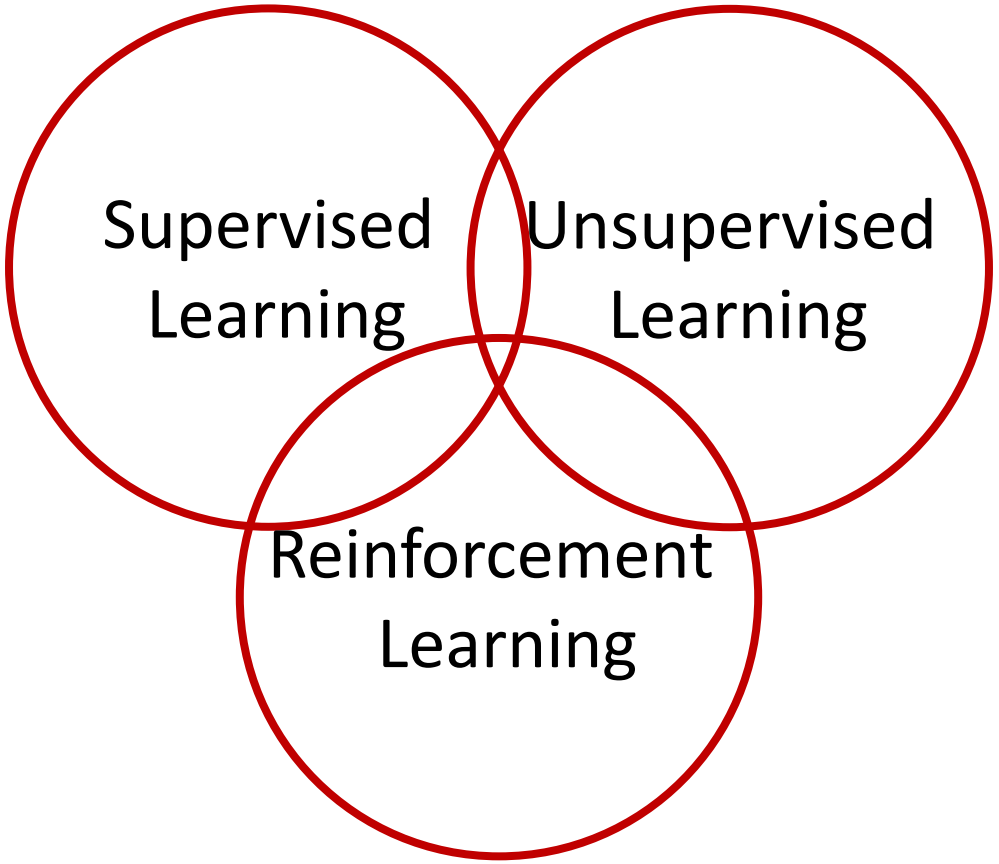
Machine Learning :

- **Learning** is **process** by which a system improves its performance by from experience
- Machine Learning is concerned with computer programs that automatically improve their performance through experience.

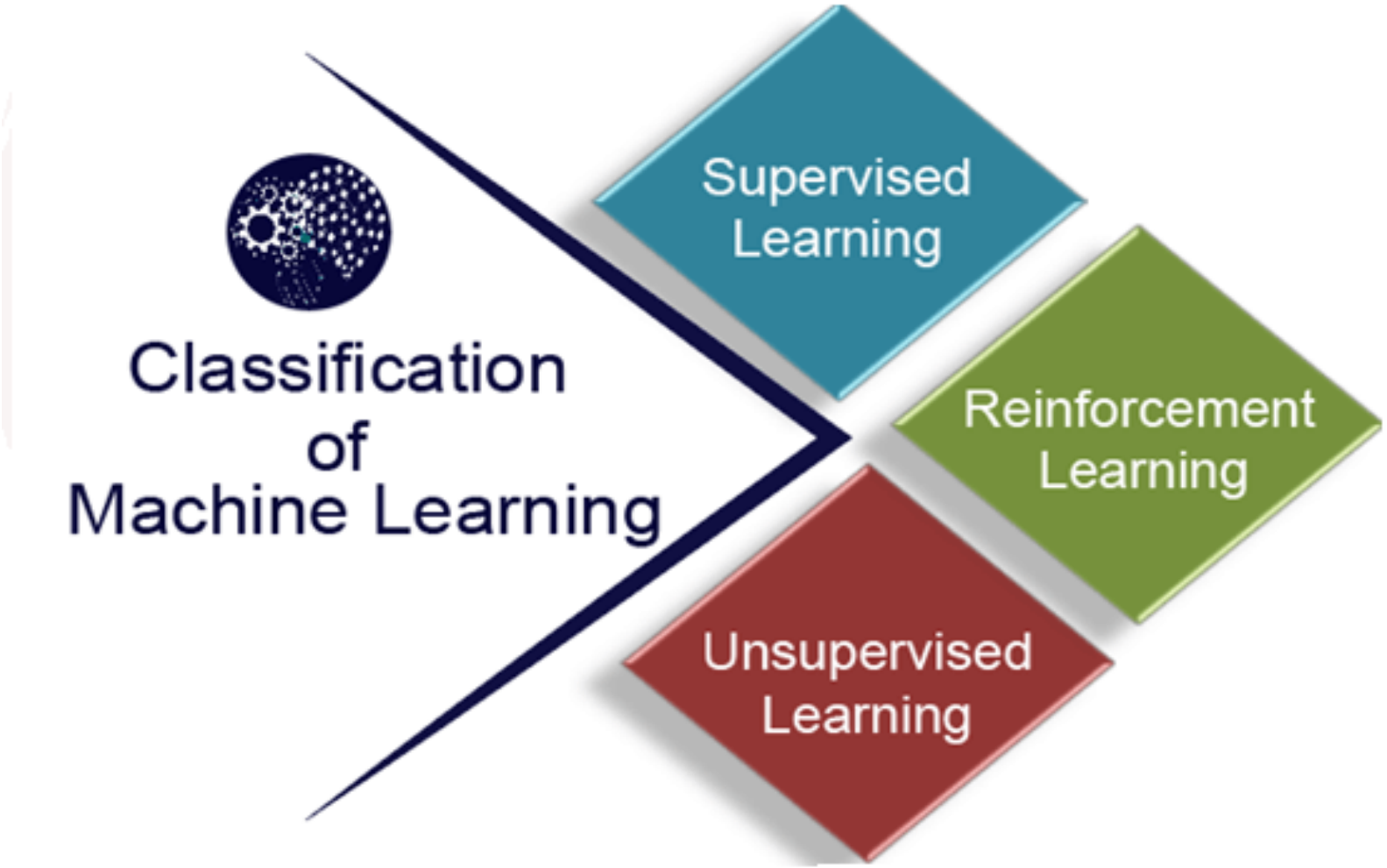
- For example, if a child is shown images with specific objects on them, they can learn to identify and differentiate between them.
- *Machine Learning works in the same way:* Through data input and certain commands, *the computer is enabled to* "learn" to identify certain objects (persons, objects, etc.) and to distinguish between them.

Taxonomy of Machine Learning

(A Simplistic View Based on Tasks)



can also be viewed as tools/methods


Classification
of
Machine Learning
Supervised
Learning
Reinforcement
Learning
Unsupervised
Learning

Supervised Learning

Supervised Learning

- Supervised learning is a type of machine learning method in which we provide sample labeled data to the machine learning system in order to train it, and on that basis, it predicts the output.

- The system creates a model using labeled data to understand the datasets and learn **about each data,** once the training and processing are done then we test the model by providing a sample data to check whether it is predicting the exact output or not.

- The goal of supervised learning **is to map input data with the output data.**
- The supervised learning is based on supervision, and it is the same as when a student learns things in the supervision of the teacher.
- The example of supervised learning is **spam filtering**.
- Supervised learning can be grouped further in two categories of algorithms:
- **Classification**
- **Regression**

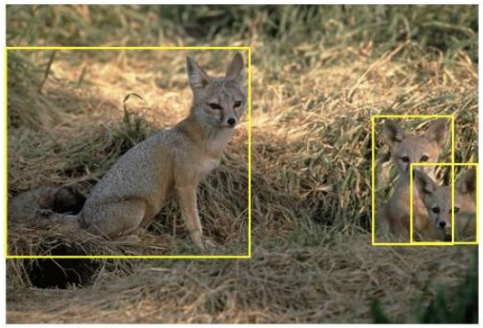
Supervised Learning in Computer Vision

- Image Classification
  - x = raw pixels of the image, y = the main object

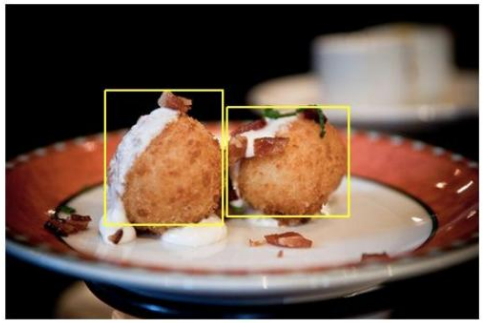
ImageNet Large Scale Visual Recognition Challenge. Russakovsky et al.'2015

Supervised Learning in Computer Vision

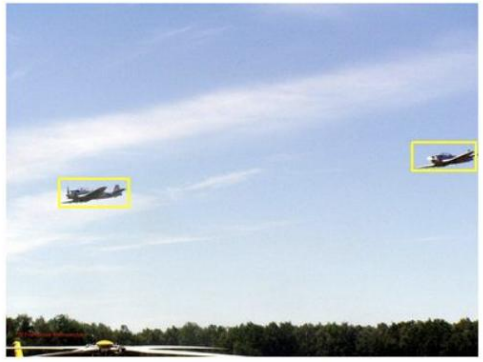
- Object localization and detection
  - x = raw pixels of the image, y = the bounding boxes



kit fox



croquette



airplane

frog

ImageNet Large Scale Visual Recognition Challenge. Russakovsky et al.'2015

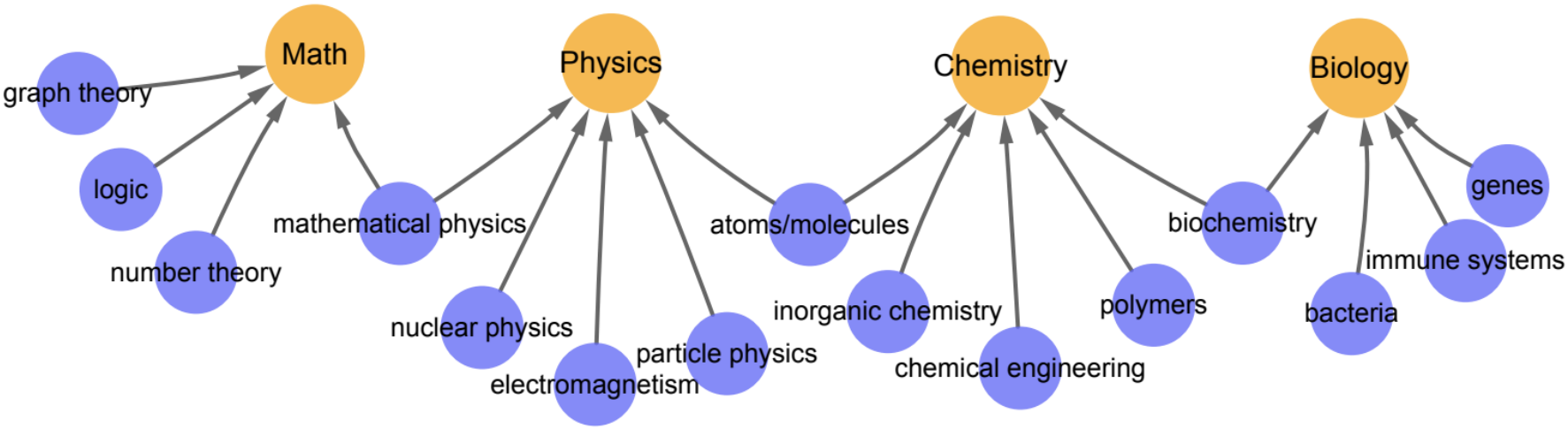
Unsupervised Learning

Unsupervised Learning

- Unsupervised learning is a learning method in which a machine learns without any supervision.
- The training is provided to the machine with the set of data that has not been labelled, classified, or categorized, and the algorithm needs to act on that data without any supervision.
- The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns.

- In unsupervised learning, we don't have a predetermined result.
- The machine tries to find useful insights from the huge amount of data.
- It can be further classifieds into two categories of algorithms:
- **Clustering**
- **Association**

Clustering Words with Similar Meanings (Hierarchically)



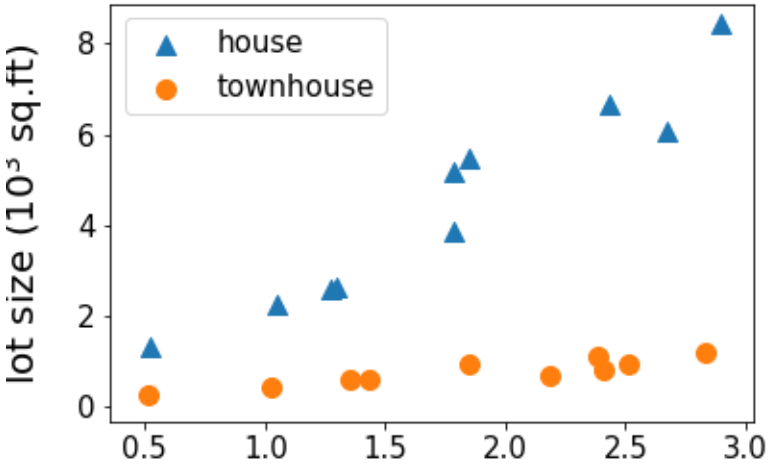
	logic deductive propositional semantics	graph subgraph bipartite vertex	boson massless particle higgs	polyester polypropylene resins epoxy	acids amino biosynthesis peptide
tag	*logic*	*graph theory*	*particle physics*	*polymer*	*biochemistry*

[Arora-Ge-Liang-M.-Risteski, TACL'17,18]

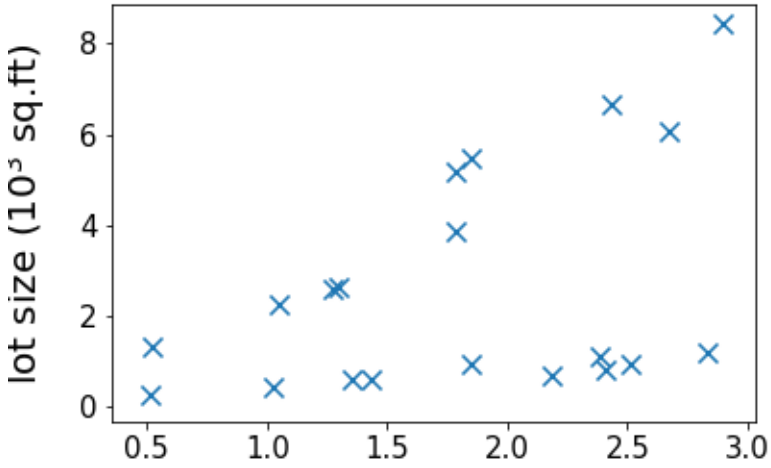
Unsupervised Learning

- Dataset contains no labels: $x^{(1)}, \ldots x^{(n)}$
- Goal (vaguely-posed): to find interesting structures in the data

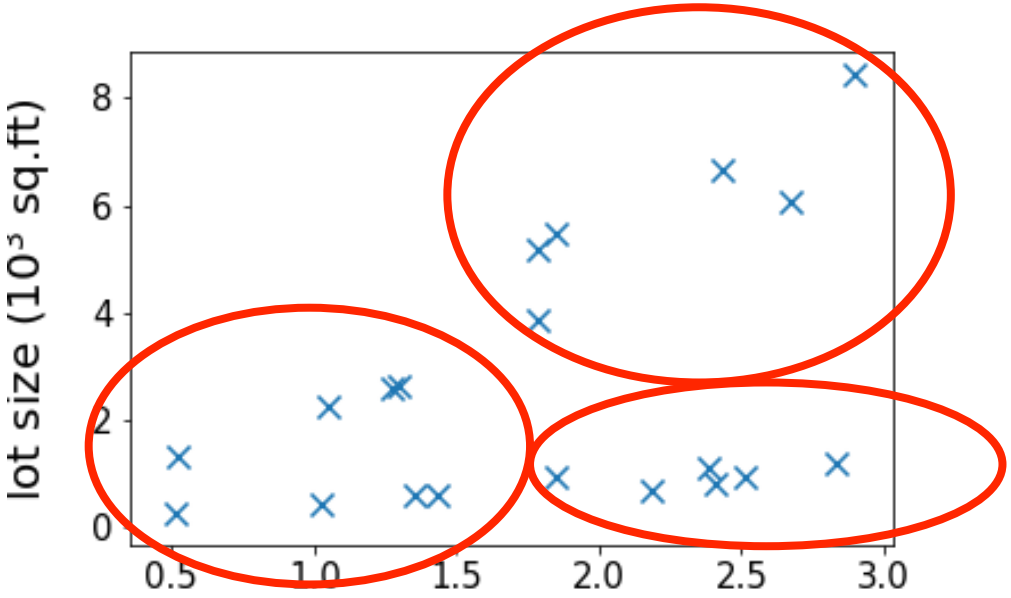
supervised



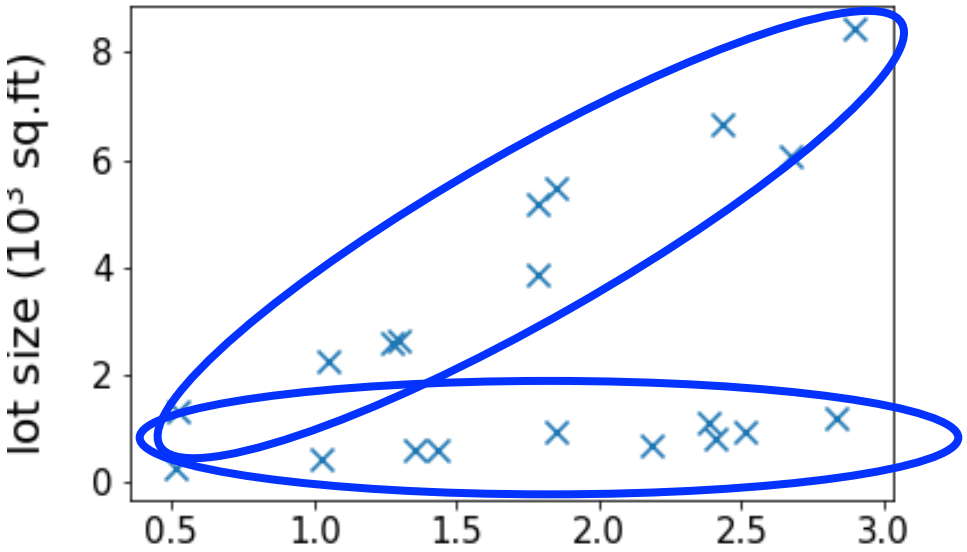
unsupervised



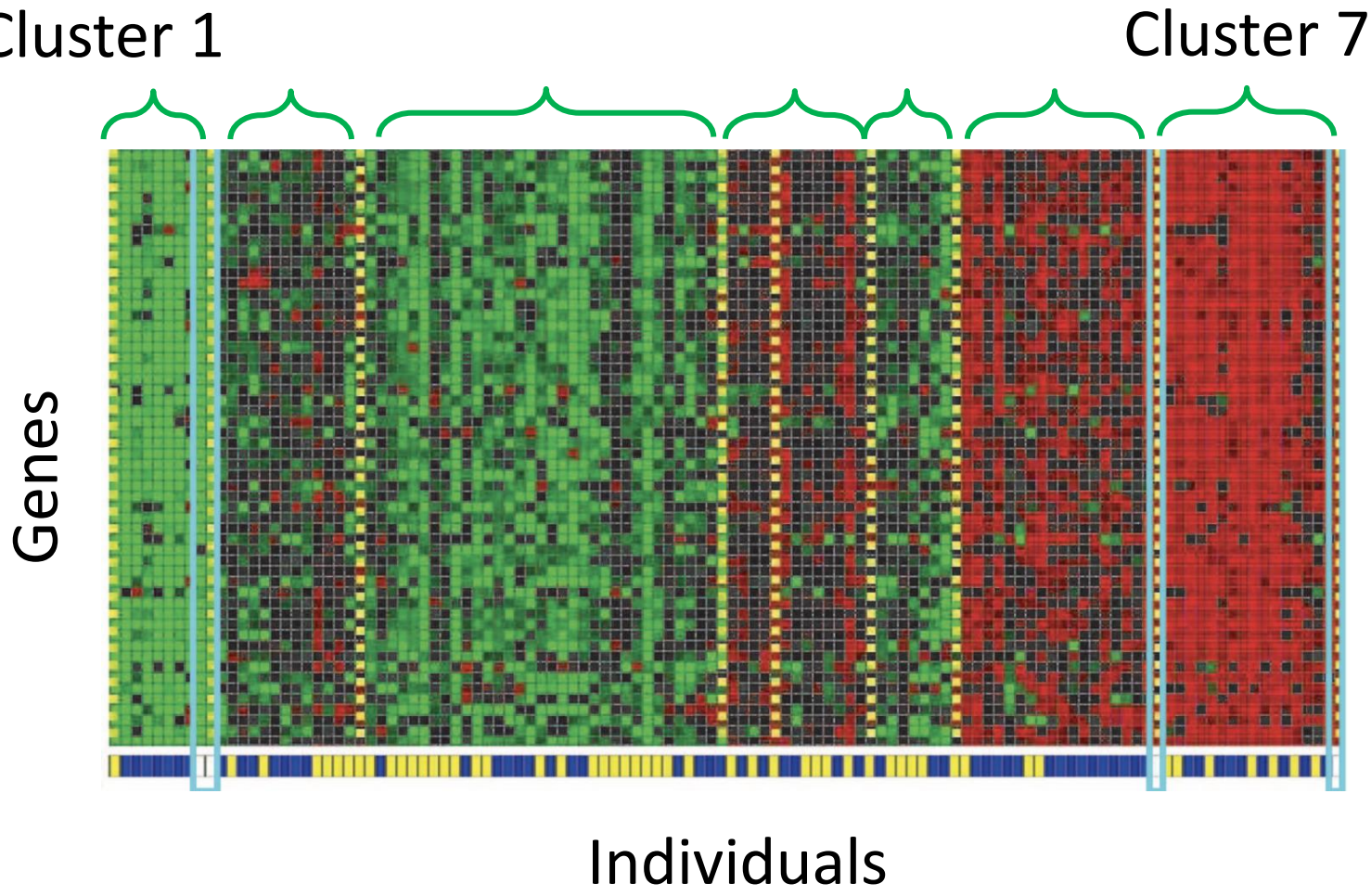
Clustering



Clustering



Clustering Genes



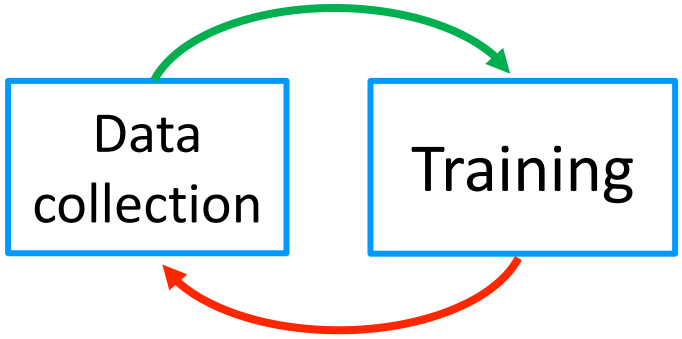
Identifying Regulatory Mechanisms using Individual Variation Reveals Key Role for Chromatin Modification. [Su-In Lee, Dana Pe'er, Aimee M. Dudley, George M. Church and Daphne Koller. ’06]

