Basics: Machine Learning

Week 4

Dr. Nouman M Durrani

Acknowledgement to all authors whose materials have been used

Contents of this Week

- Introduction to Machine Learning
    - ML Examples
- What is a Model?
- Types of Machine Learning

    (a) Supervised Learning
    - Examples
    - Workflow
    - Classification
    - Regression
    - Examples

    (b) Unsupervised Learning
    - Examples
    - Workflow
    - Clustering
    - Association
    - Examples

    (c) Reinforcement Learning
    - Examples

- The clustering Problem
- **k-means Clustering**
- k-means Clustering Solved Example
- Examples
- k-means++
- **kNN Classification**
- kNN Classification Solved Example

Week 9

- Hierarchical clustering
- Naive Bayes Classifier

Machine Learning

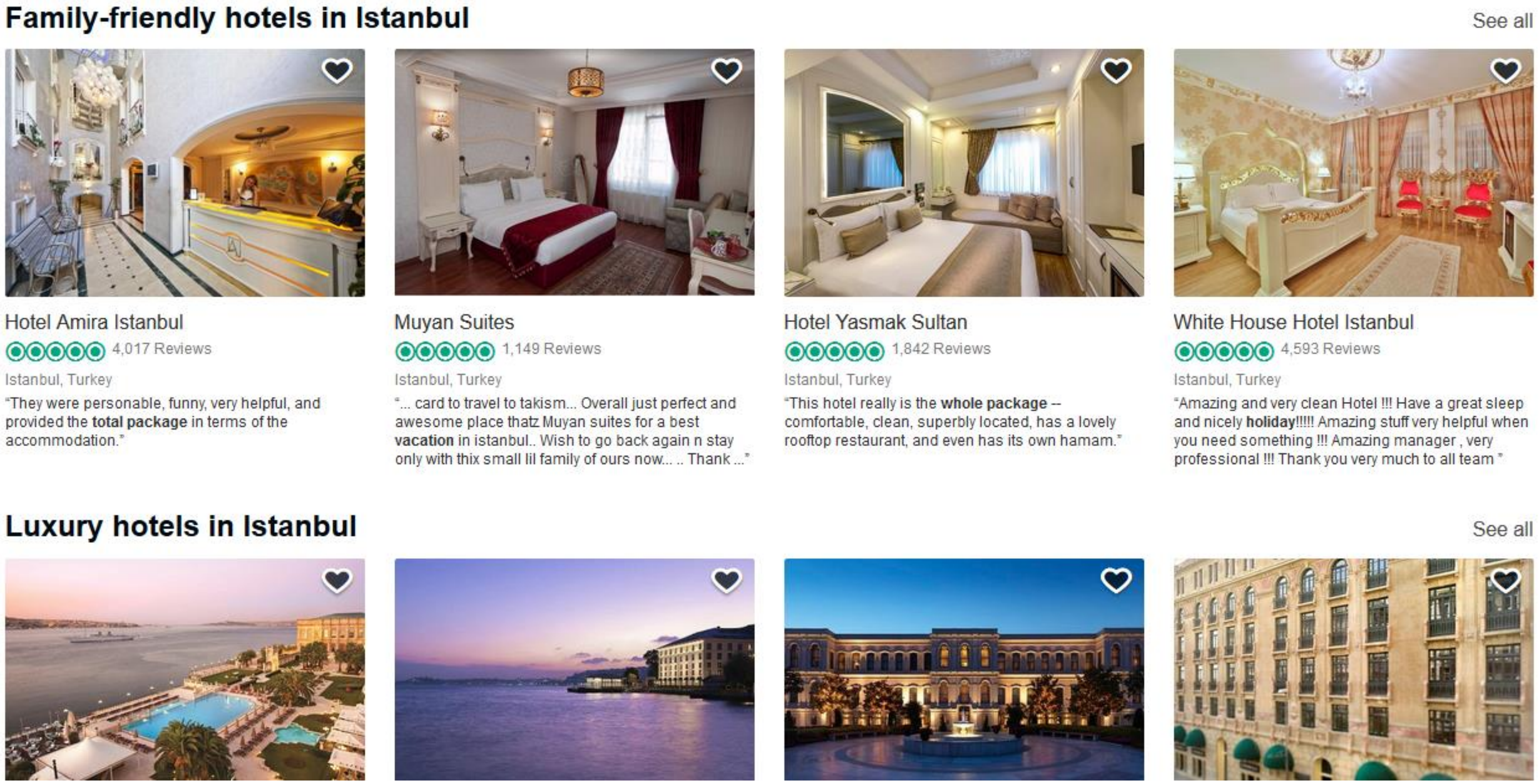
- The term Machine Learning was coined by Arthur Samuel in 1959:

“Machine Learning algorithms enable the computers to learn from data, and even improve themselves, without being explicitly programmed”.

- In 1997, Tom Mitchell gave a mathematical and relational definition that:

“A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance (as measured by P) on T improves with experience E”.