Reinforcement Learning

Reinforcement Learning

- The algorithm can collect **data interactively**

Try the strategy and collect feedbacks



Improve the strategy based on the feedbacks

Reinforcement Learning

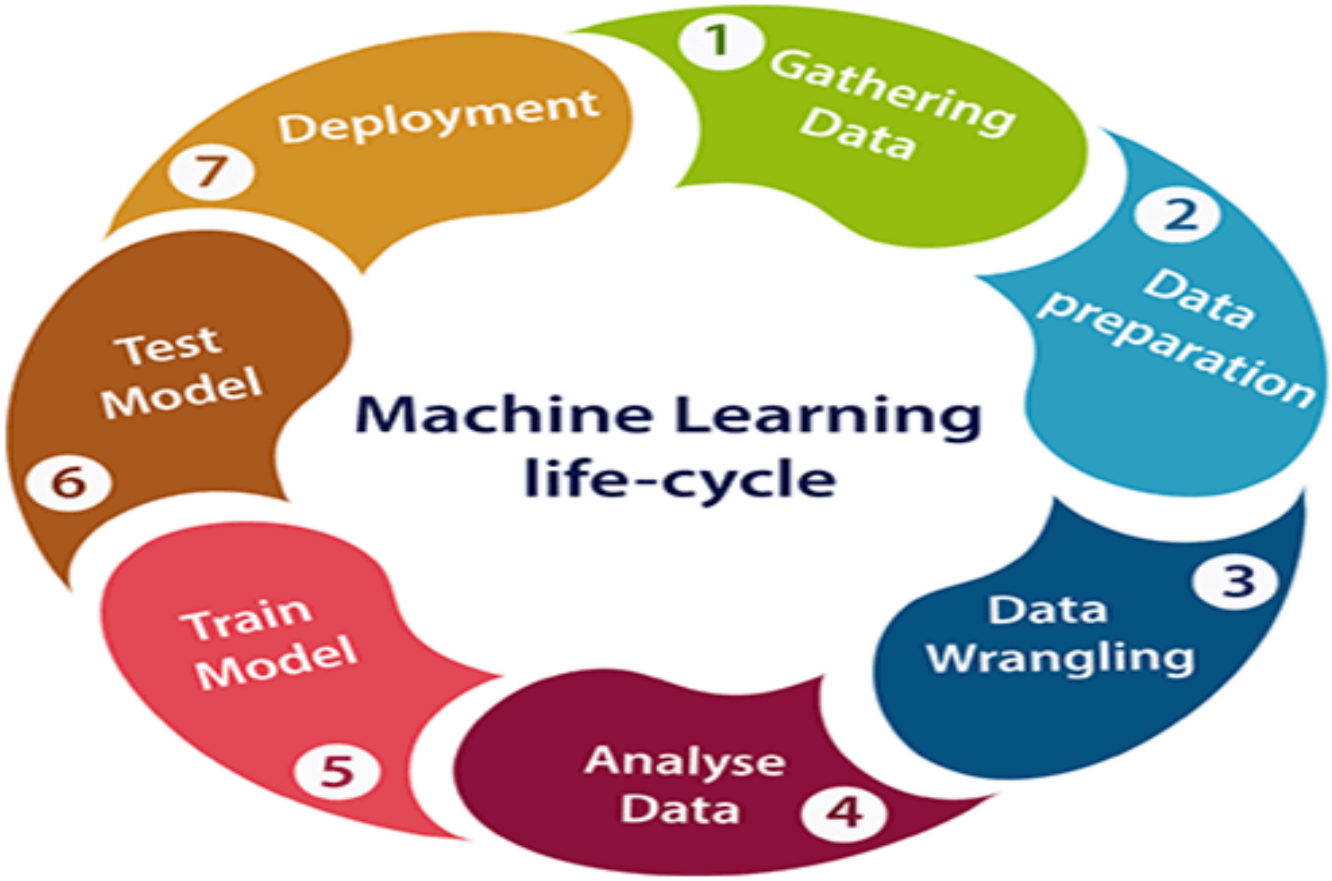
- Reinforcement learning is a feedback-based learning method, in which a learning agent gets a reward for each right action and gets a penalty for each wrong action.
- The agent learns automatically with these feedbacks and improves its performance.
- In reinforcement learning, the agent interacts with the environment and explores it.
- The goal of an agent is to get the most reward points, and hence, it improves its performance.

- The robotic dog, which automatically learns the movement of his arms, is an example of Reinforcement learning.

Issues in Machine Learning

- 1. Data Collection
- 2. Less Amount of Training Data
- 3. Non-representative Training Data
- 4. Poor Quality of Data
- 5. Irrelevant/Unwanted Features
- 6. Overfitting the Training Data
- 7. Underfitting the Training data
- 8. Offline Learning & Deployment of the model

Machine Learning Life Cycle



- Machine learning life cycle involves seven major steps, which are given below:
- **Gathering Data**
- **Data preparation**
- **Data Wrangling**
- **Analyse Data**
- **Train the model**
- **Test the model**
- **Deployment**

1. Gathering Data:

- Data Gathering is the first step of the machine learning life cycle. The goal of this step is to identify and obtain all data-related problems.
- In this step, we need to identify the different data sources, as data can be collected from various sources such as **files**, **database**, **internet**, or **mobile devices**. It is one of the most important steps of the life cycle. The quantity and quality of the collected data will determine the efficiency of the output. The more will be the data, the more accurate will be the prediction.

- This step includes the below tasks:
- **Identify various data sources**
- **Collect data**
- **Integrate the data obtained from different sources**
- By performing the above task, we get a coherent set of data, also called as a **dataset**. It will be used in further steps.

2. Data preparation

- After collecting the data, we need to prepare it for further steps. Data preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training.
- In this step, first, we put all data together, and then randomize the ordering of data.
- This step can be further divided into two processes:

- **Data exploration:**
  It is used to understand the nature of data that we have to work with.
- We need to understand the characteristics, format, and quality of data.
  A better understanding of data leads to an effective outcome.
- In this, we find Correlations, general trends, and outliers.
- **Data pre-processing:**
  Now the next step is preprocessing of data for its analysis.

3. Data Wrangling

- Data wrangling is the process of cleaning and converting raw data into a useable format.
- It is the process of cleaning the data, selecting the variable to use, and transforming the data in a proper format to make it more suitable for analysis in the next step.
- It is one of the most important steps of the complete process.
- Cleaning of data is required to address the quality issues.

- It is not necessary that data we have collected is always of our use as some of the data may not be useful. In real-world applications, collected data may have various issues, including:
- **Missing Values**
- **Duplicate data**
- **Invalid data**
- **Noise** So, we use various filtering techniques to clean the data.It is mandatory to detect and remove the above issues because it can negatively affect the quality of the outcome.

4. Data Analysis

- Now the cleaned and prepared data is passed on to the analysis step. This step involves:
- **Selection of analytical techniques**
- **Building models**
- **Review the result**

- The aim of this step is to build a machine learning model to analyze the data using various analytical techniques and review the outcome.
- It starts with the determination of the type of the problems, where we select the machine learning techniques such as **Classification**, **Regression**, **Cluster analysis**, **Association**, etc. then build the model using prepared data, and evaluate the model.
- Hence, in this step, we take the data and use machine learning algorithms to build the model.

5. Train Model

- Now the next step is to train the model, in this step we train our model to improve its performance for better outcome of the problem.
- We use datasets to train the model using various machine learning algorithms. Training a model is required so that it can understand the various patterns, rules, and, features.

6. Test Model

- Once our machine learning model has been trained on a given dataset, then we test the model.
- In this step, we check for the accuracy of our model by providing a test dataset to it.
- Testing the model determines the percentage accuracy of the model as per the requirement of project or problem.

7. Deployment

- The last step of machine learning life cycle is deployment, where we deploy the model in the real-world system.
- If the above-prepared model is producing an accurate result as per our requirement with acceptable speed, then we deploy the model in the real system.
- But before deploying the project, we will check whether it is improving its performance using available data or not.
- The deployment phase is similar to making the final report for a project.

Questions?

Thank you!

TextBooks

- 1. Machine Learning, Tom M. Mitchell, vMcGraw-Hill
- 2. Kevin Murphy, Machine Learning: A Probabilistic Perspective, MIT Press, 2012
- 3. Christopher Bishop, Pattern Recognition and Machine Learning, Springer, 2007.

References

- 1. Machine Learning: An Algorithmic Perspective, Stephen Marshland, Taylor & Francis

- 2. Machine Learning: The art and science of algorithms that make sense of data, Peter Flash, Cambridge. University press.

- 3. Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning, Springer 2009

Reference Materials and URL's

- http://cs229.stanford.edu/syllabus.html

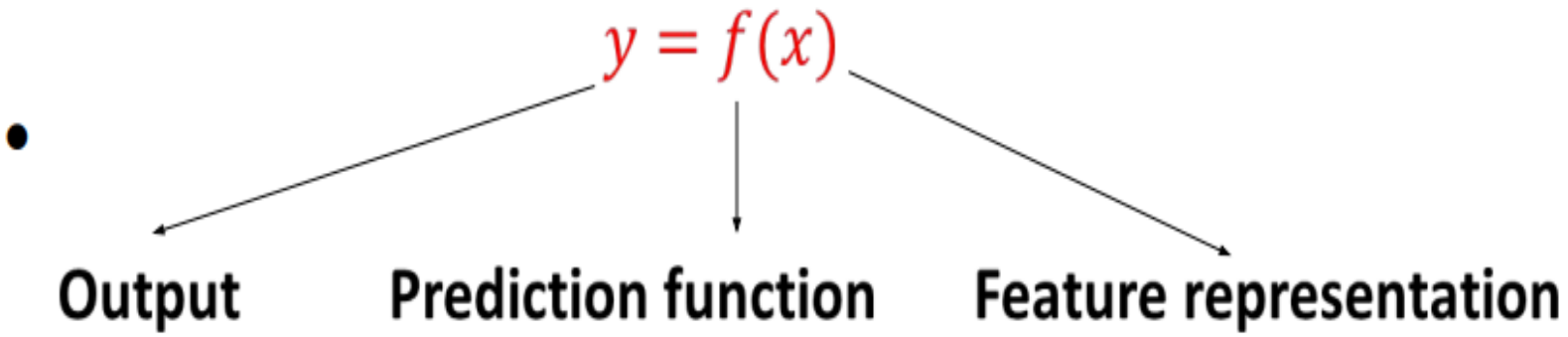
Certifications in Machine Learning

- Coursera
- Edx
- NPTEL
- Standford

Popular Master Trainers/Support

- Google
- Andrew

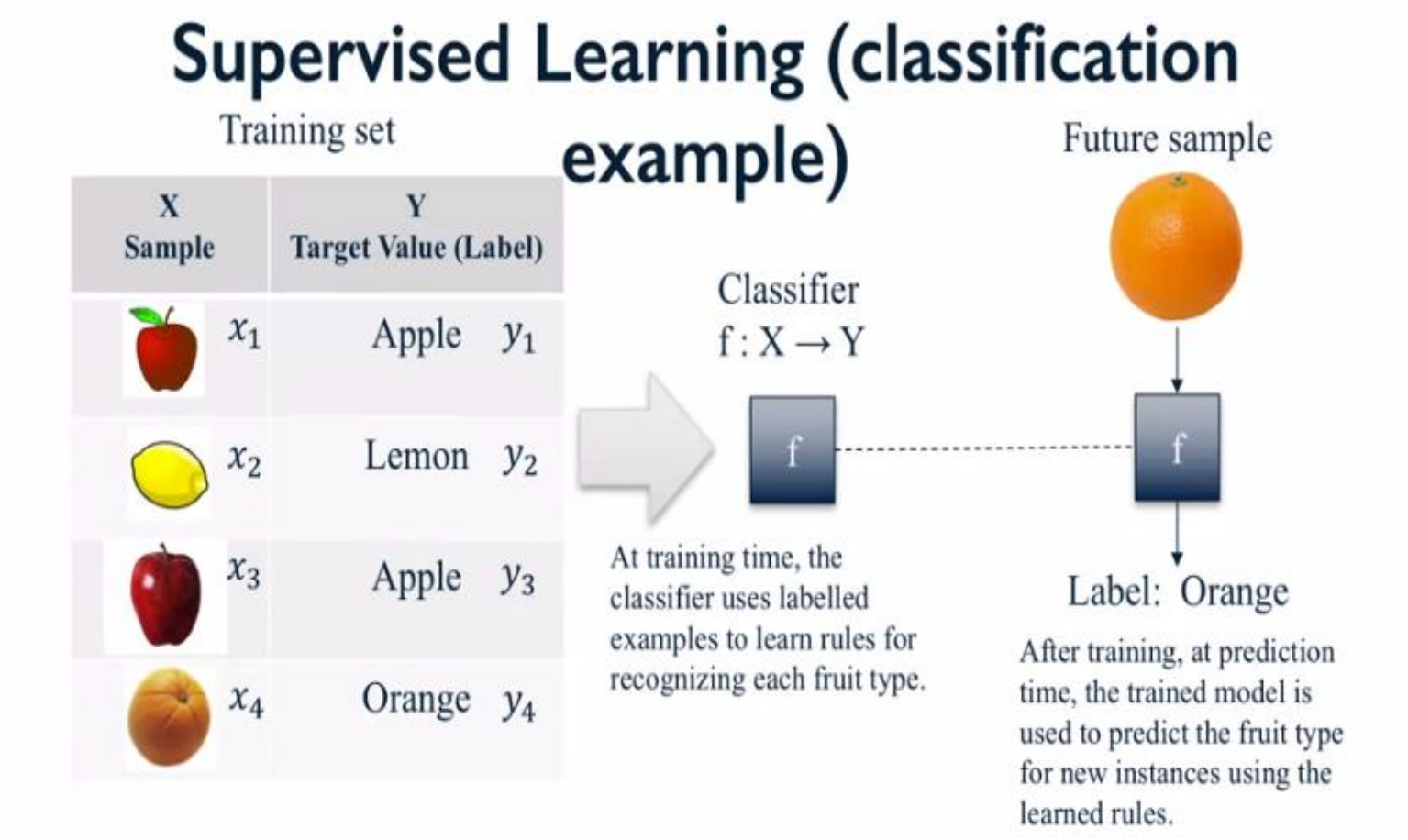
The Machine Learning Framework



Training: Given a training set, estimate the prediction function, $f(\,)$, by minimizing the prediction error

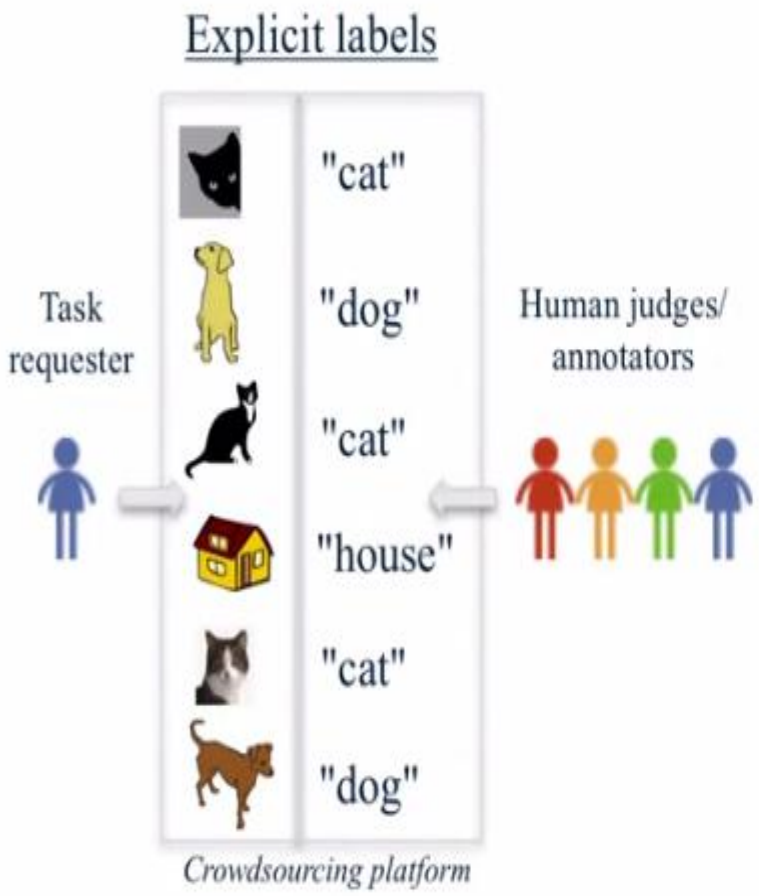
Testing: Apply $f(\,)$ to unknown test sample x and predicted value(output) is y.

Machine Learning as a training specialist



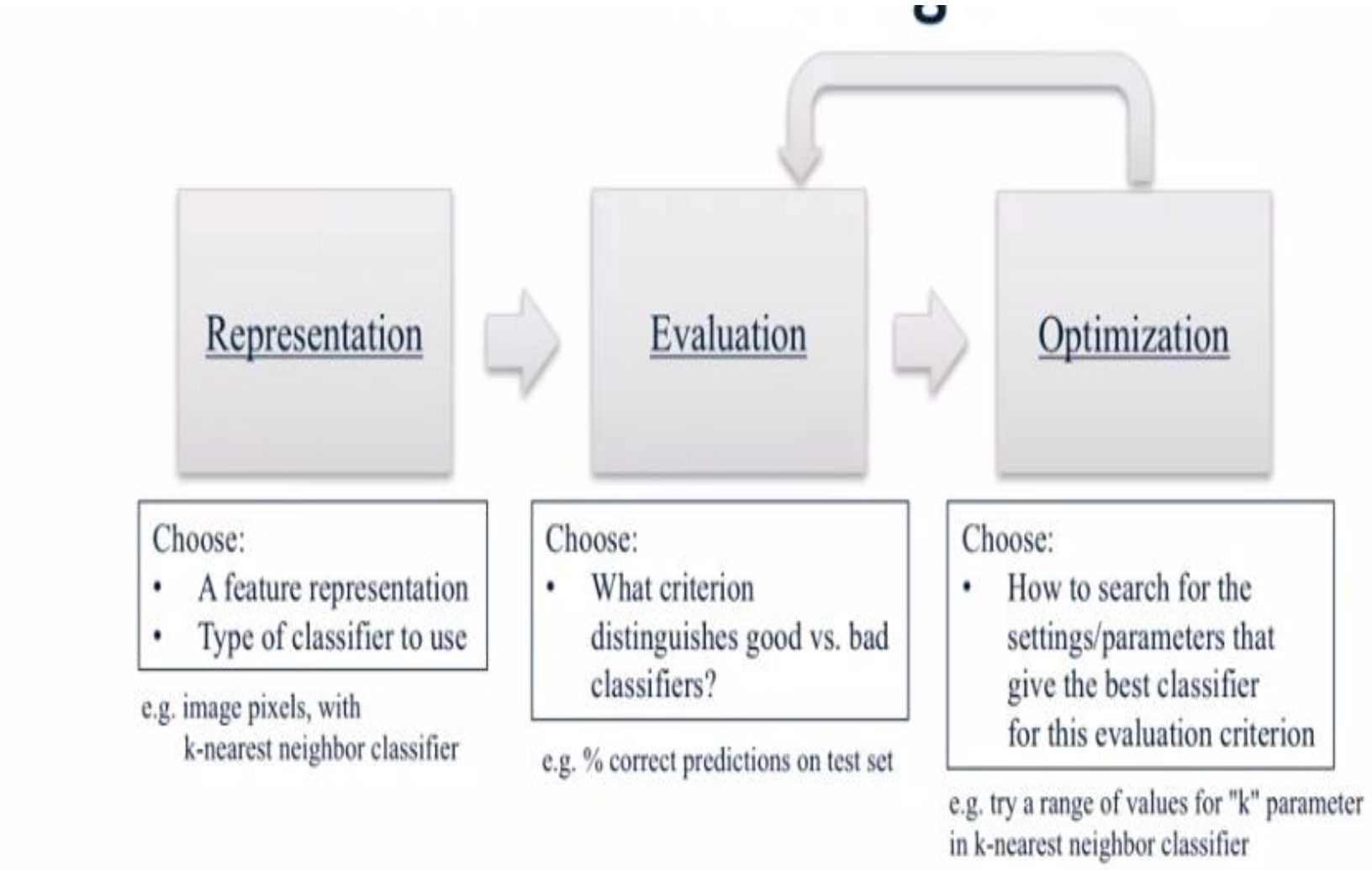
Machine Learning: What we do is adding labels

Examples of explicit and implicit label sources



Clicking and reading the "Mackinac Island" result can be an implicit label for the search engine to learn that "Mackinac Island" is especially relevant for the query `[vacations in michigan]` for that specific user.

A Basic Machine Learning Flow



Machine Learning

Feature Representations

Email

```
To:  Chris Brooks
From:  Daniel Romero
Subject:  Next course offering
Hi Daniel,
Could you please send the outline for the
next course offering?  Thanks! -- Chris
```

Feature	Count
to	1
chris	2
brooks	1
from	1
daniel	2
romero	1
the	2
...	

Feature representation

A list of words with their frequency counts

Picture

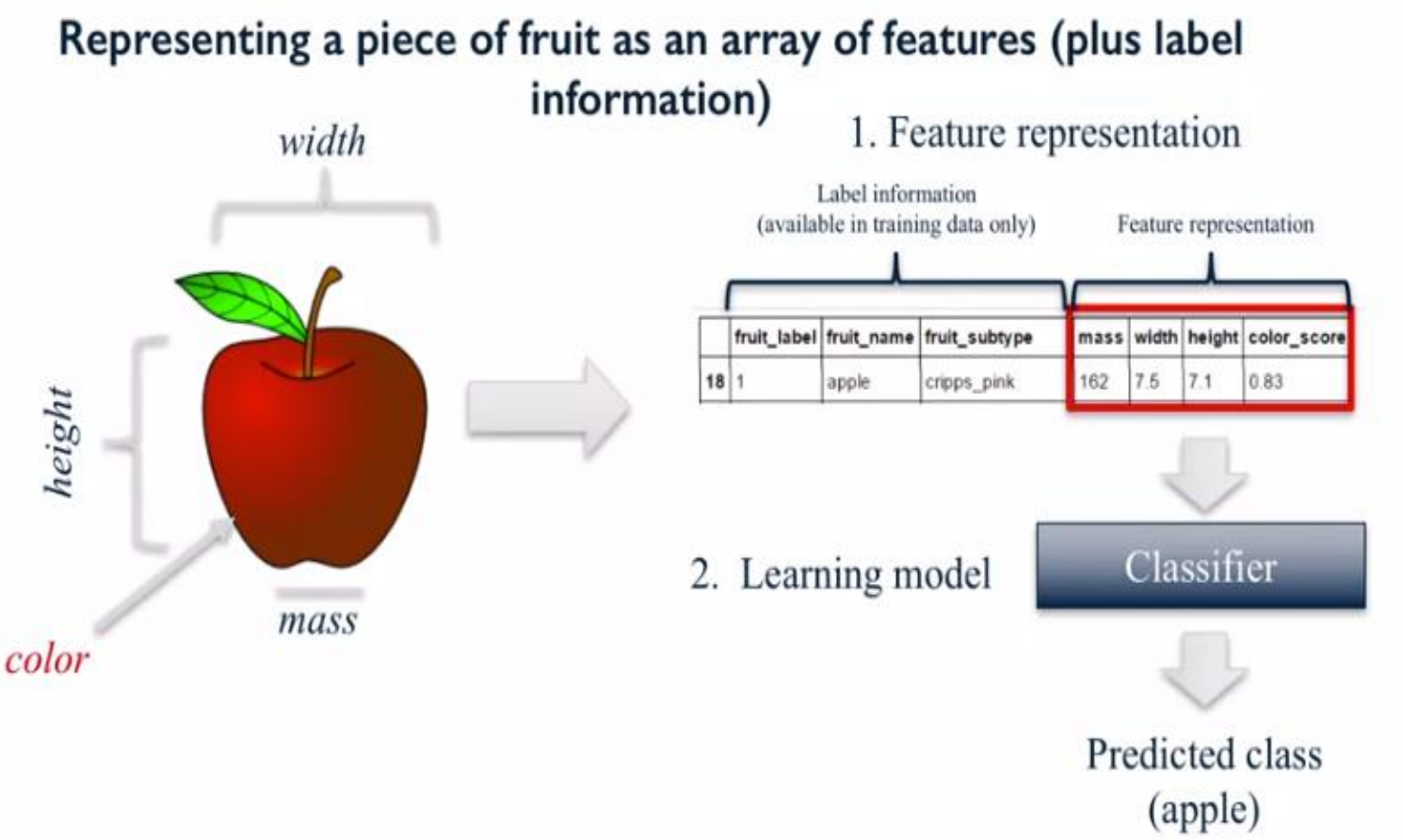
A matrix of color values (pixels)

Sea Creatures

Feature	Value
DorsalFin	Yes
MainColor	Orange
Stripes	Yes
StripeColor1	White
StripeColor2	Black

A set of attribute values

Feature Engineering



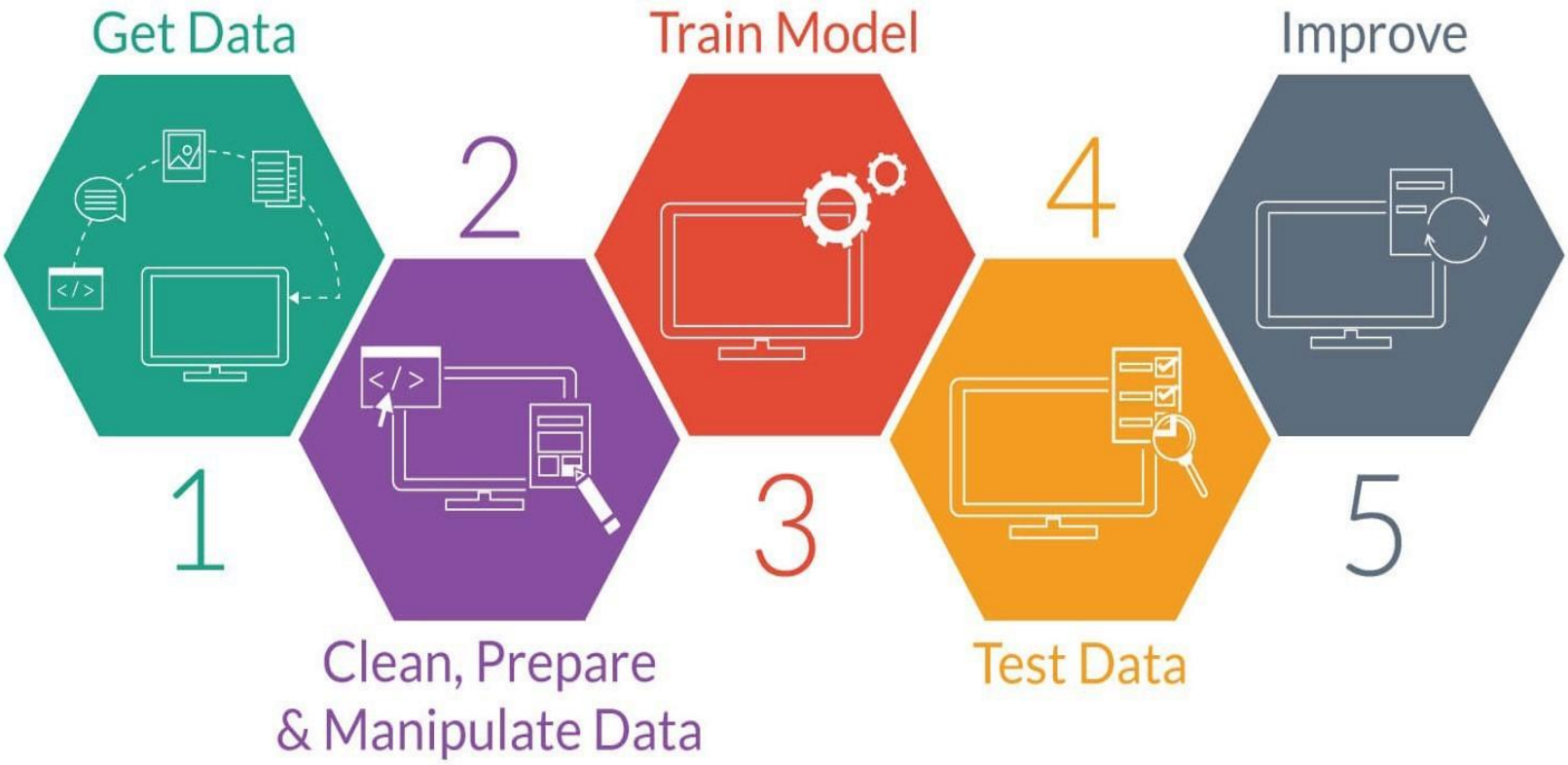
Machine Learning

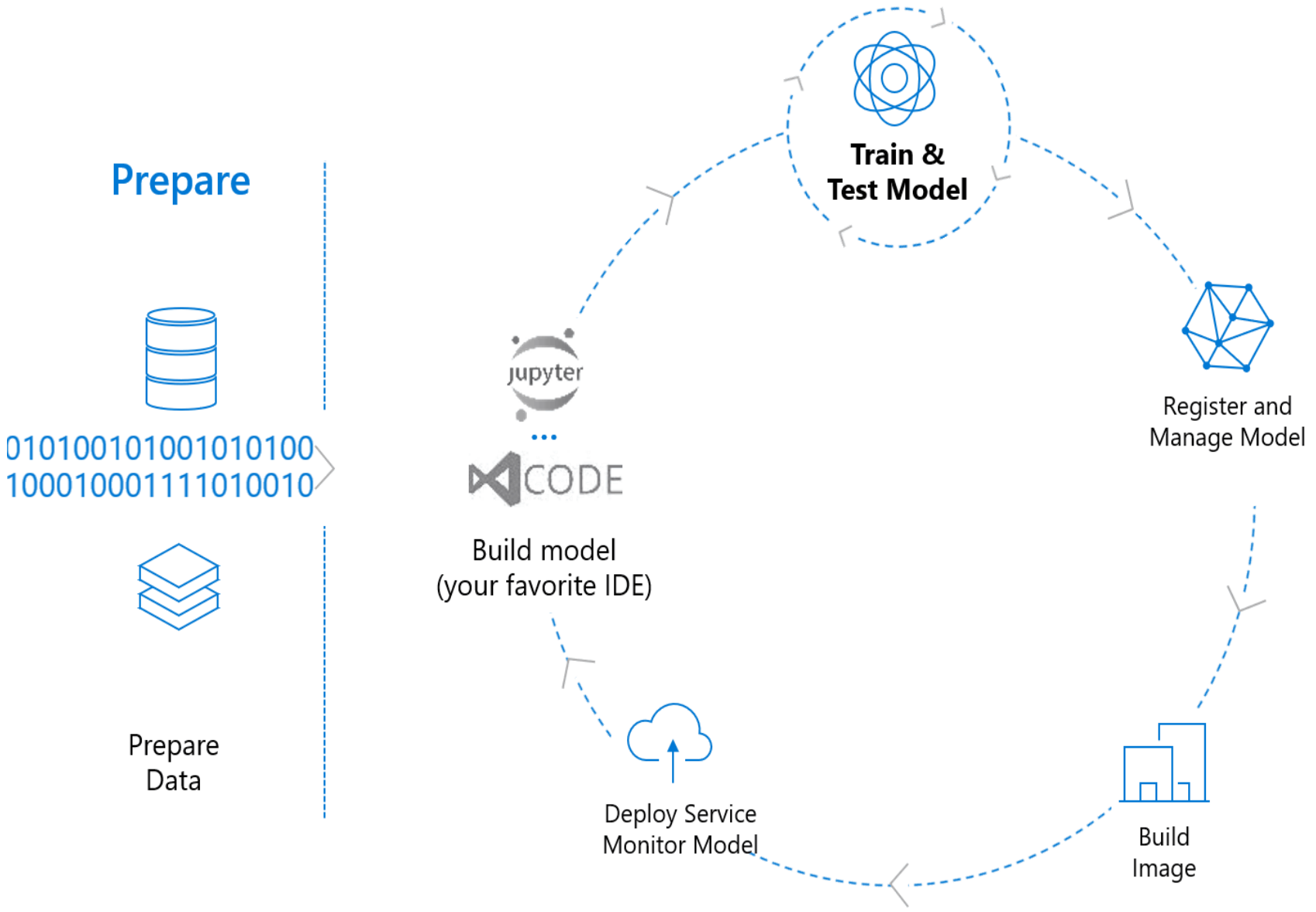
- **Why machine learning is hard?**

  Learning to identify an ‘apple’?

	Apple	Apple corporation	Peach
Colour	Red	White	Red
Type	Fruit	Logo	Fruit
Shape	Oval	Cut oval	Round

Steps used in Machine Learning

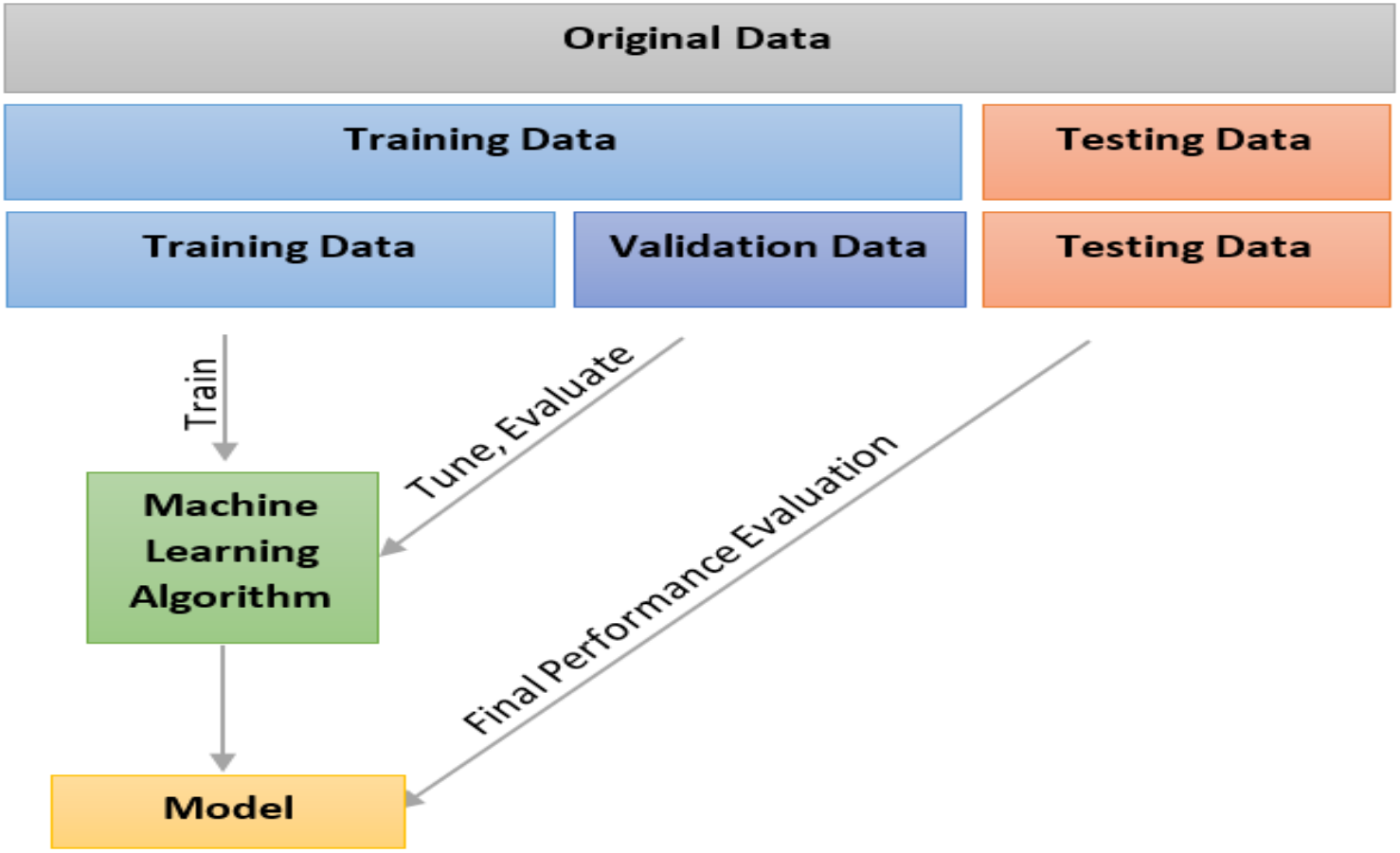



Prepare
01010010100101010010001000111101001
Prepare Data
jupyter
CODE
Build model (your favorite IDE)
Train & Test Model
Register and Manage Model
Build Image
Deploy Service Monitor Model

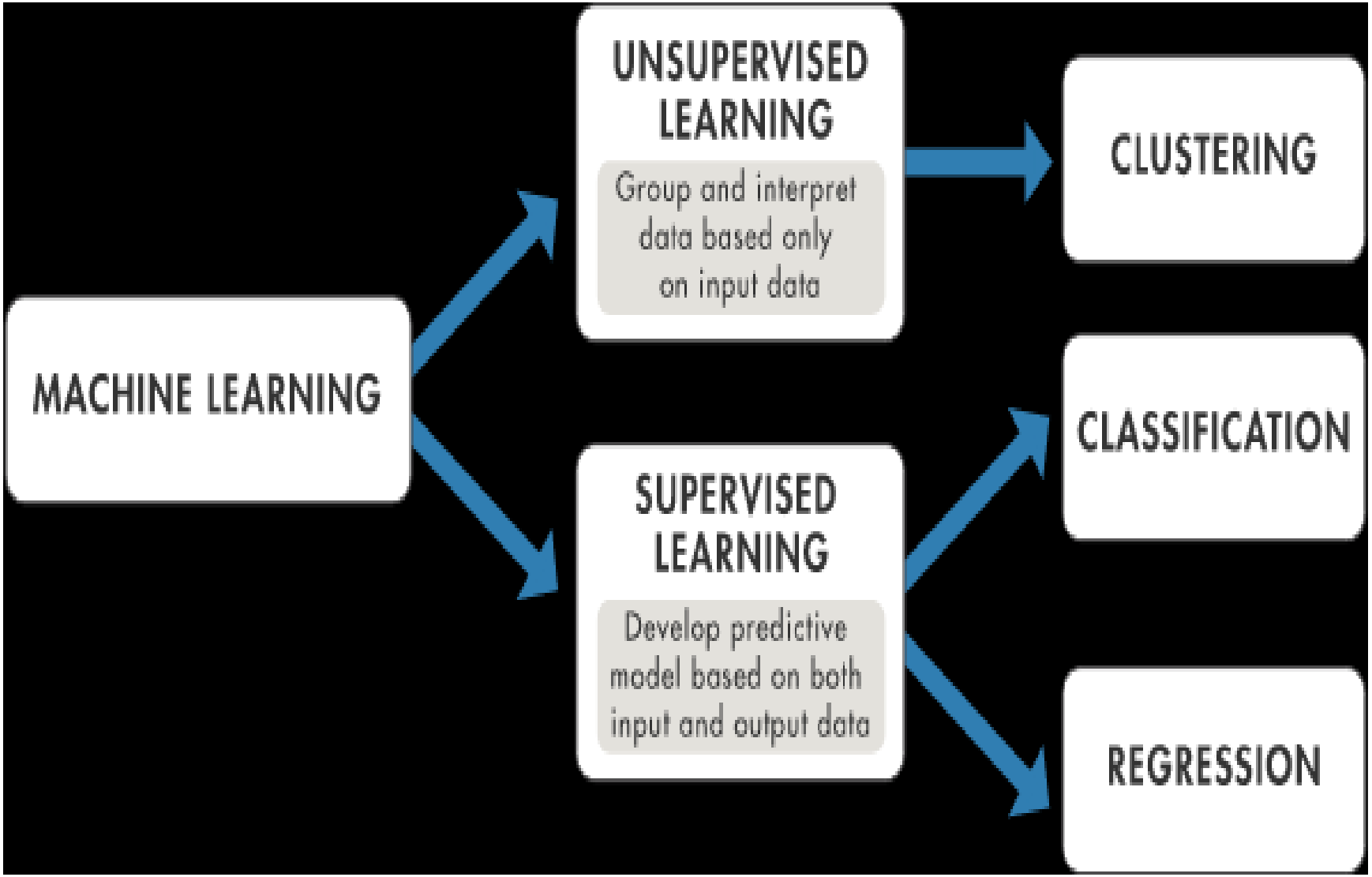
Machine Learning – Key Steps

- **Train Data** for Model Preparation
- **Validate Data** for Evaluate the Model
- **Test Data** for Model Performance Assessment(final)

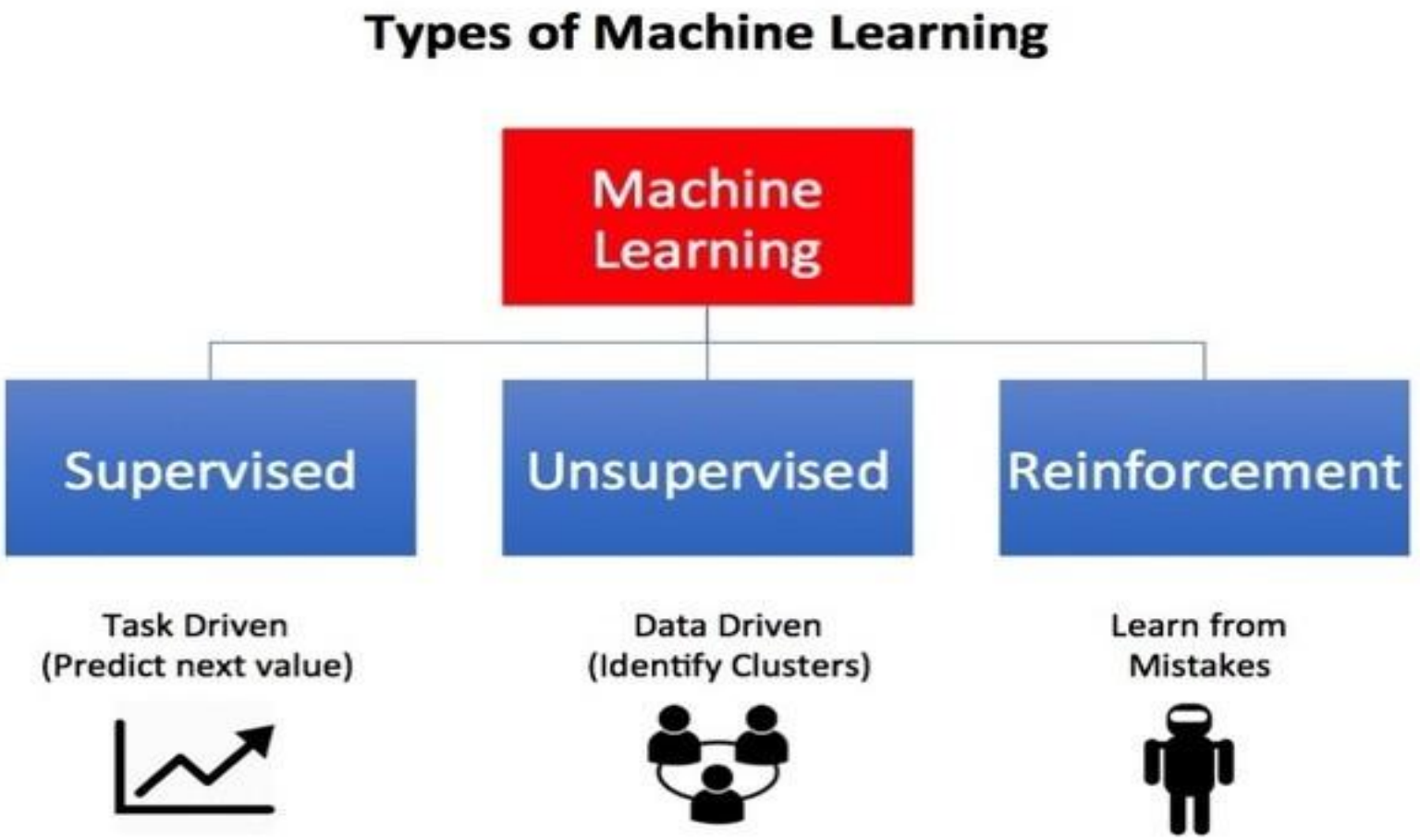
Machine Learning: How we proceed



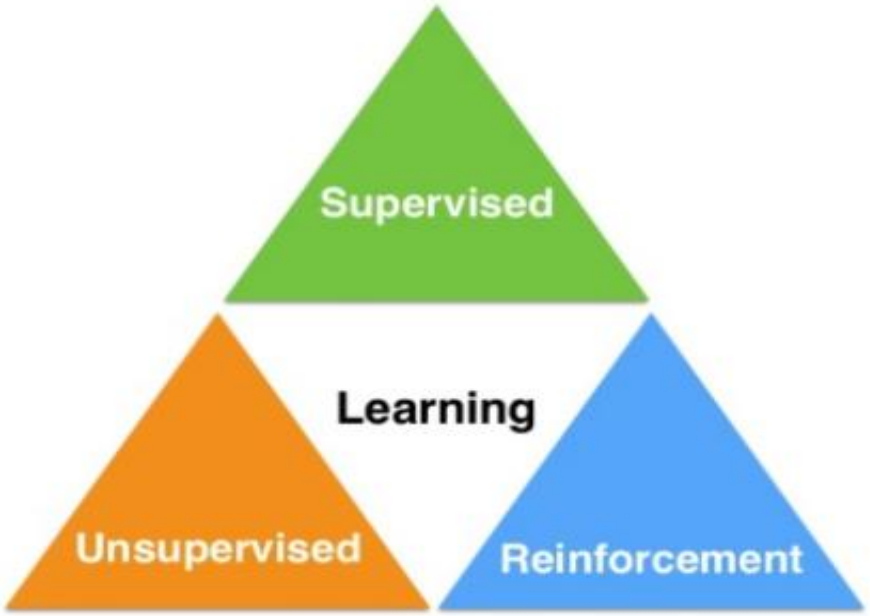
Machine Learning Techniques



Machine Learning Techniques



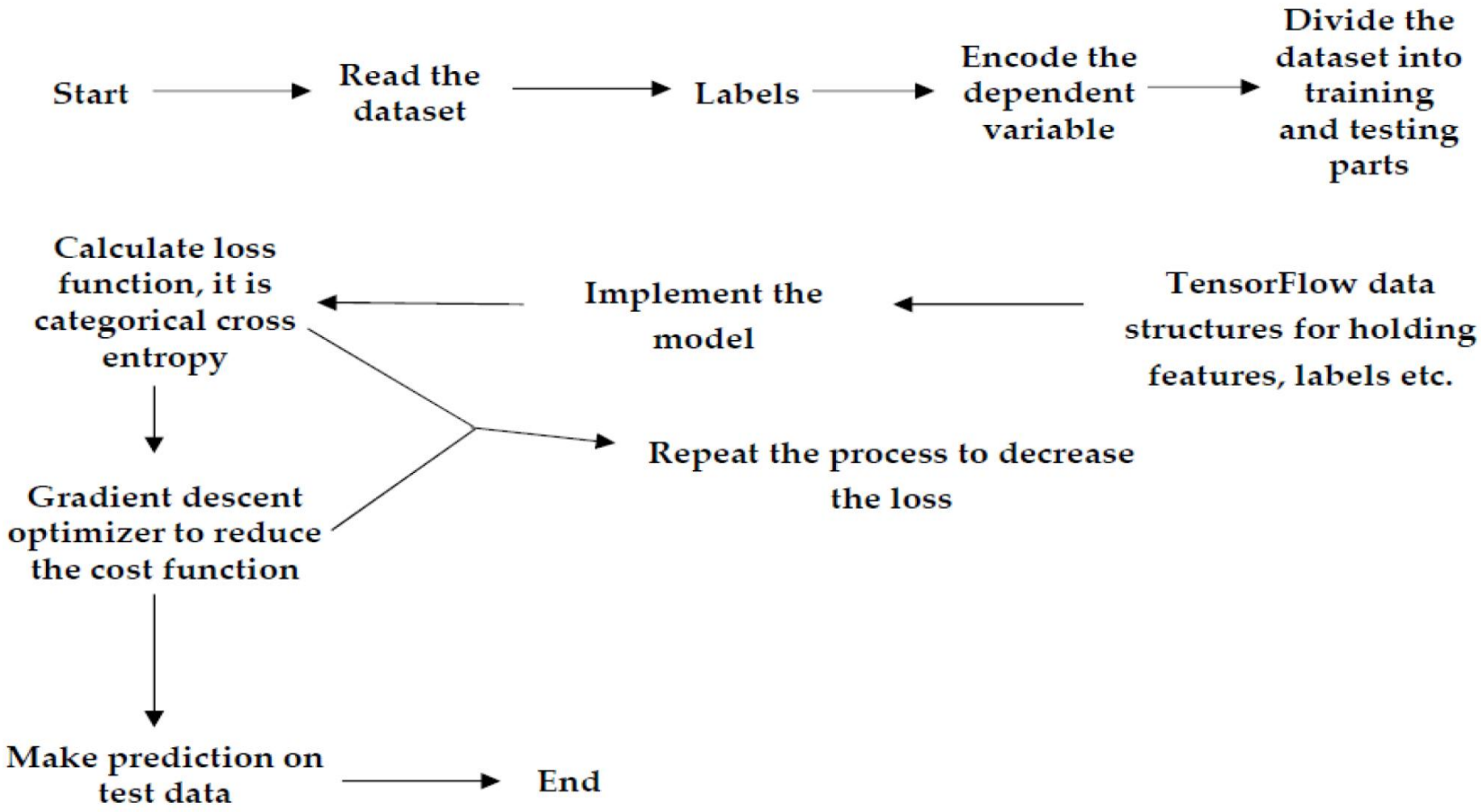
- Labeled data
- Direct feedback
- Predict outcome/future



- No labels
- No feedback
- "Find hidden structure"

- Decision process
- Reward system
- Learn series of actions

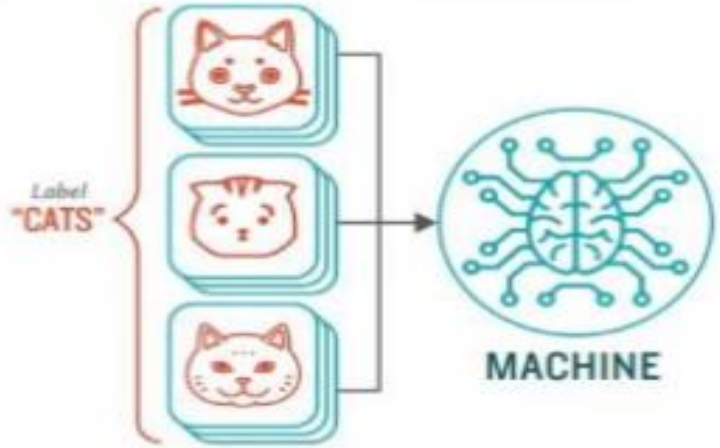
Steps in Developing a Model



How **Supervised** Machine Learning Works

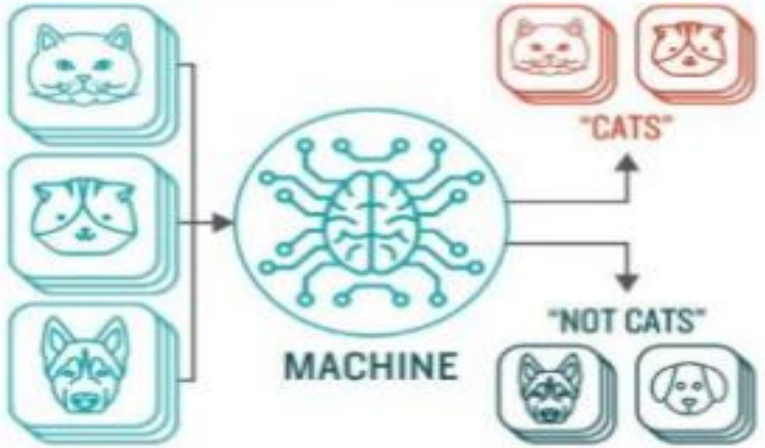
STEP 1

Provide the machine learning algorithm categorized or "labeled" input and output data from to learn

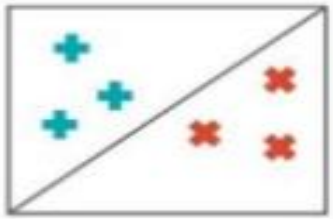


STEP 2

Feed the machine new, unlabeled information to see if it tags new data appropriately. If not, continue refining the algorithm



TYPES OF PROBLEMS TO WHICH IT'S SUITED



CLASSIFICATION

Sorting items into categories

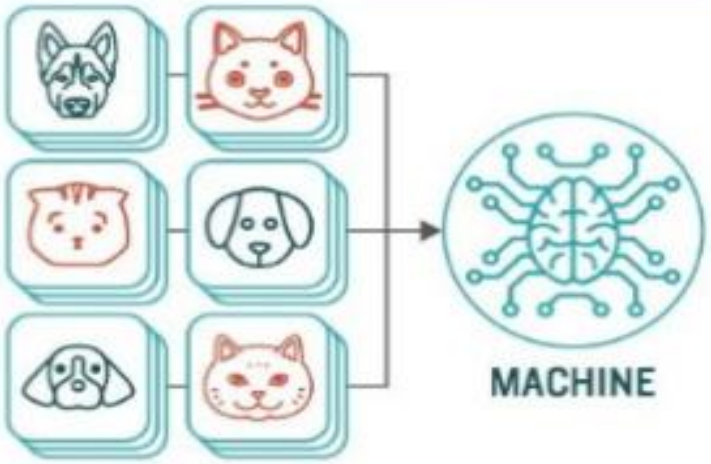
REGRESSION

Identifying real values (dollars, weight, etc.)

How **Unsupervised** Machine Learning Works

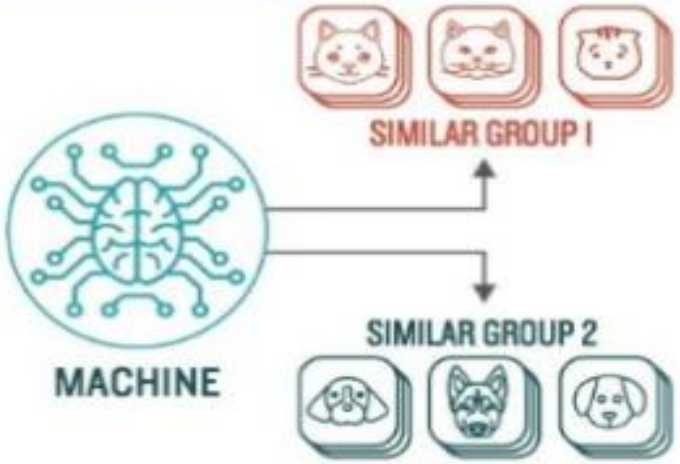
STEP 1

Provide the machine learning algorithm uncategorized, unlabeled input data to see what patterns it finds

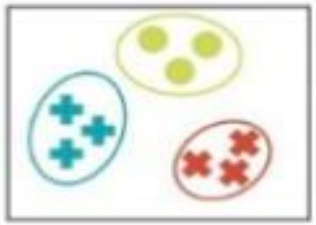


STEP 2

Observe and learn from the patterns the machine identifies



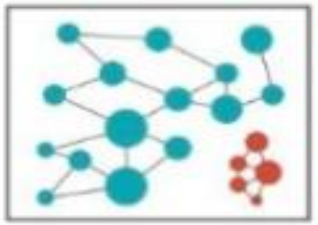
TYPES OF PROBLEMS TO WHICH IT'S SUITED



CLUSTERING

Identifying similarities in groups

For Example: Are there patterns in the data to indicate certain patients will respond better to this treatment than others?



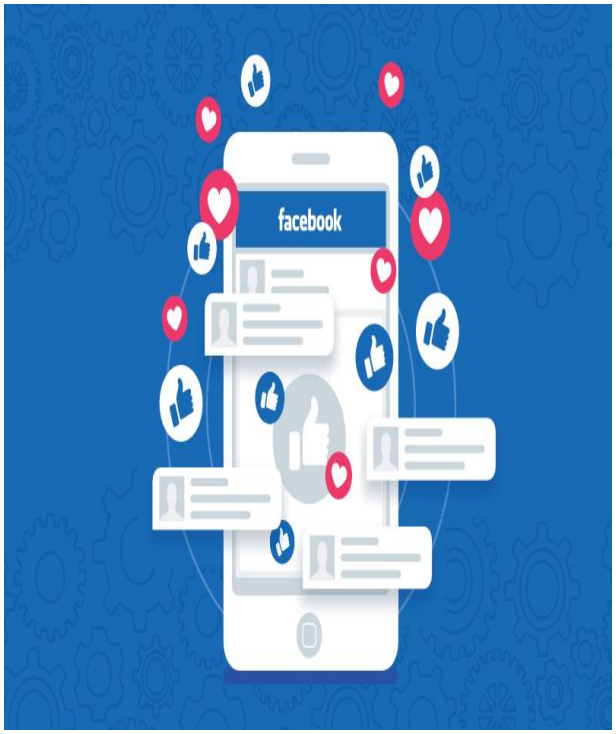
ANOMALY DETECTION

Identifying abnormalities in data

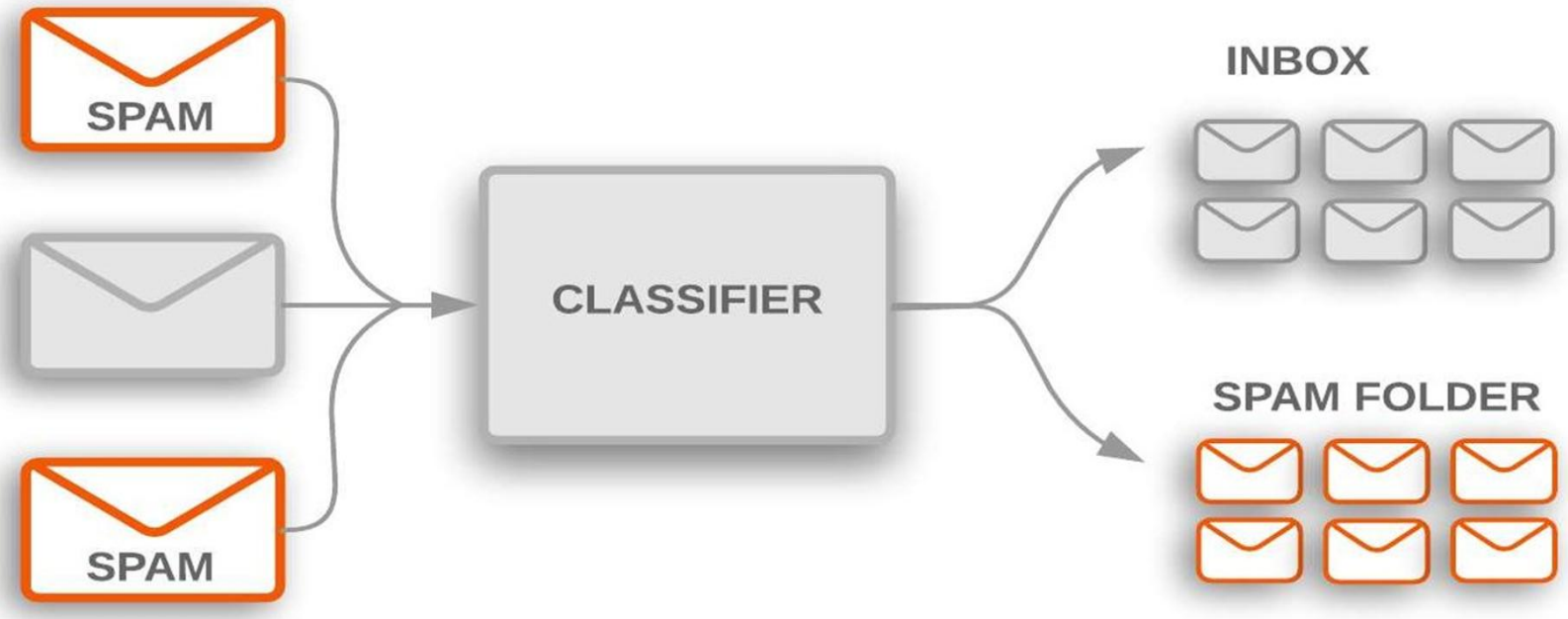
For Example: Is a hacker intruding in our network?

ML Applications-1

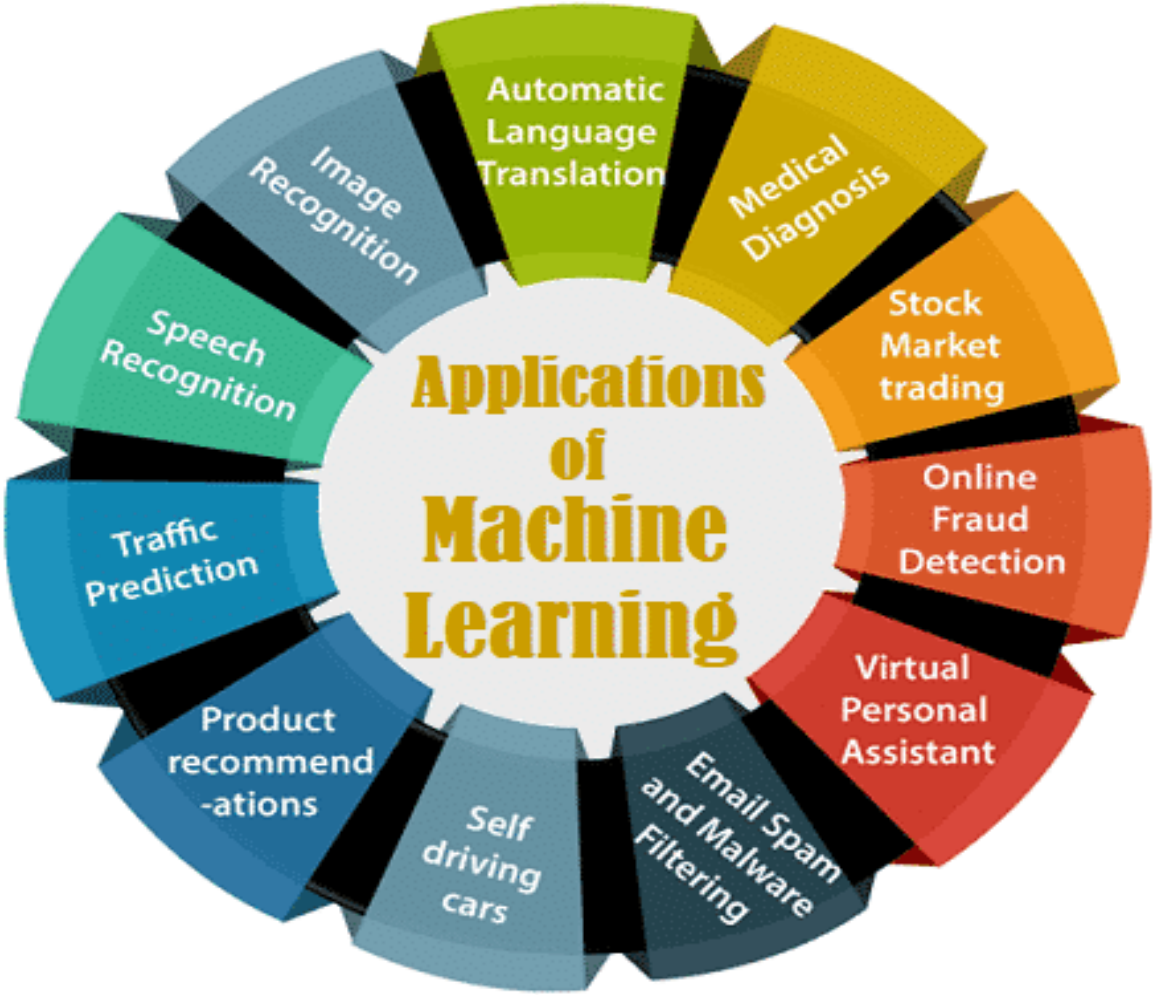
Google Machine Learning

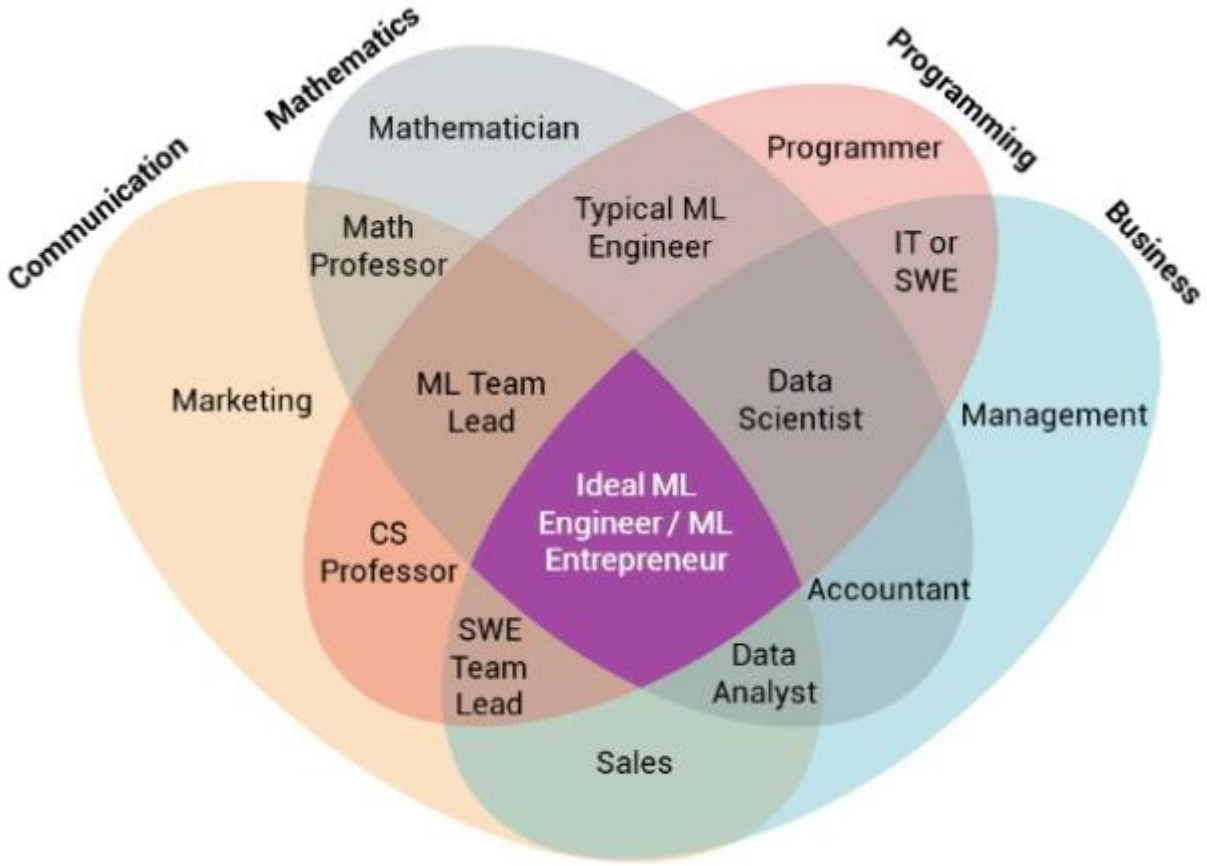


ML Applications-2



ML Applications-3




Communication
Mathematics
Programming
Business
Mathematician
Programmer
Math Professor
Typical ML Engineer
IT or SWE
Marketing
ML Team Lead
Data Scientist
Management
Ideal ML Engineer / ML Entrepreneur
CS Professor
Accountant
SWE Team Lead
Data Analyst
Sales

What is a parameter in a machine learning model?

- A model parameter is a configuration variable that is internal to the model and whose value can be estimated from the given data.
- They are required by the model when making predictions.
- Their values define the skill of the model on your problem.
- They are estimated or learned from historical training data.
- They are often not set manually by the practitioner.
- They are often saved as part of the learned model.

The examples of model parameters include:

- The weights in an artificial neural network.
- The support vectors in a support vector machine.
- The coefficients in linear regression or logistic regression.

Parametric and Non Parametric

- *A learning model that summarizes data with a set of parameters of fixed size (independent of the number of training examples) is called a parametric model.*
- The algorithms involve two steps:
- Select a form for the function.
- Learn the coefficients for the function from the training data.

- An example of this would be the approach used in linear regression algorithms, where the simplified functional form can be something like:
- **$B0+B1*X1+B2*X2=0$**
- This assumption greatly simplifies the learning process; after selecting the initial function, the remaining problem is simply to estimate the coefficients B0, B1, and B2 using different samples of input variables X1 and X2.

- This assumption greatly simplifies the learning process; after selecting the initial function, the remaining problem is simply to estimate the coefficients B0, B1, and B2 using different samples of input variables X1 and X2.

- **Examples of parametric machine learning algorithms** include:
- Logistic Regression
- Linear Discriminant Analysis
- Perceptron
- Naive Bayes
- Simple Neural Networks

Benefits:

- **Simpler** and **easier** to understand; easier to interpret the results
- **Faster** while learning from data
- **Less training data** required to learn the mapping function, working well even if the fit to data is not perfect

Limitations:

- **Highly constrained** to the specified form of the simplified function
- **Limited complexity** of the problems they are suitable for
- **Poor fit** in practice, unlikely to match the underlying mapping function.

Non-parametric Machine Learning Algorithms

- Non-parametric algorithms do not make assumptions regarding the form of the mapping function between input data and output. Consequently, they are free to learn any functional form from the training data.
- A simple example is the K-nearest neighbors (KNN) algorithm, KNN does not make any assumptions about the functional form, but instead uses the pattern that points have similar output when they are close.

- **Examples of popular nonparametric machine learning algorithms** are:
- k-Nearest Neighbors
- Decision Trees like CART and C4.5
- Support Vector Machines

- Consider decision tree algorithms.
- If we increase the number of instances, then the decision tree that is going to be built becomes more complex.
- The more decision rules could be created based on those new instances inherently.

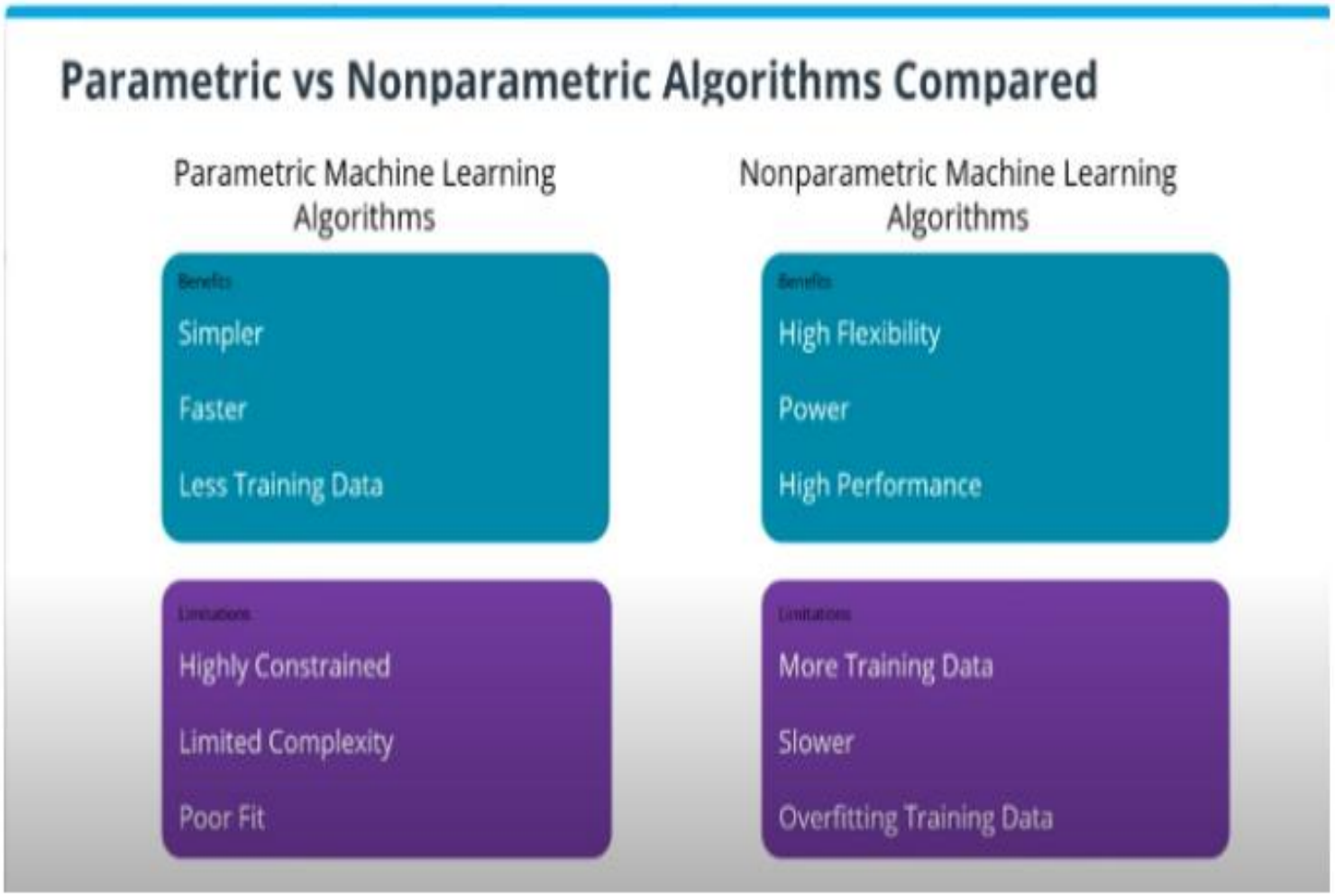
Benefits:

- **High flexibility**, in the sense that they are capable of fitting a large number of functional forms
- **Power** by making weak or no assumptions on the underlying function
- **High performance** in the prediction models that are produced

Limitations:

- **More training data** is required to estimate the mapping function
- **Slower** to train, generally having far more parameters to train
- **Overfitting** the training data is a risk; overfitting makes it harder to explain the resulting predictions

Comparison



Picture from: Udacity course

- Parametric methods make large assumptions about the mapping of the input variables to the output variable and in turn are faster to train, require less data but may not be as powerful.
- Nonparametric methods make few or no assumptions about the target function and in turn require a lot more data, are slower to train and have a higher model complexity but can result in more powerful models.

Daily Assessment

- What is Machine learning?

  a) The autonomous acquisition of knowledge through the use of computer programs

  b) The autonomous acquisition of knowledge through the use of manual programs

  c) The selective acquisition of knowledge through the use of computer programs

  d) The selective acquisition of knowledge through the use of manual programs

- Answer: a
  Explanation: Machine learning is the autonomous acquisition of knowledge through the use of computer programs.

- In Model based learning methods, an iterative process takes place on the ML models that are built based on various model parameters, called ?

- A. mini-batches
  B. optimizedparameters
  C. hyperparameters
  D. superparameters

- Ans : C
  Explanation: In Model based learning methods, an iterative process takes place on the ML models that are built based on various model parameters, called hyperparameters.

- **What is meant by** ‘Training set’ and ‘Test Set’**?**
- Differentiate between inductive learning and deductive learning?
- What is the difference between Data Mining and Machine Learning?
- Differentiate supervised and unsupervised machine learning.

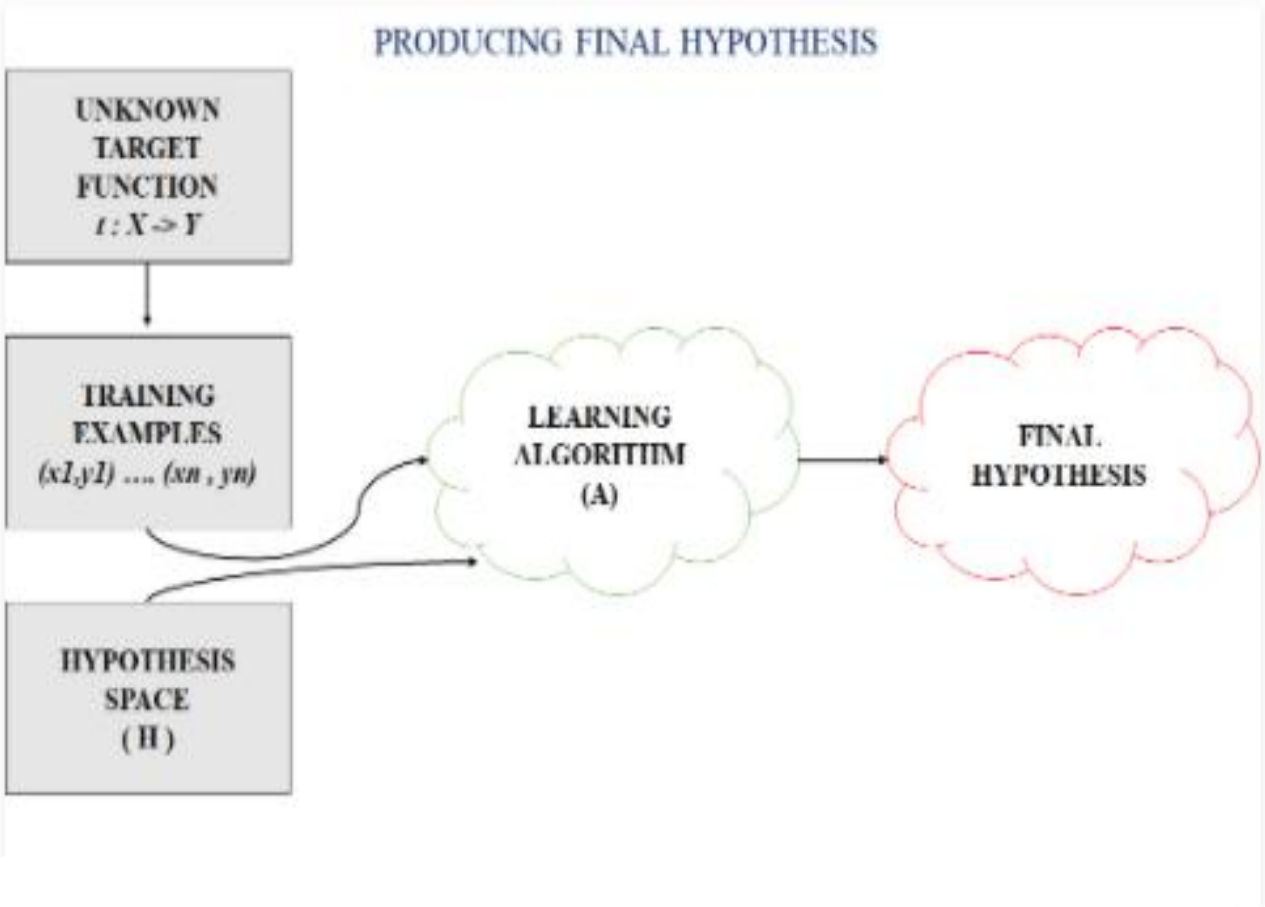
- What are the different types of Algorithm methods in Machine Learning?
- What do you understand by Reinforcement Learning technique?
- What is the trade-off between bias and variance?

- What are the common ways to handle missing data in a dataset?
- What do you understand by ILP? Inductive Logic Programming
- Is it a part of ML
- Describe the classifier in machine learning

Hypothesis Space and Inductive Bias

- A **hypothesis** **is an assumption,** an idea that is proposed for the sake of argument so that it can be tested to see if it might be true. ...
- A **hypothesis** is usually tentative; it's an assumption or suggestion made strictly for the objective of being tested.

- A statistical **hypothesis** is an explanation about the relationship between data populations that is interpreted probabilistically. ...
- A **machine learning hypothesis** is a candidate model that approximates a target function for mapping inputs to outputs.


PRODUCING FINAL HYPOTHESIS
UNKNOWN TARGET FUNCTION
t : X -> Y
TRAINING EXAMPLES
(x1,y1) .... (xn , yn)
HYPOTHESIS SPACE
(H)
LEARNING ALGORITHM
(A)
FINAL HYPOTHESIS

- **Hypothesis Space (H):**
  Hypothesis space is the set of all the possible legal hypothesis.

- This is the set from which the machine learning algorithm would determine the best possible (only one) which would best describe the target function or the outputs.

Hypothesis (h):

- A hypothesis is a function that best describes the target in supervised machine learning.
- The hypothesis that an algorithm would come up depends upon the data and also depends upon the restrictions and bias that we have imposed on the data.
- To better understand the Hypothesis Space and Hypothesis consider the following coordinate that shows the distribution of some data:

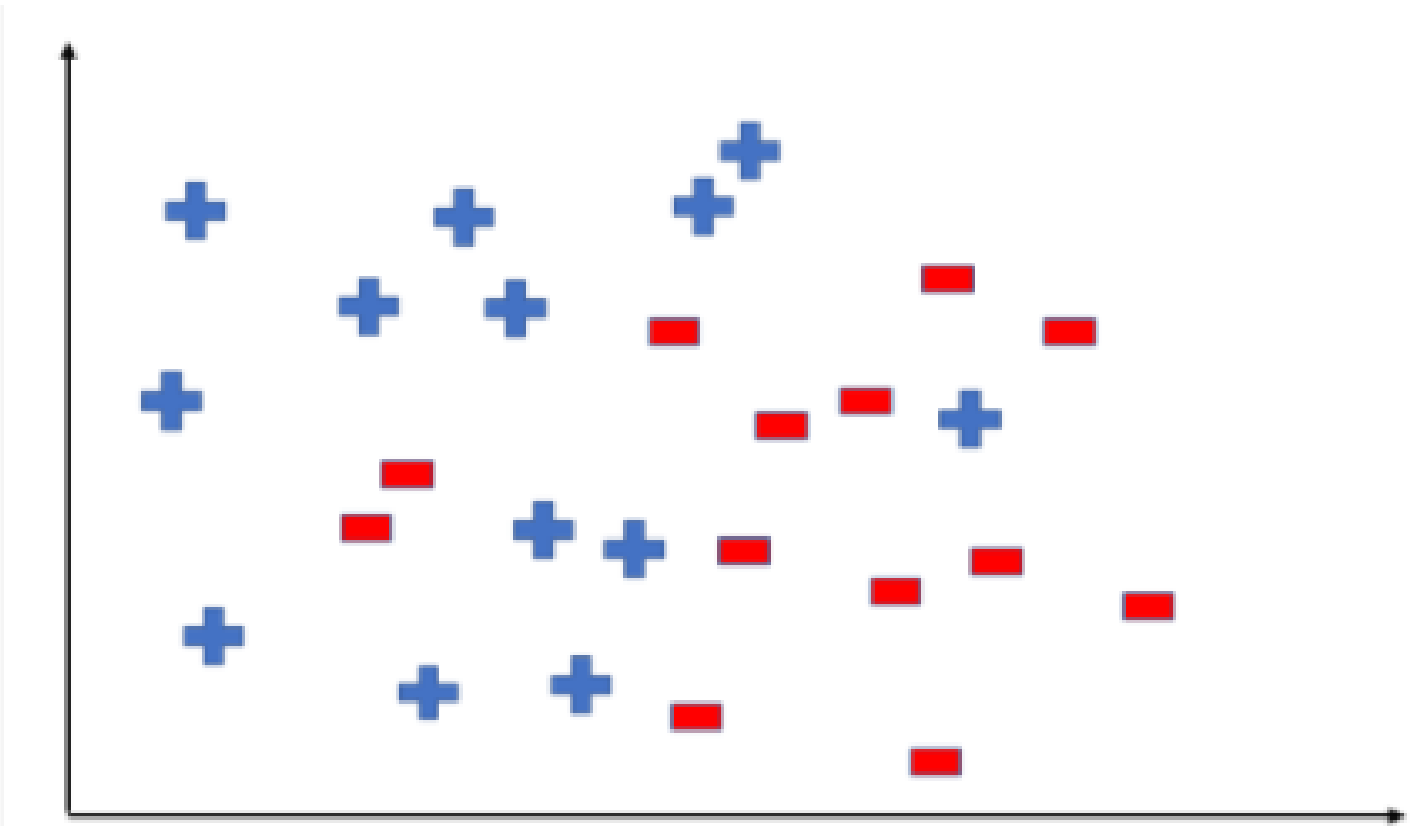
- **Terminology**

  Example (x,y) : Instance x with label y=f(x)
  Training Data s : Collection of examples observed by learning algorithm.
  Instance Space x : Set of all possible objects describable by features.
  Concept c : Subset of objects from X (c is unknown).
  Target Function f : Maps each instance x ε X to target label y ε Y

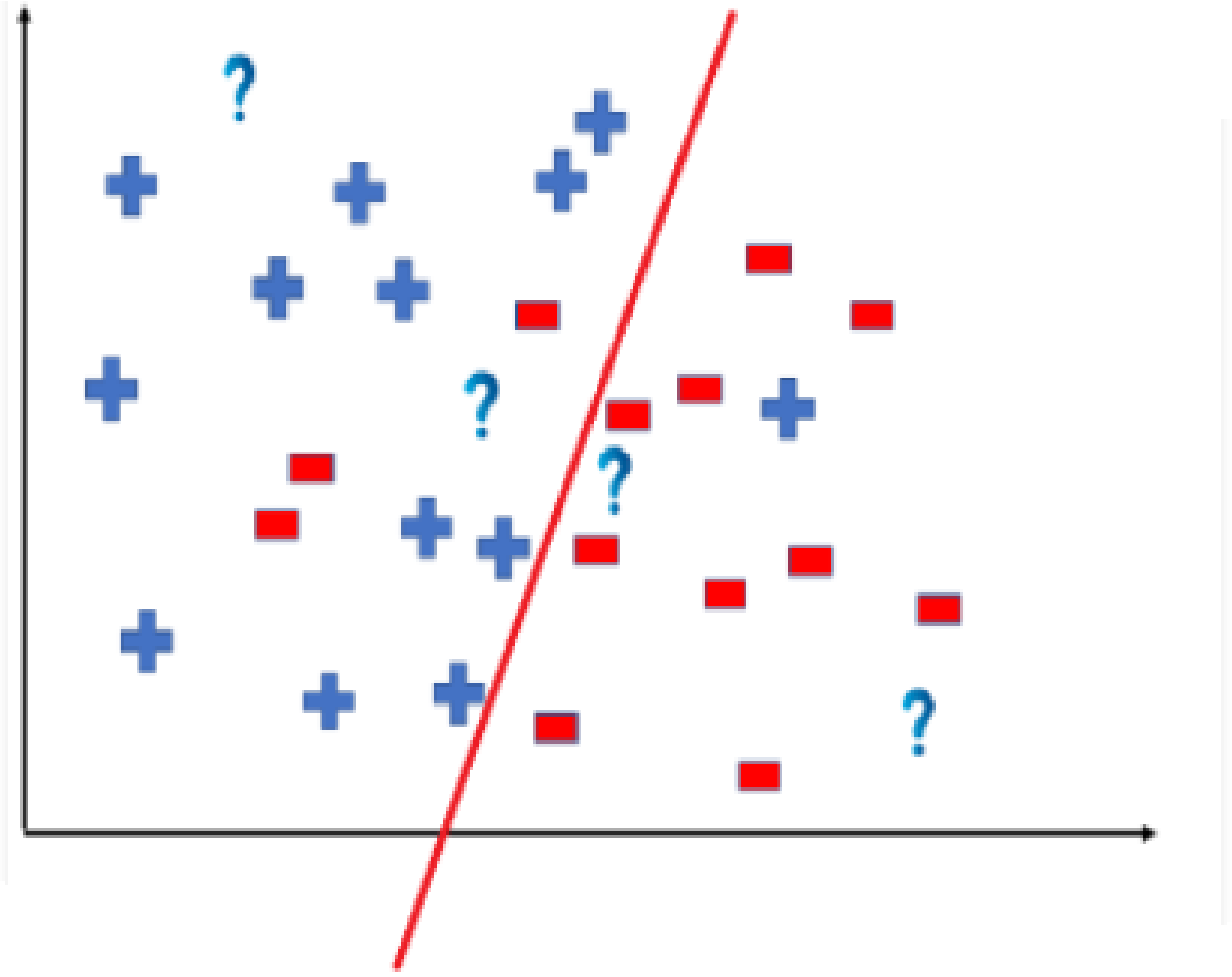
- **Supervised learning machine** explores a "**hypothesis space**".

Classifier

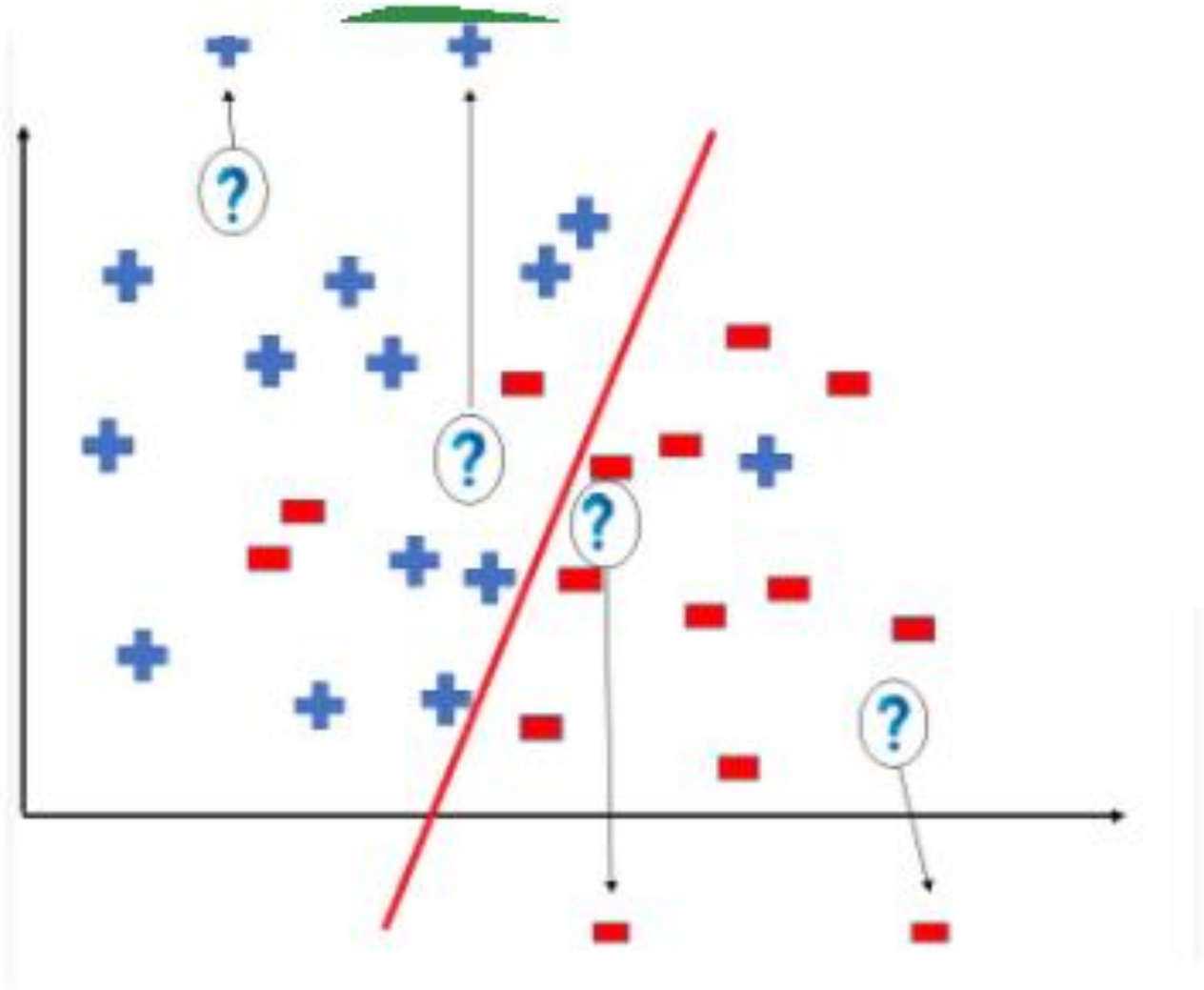
Hypothesis h : Function that approximates f.
Hypothesis Space H : Set of functions we allow for approximating f.
The set of hypotheses that can be produced, can be restricted further by specifying a language bias.
Input : Training set $S \subseteq X$
Output : A hypothesis $h \subseteq H$



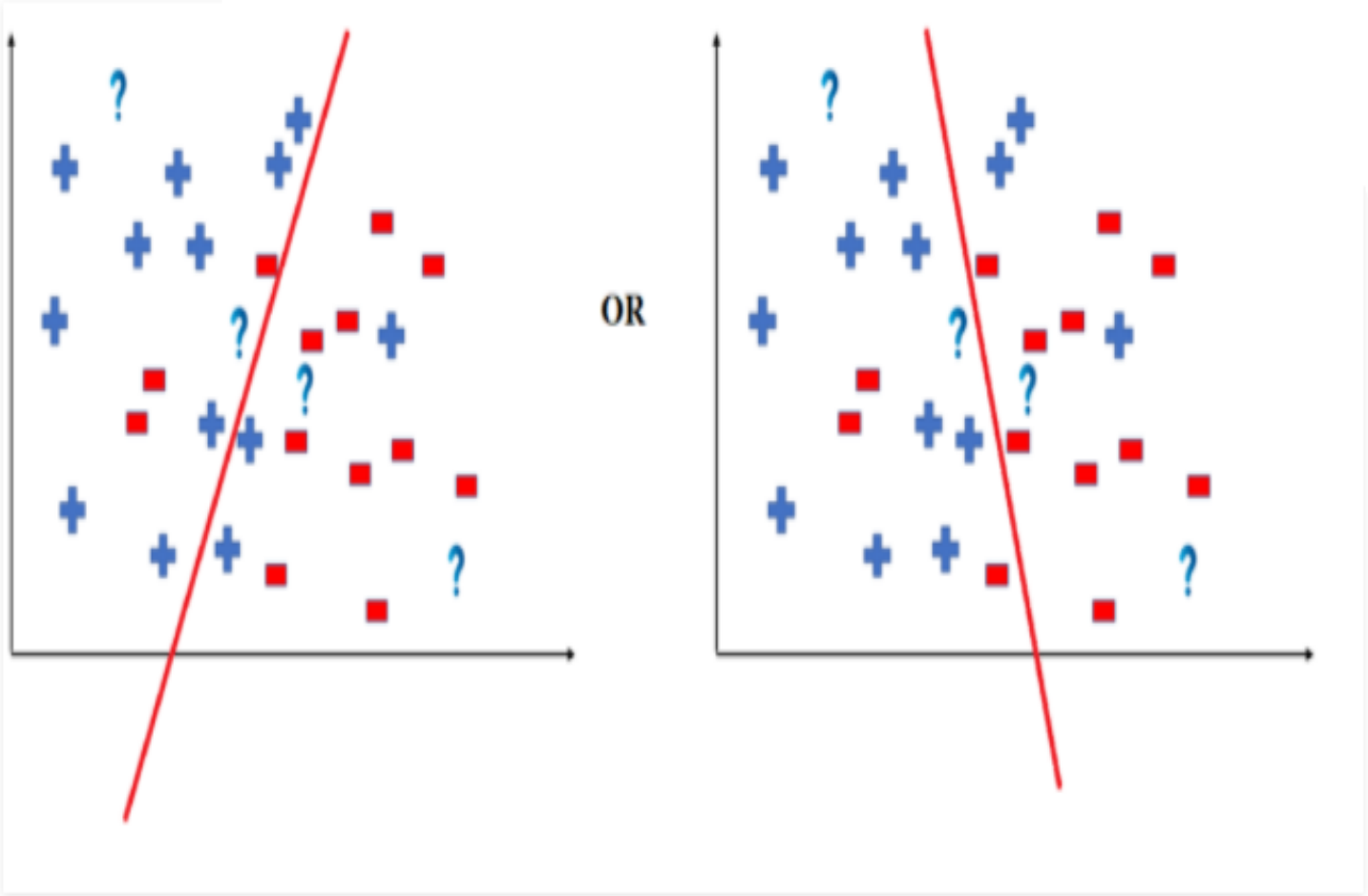
We can predict the outcomes by dividing the coordinate as shown below:



So the test data would yield the following result:

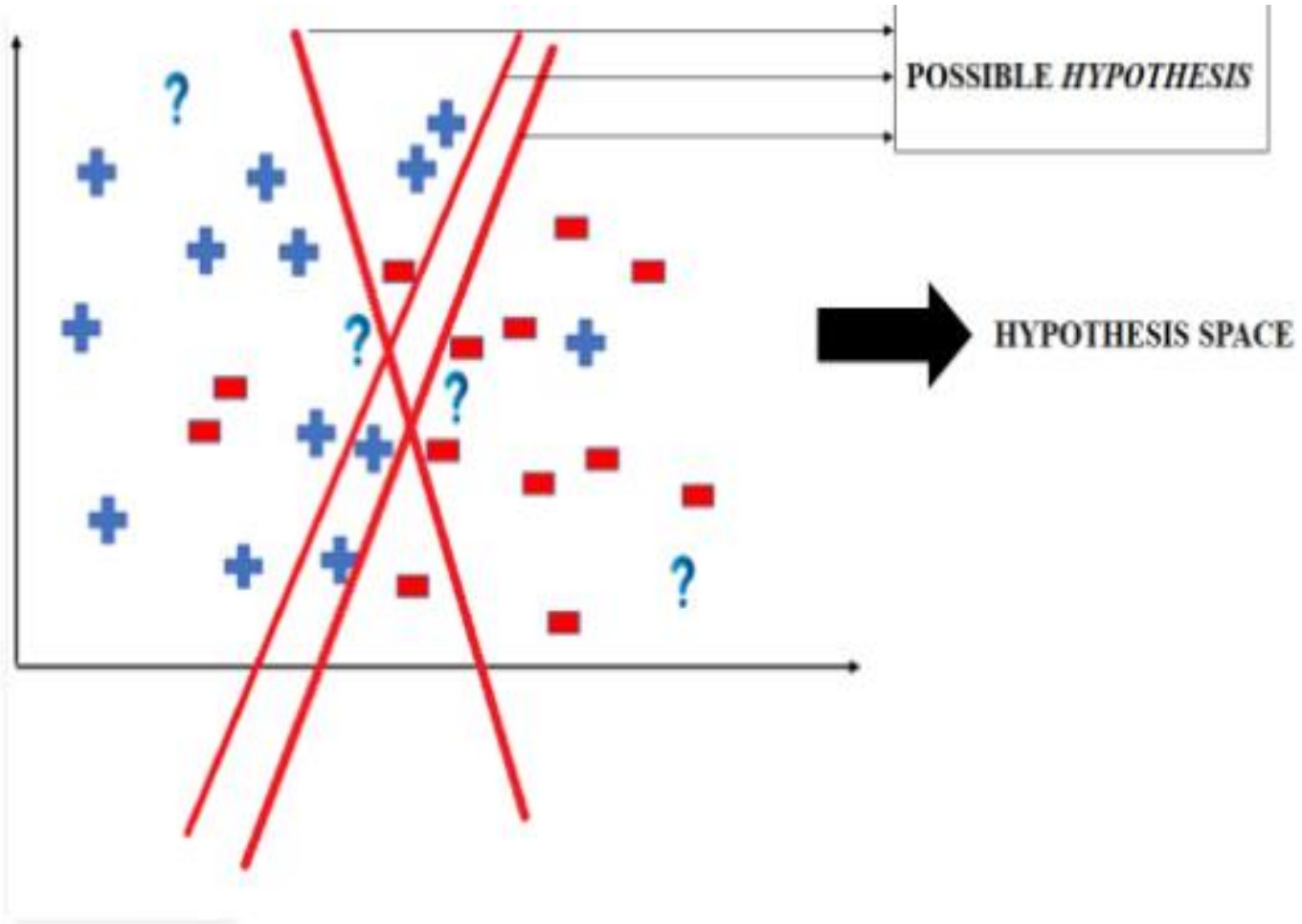


we could have divided the coordinate plane as:



- The way in which the coordinate would be divided depends on the **data, algorithm and constraints**.
- All these legal possible ways in which we can divide the coordinate plane to predict the outcome of the test data composes of the Hypothesis Space.
- Each **individual possible way is known as the hypothesis.**

Hence, in this example the hypothesis space would be like:



Note

- A good hypothesis is testable; it can be either true or false.
- A good hypothesis fits the evidence and can be used to make predictions about new observations or new situations.
- The hypothesis that best fits the evidence and can be used to make predictions is called a theory, or is part of a theory.

- If the null hypothesis is rejected, then we assume the alternative hypothesis that there exists some difference between the means.
- **Null Hypothesis (H0)**: Suggests no effect.
- **Alternate Hypothesis (H1)**: Suggests some effect.

Already discussed

- **h (*hypothesis*)**: A single hypothesis, e.g. an instance or specific candidate model that maps inputs to outputs and can be evaluated and used to make predictions.
- **H (*hypothesis set*)**: A space of possible hypotheses for mapping inputs to outputs that can be searched, often constrained by the choice of the framing of the problem, the choice of model and the choice of model configuration.

- We can summarize the three definitions again as follows:
- **Hypothesis in Science**: Provisional explanation that fits the evidence and can be confirmed or disproved.
- **Hypothesis in Statistics**: Probabilistic explanation about the presence of a relationship between observations.
- **Hypothesis in Machine Learning**: Candidate model that approximates a target function for mapping examples of inputs to outputs

Bias

- Bias is one type of error which occurs due to wrong assumptions about data such as assuming data is linear when in reality, data follows a complex function.
- On the other hand, variance gets introduced with high sensitivity to variations in training data.
- This also is one type of error since we want to make our model robust against noise.

- Let’s say, f(x) is the function which our given data follows.
- We will build few models which can be denoted as .

- Each point on this function is a random variable having number of values equal to number of models.
- To correctly approximate the true function f(x), we take expected value of .

$$\hat{f}(x) : E\left[\hat{f}(x)\right].$$

Bias : $f - E[\hat{f}]$

Variance : $E[\hat{f^2}] - E[\hat{f}]] = E[(\hat{f} - E[\hat{f}])^2]$

- A **low bias** model incorporates fewer assumptions about the target function.
- A linear algorithm often has high **bias**, which makes them learn fast.
- In linear regression analysis, **bias** refers to the error that is introduced by approximating a real-life problem, which may be complicated, by a much simpler model.

Inductive learning

- **Inductive Learning :** Inducing a general function from training examples
  - Construct hypothesis h to agree with c on the training example.
  - A hypothesis is consistent if it agrees with all training examples.
  - A hypothesis is said to generalize well if it correctly predicts the value of y for novel example

- Evaluation. Training and test sets, cross validation.

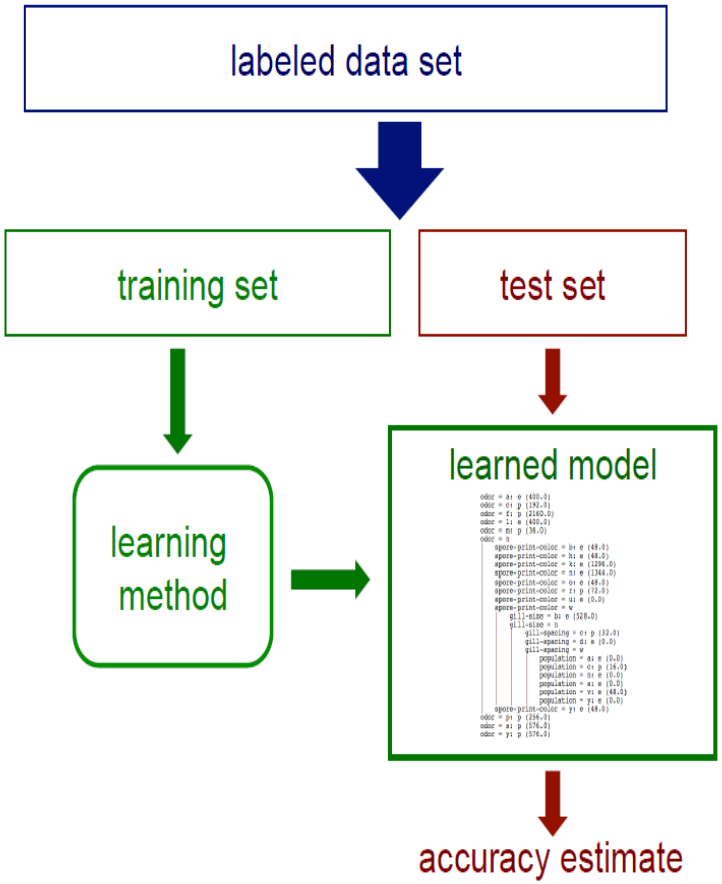
Evaluating Machine Learning Methods

Today, We shall see the following concepts

- test sets
- learning curves
- validation (tuning) sets
- stratified sampling
- cross validation
- internal cross validation
- confusion matrices
- TP, FP, TN, FN

Test sets

How can we get an unbiased estimate of the accuracy of a learned model?



- when learning a model, you should pretend that you don't have the test data yet (it is "in the mail")*
- • if the test-set labels influence the learned model in any way, accuracy estimates will be biased

* In some applications it is reasonable to assume that you have access to the feature vector (i.e. ***x***) but not the *y* part of each test instance.

Learning curves

How does the accuracy of a learning method change as a function of the training-set size?

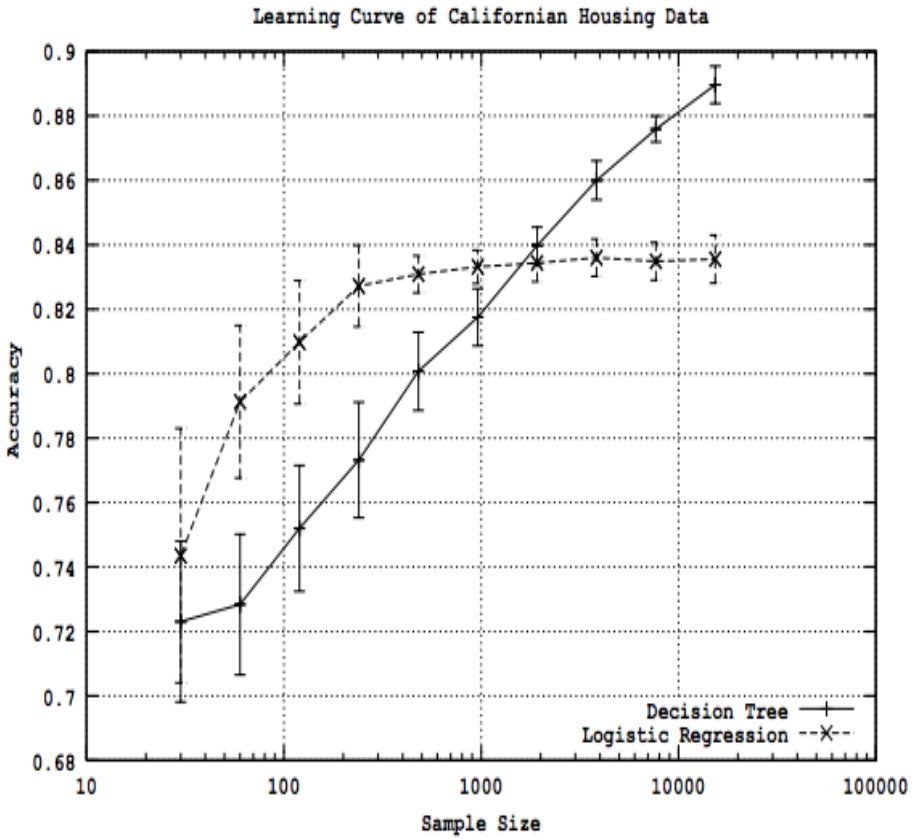


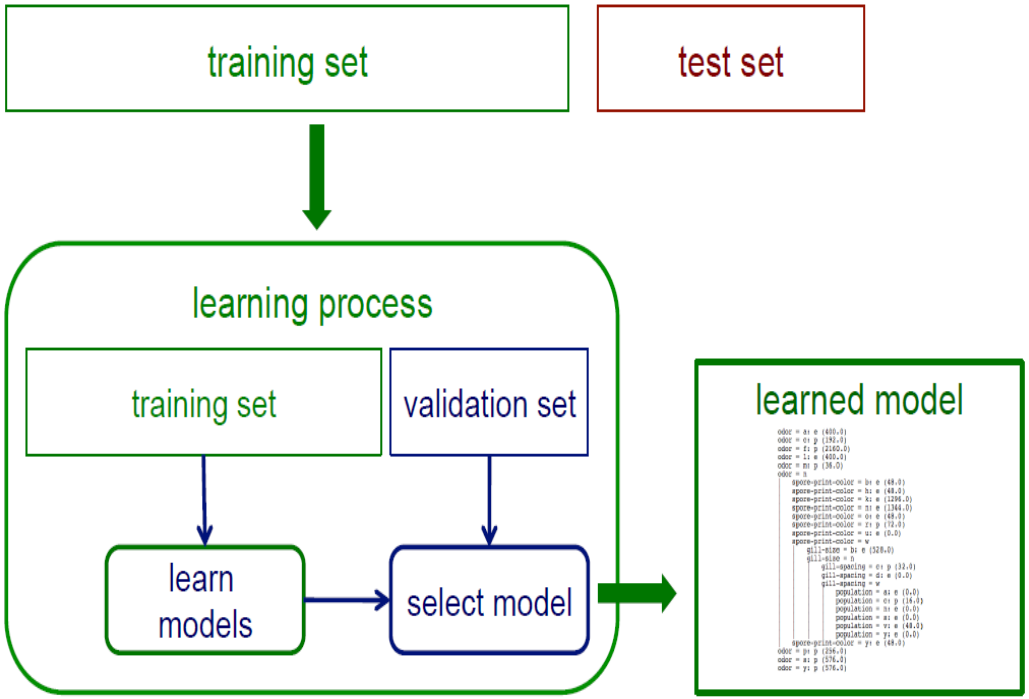
Figure from Perlich et al. *Journal of Machine Learning Research*, 2003

given training/test set partition

- for each sample size s on learning curve
  - (optionally) repeat n times
    - randomly select s instances from training set
    - learn model
    - evaluate model on test set to determine accuracy a
    - plot (s, a) or (s, avg. accuracy and error bars)

Validation (tuning) sets

Suppose we want unbiased estimates of accuracy during the learning process (e.g. to choose the best level of decision-tree pruning or find the best parameters in deep learning)?



Partition training data into separate training/validation sets

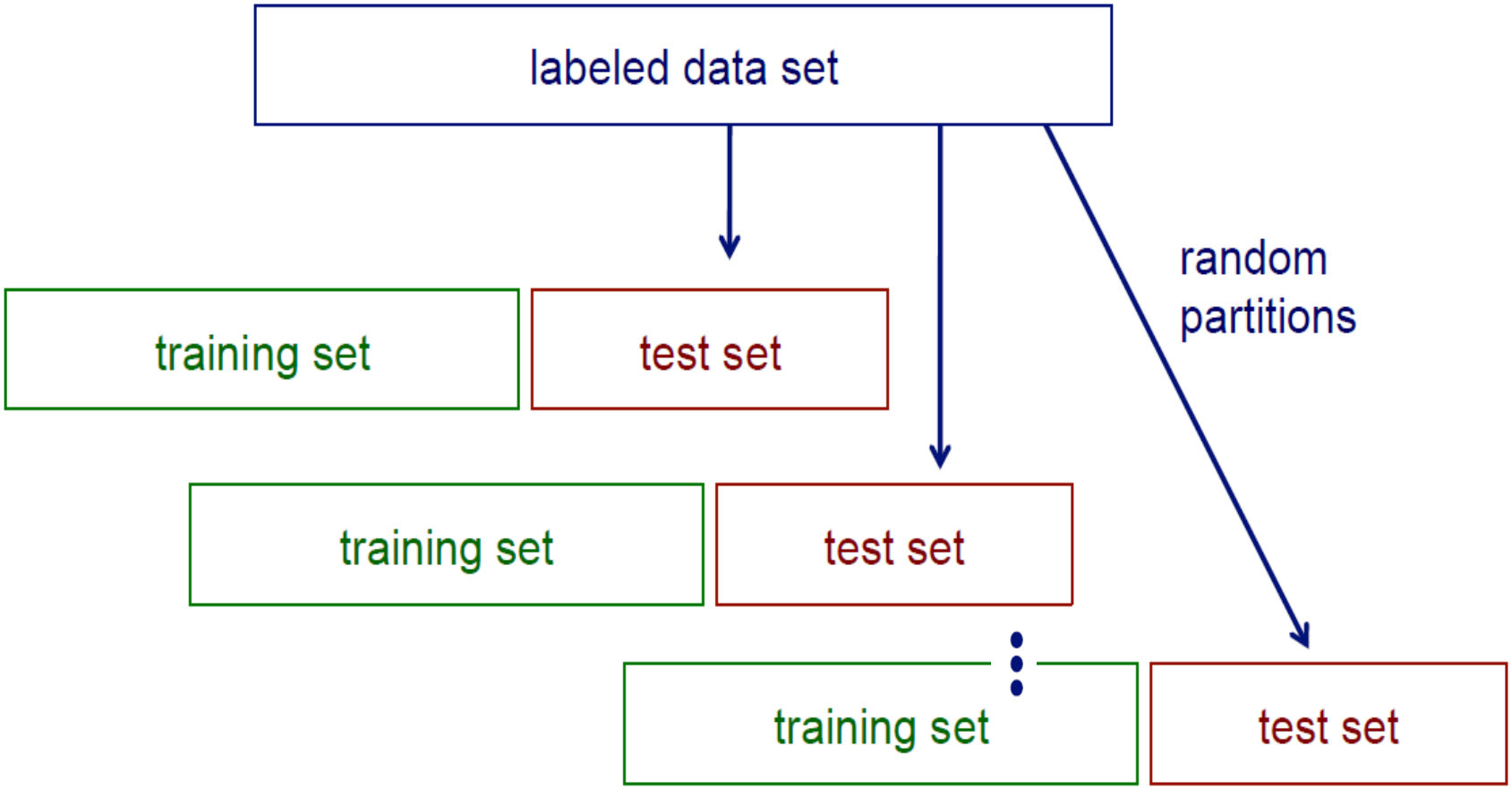
Limitations of using a single training/test partition

- we may not have enough data to make sufficiently large training and test sets
  - a larger test set gives us more reliable estimate of accuracy (i.e. a lower variance estimate)
  - but... a larger training set will be more representative of how much data we actually have for learning process

- a single training set doesn't tell us how sensitive accuracy is to a particular training sample

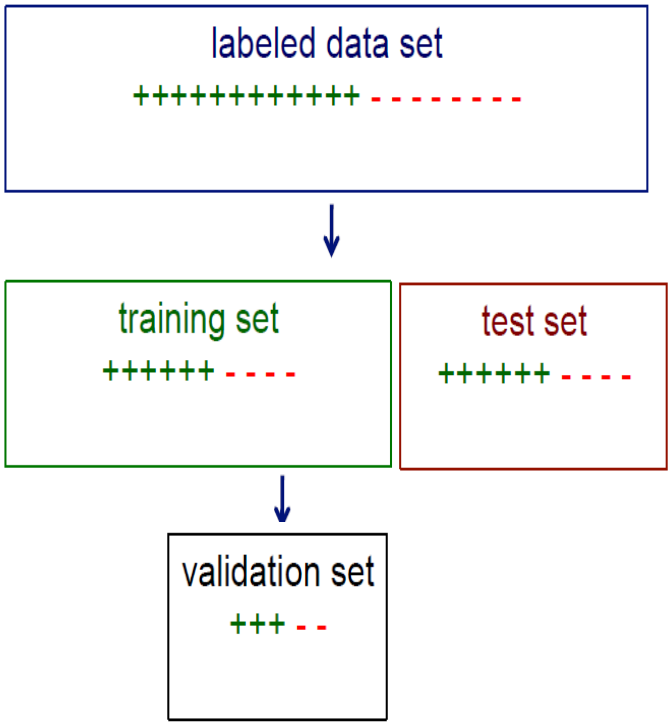
Random resampling

We can address the second issue by repeatedly randomly partitioning the available data into training and set sets.



Stratified sampling

When randomly selecting training or validation sets, we may want to ensure that class proportions are maintained in each selected set

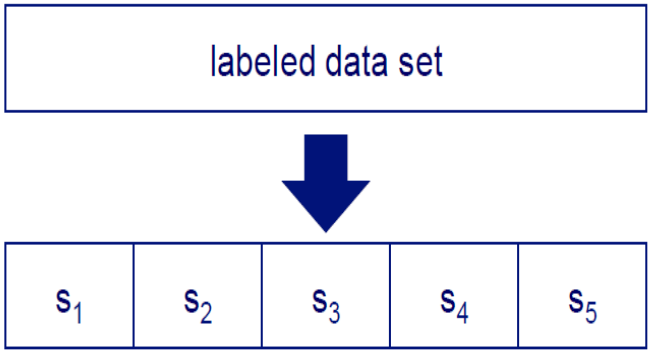


This can be done via stratified sampling: first stratify instances by class, then randomly select instances from each class proportionally.

Cross validation

partition data

into *n* subsamples



iteratively leave **one subsample out** for the test set, train on the rest

iteration	train on	test on
1	s_2 s_3 s_4 s_5	s_1
2	s_1 s_3 s_4 s_5	s_2
3	s_1 s_2 s_4 s_5	s_3
4	s_1 s_2 s_3 s_5	s_4
5	s_1 s_2 s_3 s_4	s_5

Cross validation example

Suppose we have 100 instances, and we want to estimate accuracy with cross validation

iteration	train on	test on	correct
1	S_2 S_3 S_4 S_5	S_1	11 / 20
2	S_1 S_3 S_4 S_5	S_2	17 / 20
3	S_1 S_2 S_4 S_5	S_3	16 / 20
4	S_1 S_2 S_3 S_5	S_4	13 / 20
5	S_1 S_2 S_3 S_4	S_5	16 / 20

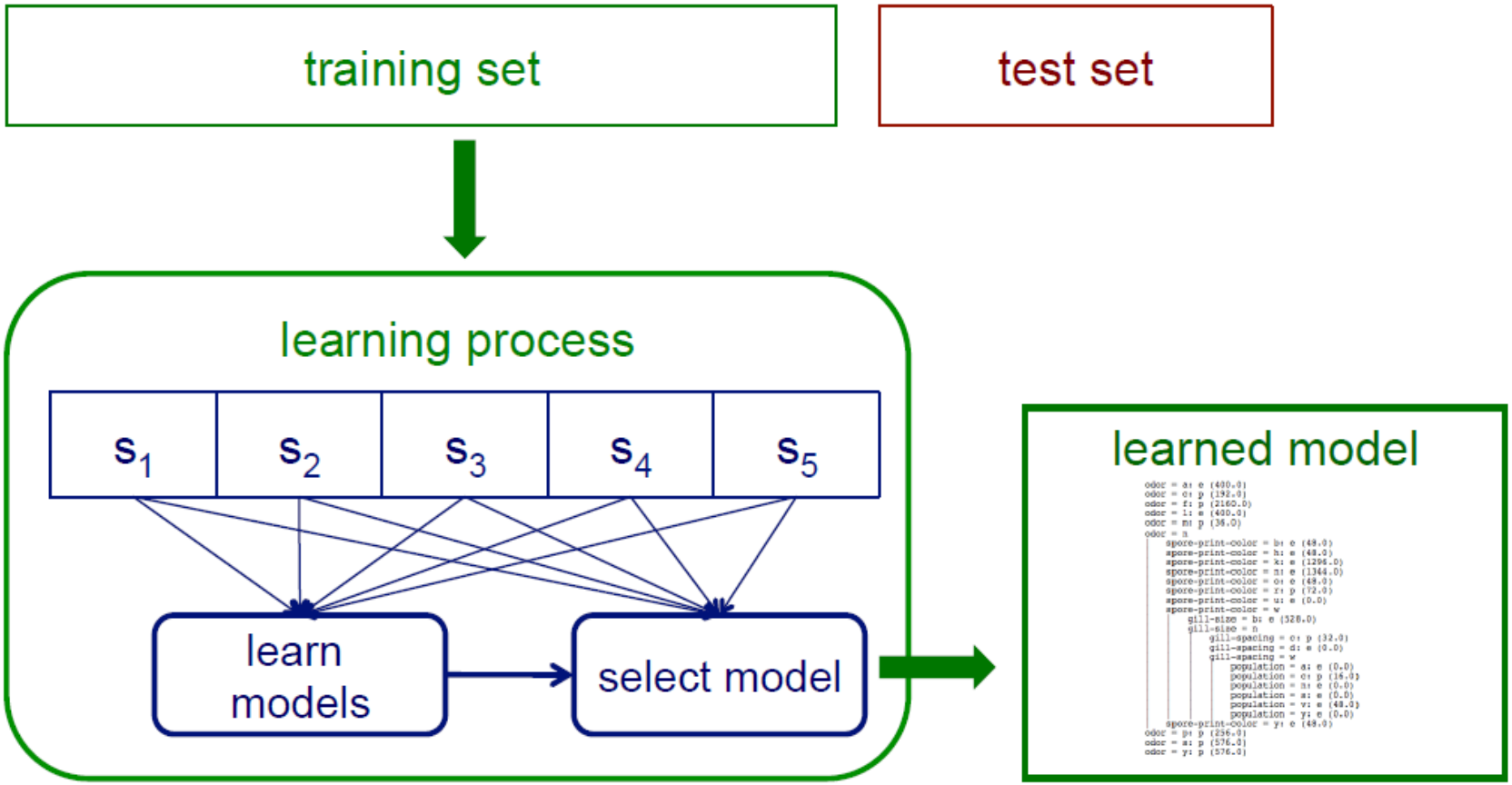
accuracy = 73/100 = 73%

Cross validation

- 10-fold cross validation is common, **but smaller values of *n*** are often used when learning takes a lot of time
- in *leave-one-out* cross validation, *n* = # instances
- in *stratified* cross validation, stratified sampling is used when partitioning the data
- **CV** makes efficient use of the available data for testing
- note that whenever we use multiple training sets, as in CV and random resampling, we are evaluating a learning method as opposed to an individual learned model

Internal cross validation

Instead of a single validation set, we can use **cross-validation** within a training set to select a model (e.g. to choose the best level of decision-tree pruning or find the optimal parameters of a deep learning model)



Example: using internal cross validation to select k in k-NN

given a training set

1. partition training set into n folds, $s_1 \ldots s_n$
2. for each value of k considered

   for $i = 1$ to n

   learn k-NN model using all folds but s_i

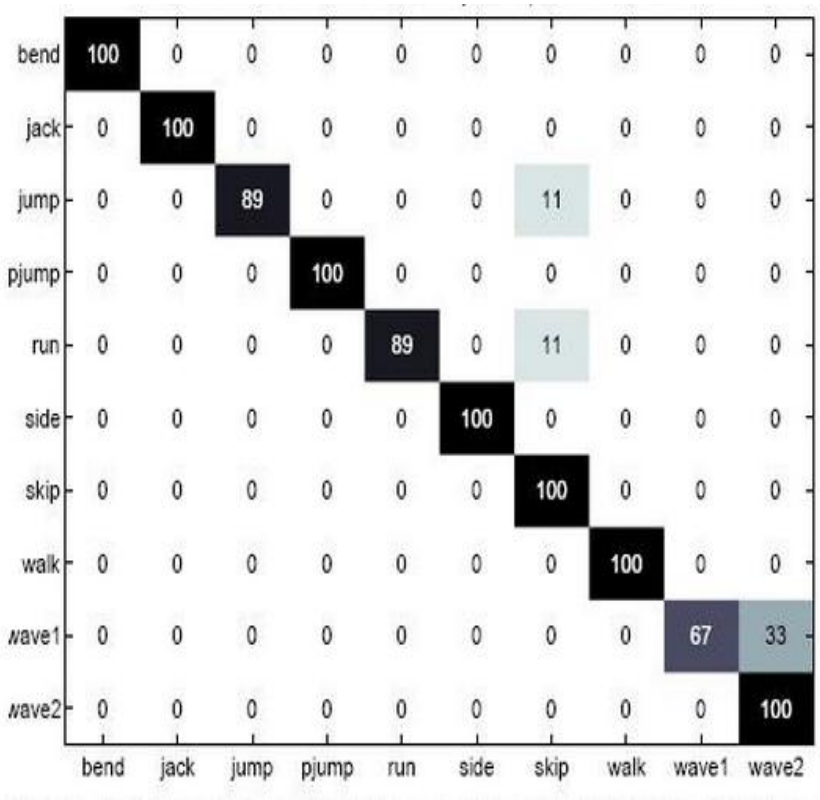
   evaluate accuracy on s_i
3. select k that resulted in best accuracy for $s_1 \ldots s_n$
4. learn model using entire training set and selected k

the steps inside the box are run independently for each training set (i.e. if we're using 10-fold CV to measure the overall accuracy of our k-NN approach, then the box would be executed 10 times)

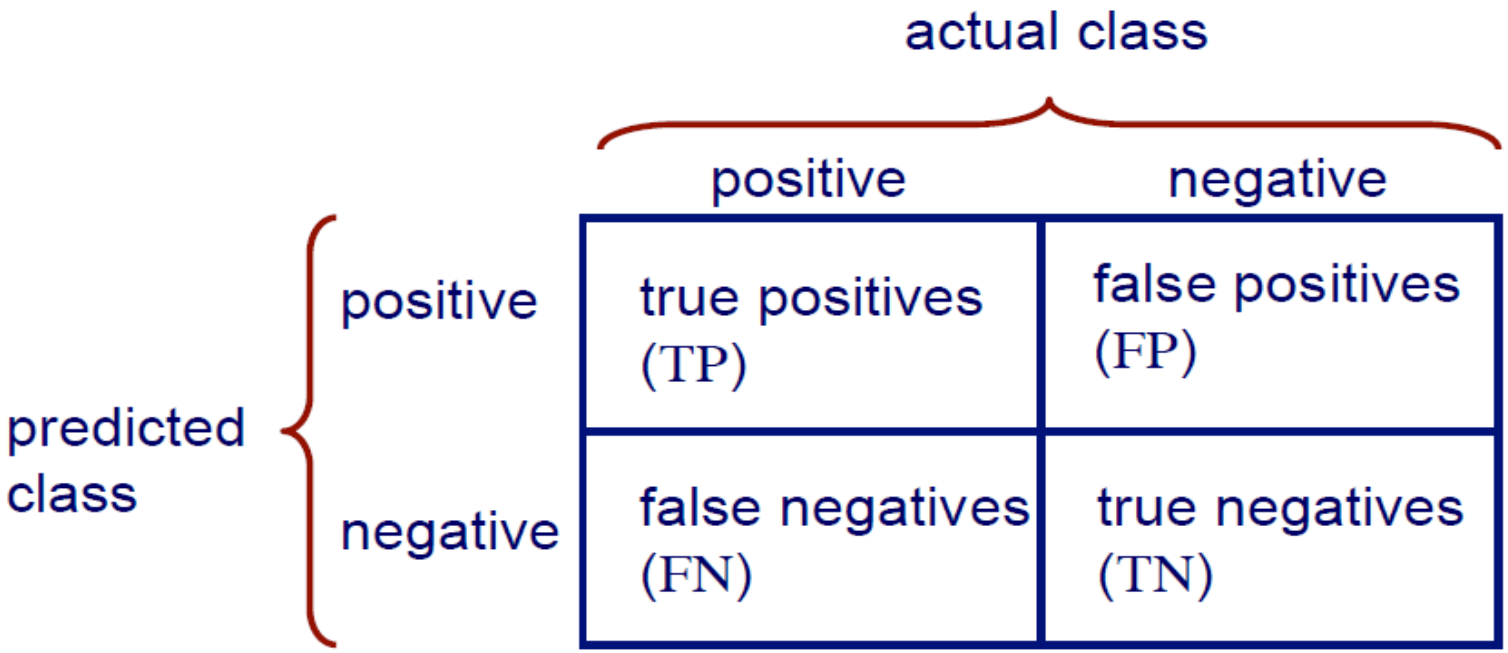
Confusion matrices

How can we understand what types of mistakes a learned model makes?

activity recognition from video



Confusion matrix for 2-class problems

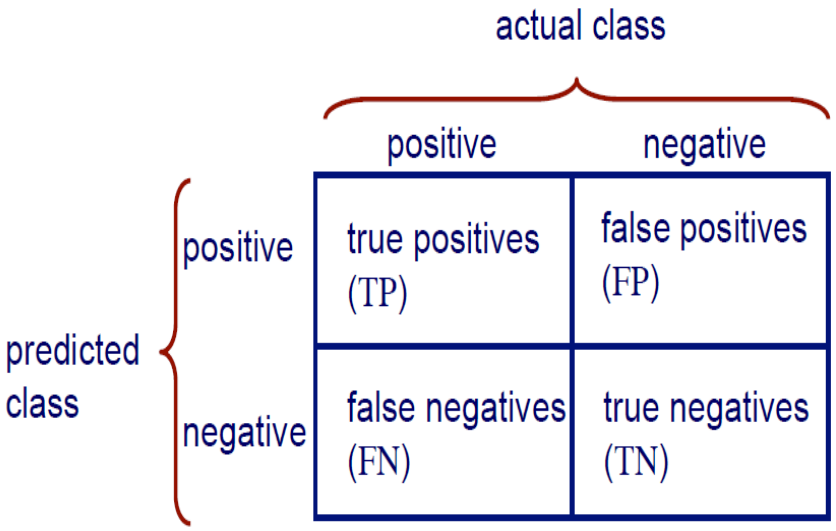


$$\text{accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Is accuracy an adequate measure of predictive performance?

- accuracy may not be useful measure in cases where
  - there is a large class **skew**
    - Is 98% accuracy good if 97% of the instances are negative?
  - there are differential misclassification costs – say, getting a positive wrong costs more than getting a negative wrong
    - Consider a medical domain in which a false positive results in an extraneous test but a false negative results in a failure to treat a disease
- we are most interested in a subset of high-confidence predictions

Other accuracy metrics



$$\text{true positive rate (recall)} = \frac{TP}{\text{actual pos}} = \frac{TP}{TP + FN}$$

$$\text{false positive rate} = \frac{FP}{\text{actual neg}} = \frac{FP}{TN + FP}$$

Others???

Sensitivity?

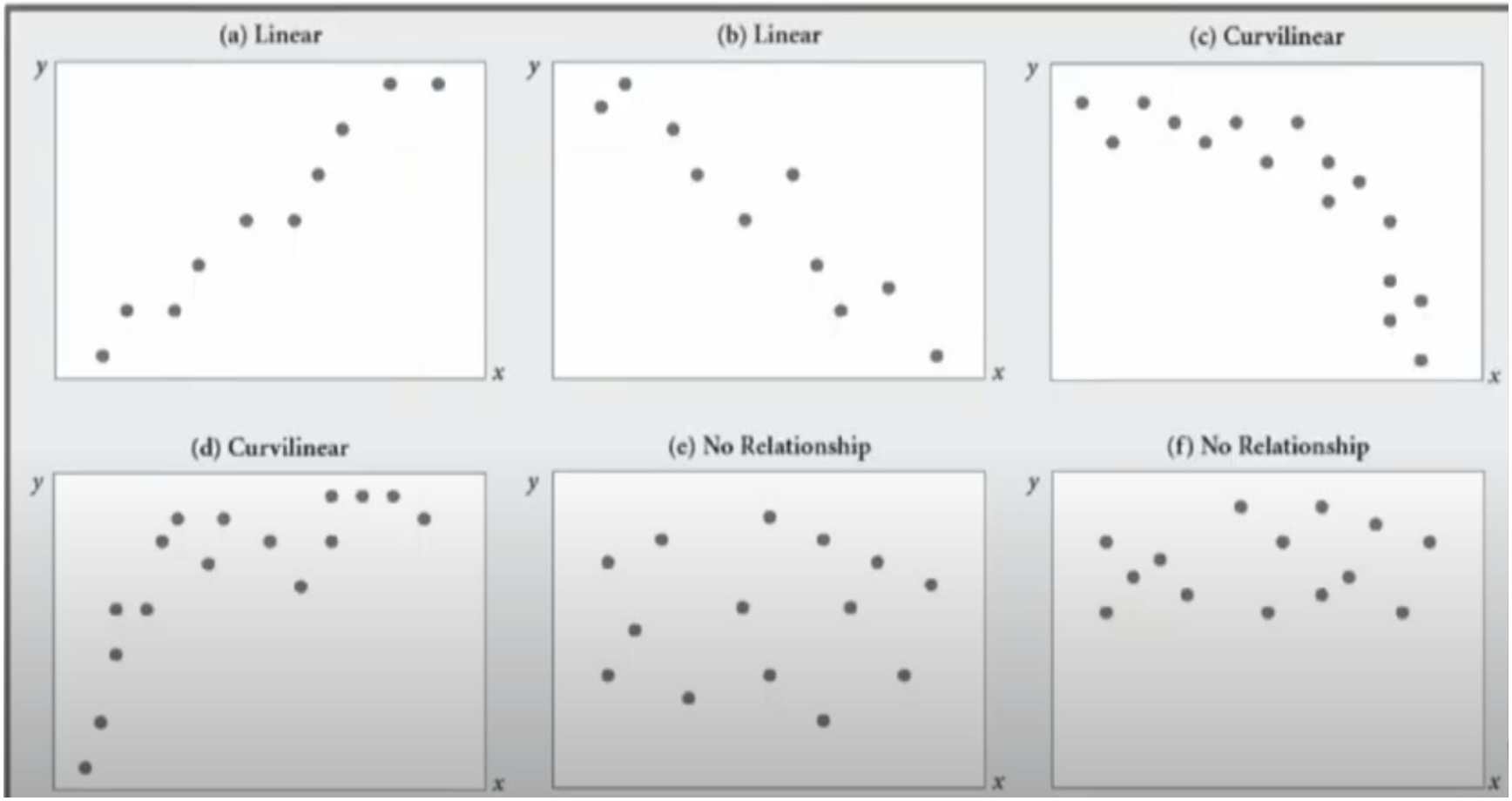
Specificity?

Linear Regression: Introduction, Linear Models for Regression.

Regression Analysis

- Regression is a well-known statistical technique to model the predictive relationship between one or more independent variables (DVs) and one dependent variable.
- The objective is to find the best-fitting curve for a dependent variable in a multidimensional space, with each independent variable being a dimension.
- The curve could be a straight line, or it could be a nonlinear curve.
- The quality of fit of the curve to the data can be measured by a coefficient of correlation (*r*), which is the square root of the amount of variance explained by the curve.

Visual Look at Relationships



Regression Analysis

Regression analysis is used to:

- Predict the value of a dependent variable based on the value of at least one independent variable
- Explain the impact of changes in an independent variable on the dependent variable

Dependent variable: the variable we wish to explain

Independent variable: the variable used to explain the dependent variable

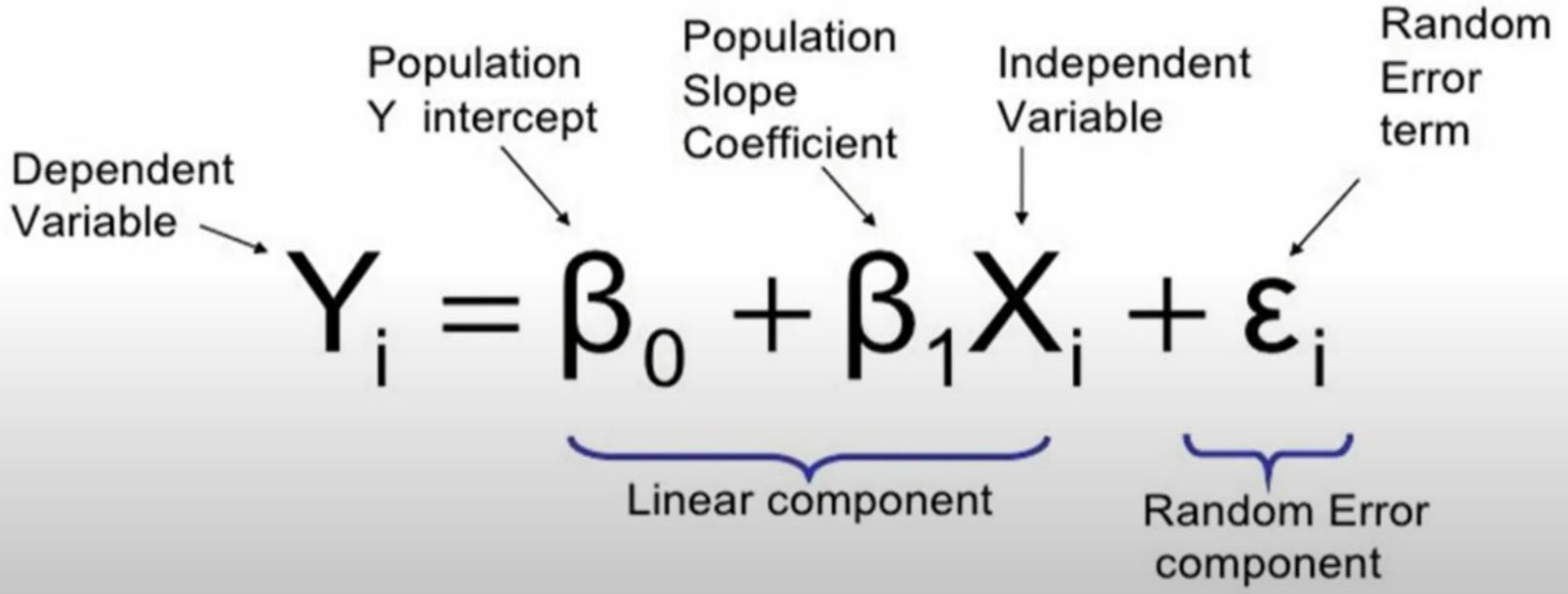
The key steps for regression are simple:

1. List all the variables available for making the model.
2. Establish a dependent variable of interest.
3. Examine visual (if possible) relationships between variables of interest.
4. Find a way to predict dependent variable using the other variables.

- The regression model is described as a linear equation that follows.
- *y* is the dependent variable, that is, the variable being predicted.
- *x* is the independent variable, or the predictor variable.
- There could be many predictor variables (such as *x*1, *x*2, . . .) in a regression equation.
- However, there can be only one dependent variable (*y*) in the regression equation.

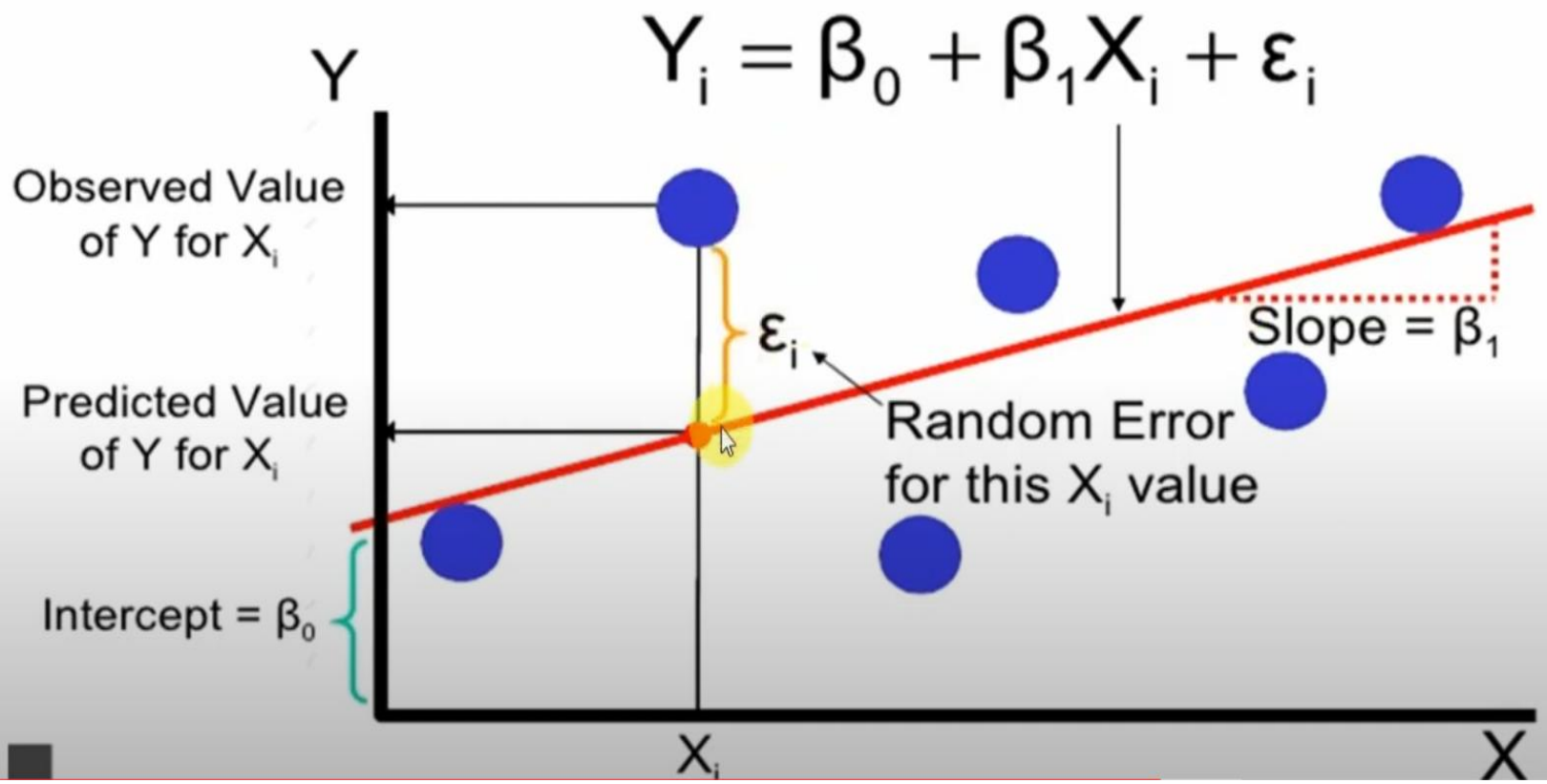
$$y = \beta_0 + \beta_1 x + \varepsilon$$

Simple Linear Regression Model



Simple Linear Regression Model

(continued)



The Multiple Regression Model

Idea: Examine the linear relationship between
1 dependent (Y) & 2 or more independent variables (X_i)

Multiple Regression Model with k Independent Variables:

Y-intercept | Population slopes | Random Error

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

Linear Regression numerical example with multiple independent variables:

Predict the value of Y given X_1 and X_2

SUBJECT	Y	X_1	X_2
1	-3.7	3	8
2	3.5	4	5
3	2.5	5	7
4	11.5	6	3
5	5.7	2	1
6	?	3	2

Linear Regression 2 Independent Variable

$$a = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2$$

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

Linear Regression 2 independent variable

$$\sum x_1^2 = \sum X_1 X_1 - \frac{(\sum X_1)(\sum X_1)}{N}$$

$$\sum x_2^2 = \sum X_2 X_2 - \frac{(\sum X_2)(\sum X_2)}{N}$$

$$\sum x_1 y = \sum X_1 Y - \frac{(\sum X_1)(\sum Y)}{N}$$

$$\sum x_2 y = \sum X_2 Y - \frac{(\sum X_2)(\sum Y)}{N}$$

$$\sum x_1 x_2 = \sum X_1 X_2 - \frac{(\sum X_1)(\sum X_2)}{N}$$

Linear Regression 2 independent variable

SUBJECT	Y	X_1	X_2	X_1X_1	X_2X_2	X_1X_2	X_1Y	X_2Y
1	-3.7	3	8	9	64	24	-11.1	-29.6
2	3.5	4	5	16	25	20	14	17.5
3	2.5	5	7	25	49	35	12.5	17.5
4	11.5	6	3	36	9	18	69	34.5
5	5.7	2	1	4	1	2	11.4	5.7
$\sum$	19.5	20	24	90	148	99	95.8	45.6

$$\sum x_1^2 = \sum X_1 X_1 - \frac{(\sum X_1)(\sum X_1)}{N} = 10$$

$$\sum x_2^2 = \sum X_2 X_2 - \frac{(\sum X_2)(\sum X_2)}{N} = 32.8$$

$$\sum x_1 y = \sum X_1 Y - \frac{(\sum X_1)(\sum Y)}{N} = 17.8$$

$$\sum x_2 y = \sum X_2 Y - \frac{(\sum X_2)(\sum Y)}{N} = -48$$

$$\sum x_1 x_2 = \sum X_1 X_2 - \frac{(\sum X_1)(\sum X_2)}{N} = 3$$

Linear Regression 2 Independent Variable

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{32.8 * 17.8 - 3 * (-48)}{10 * 32.8 - 3 * 3} = 2.28$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{10 * (-48) - 3 * 17.8}{10 * 32.8 - 3 * 3} = -1.67$$

$$a = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2 = \frac{19.5}{5} - \frac{2.28 * 20}{5} - \frac{-1.67 * 24}{5} = 2.796$$

Final Regression equation or Model is:

$$\boldsymbol{Y = 2.796 + 2.28\, x_1 - 1.67 x_2}$$

Now given $x_1 = 3$ *and* $x_2 = 2$ $Y = ?$

$$Y = 2.796 + 2.28 * 3 - 1.67 * 2$$
$$= \mathbf{6.296}$$

Model Evaluation:

- Model evaluation is very important in data science or ML. It helps you to understand the performance of your model and makes it easy to present your model to other people. There are many different evaluation metrics out there but only some of them are suitable to be used for regression.

There are 3 main metrics for model evaluation in regression:

1. R Square/Adjusted R Square
2. Mean Square Error (MSE)/Root Mean Square Error (RMSE)
3. Mean Absolute Error (MAE)

- R Square/Adjusted R Square
- R Square measures how much of variability in dependent variable can be explained by the model. It is square of Correlation Coefficient(R) and that is why it is called R Square.

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

R Square is calculated by the sum of squared of prediction error divided by the total sum of square which replace the calculated prediction with mean. R Square value is between0 to 1 and bigger value indicates a better fit between prediction and actual value. R Square is a good measure to determine how well the model fits the dependent variables. However, it does not take into consideration of overfitting problem. If your regression model has many independent variables, because the model is too complicated, it may fit very well to the training data but performs badly for testing data. That is why Adjusted R Square is introduced because it will penalise additional independent variables added to the model and adjust the metric to prevent overfitting issue.