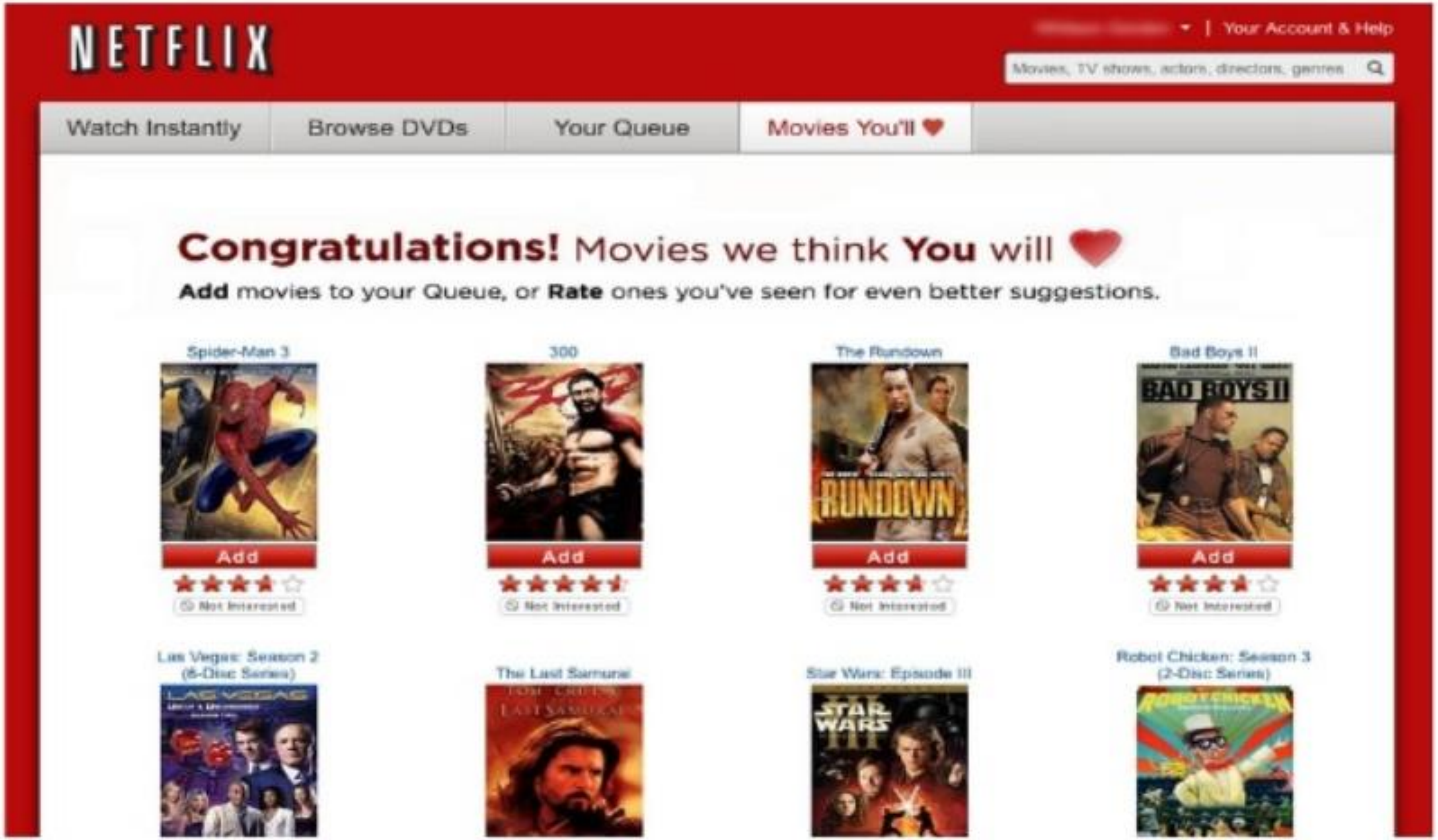
ML Examples



Example 1:

- Suppose you decide to check out trip offers for a vacation
- You browse through the travel agency website and search for a hotel
- When you look at a specific hotel, just below the hotel description there is a section titled "You might also like these hotels".
- This is a common use case of Machine Learning called "Recommendation Engine"
  - Many data points were used to train a model in order to predict what will be the best hotels to show you under that section, based on a lot of information they already know about you

Example: Netflix



ML Examples

Example 2:

- Program to predict traffic patterns at a busy intersection (task T)

- Run it through a machine learning algorithm with data (task T) about past traffic patterns (experience E) and, if it has successfully “learned”, it will then do better in predicting future traffic patterns (performance measure P).

ML Examples

Example 3: Learn to detect SPAM

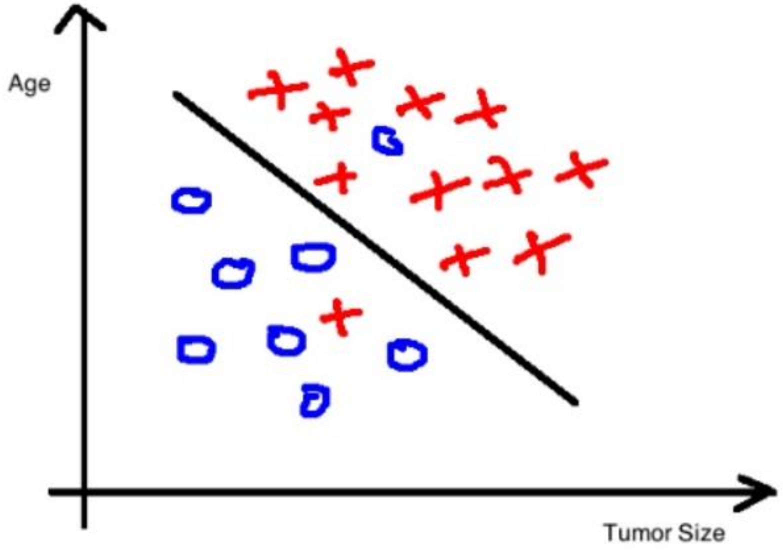
- T: Distinguish between SPAM and Non-SPAM
- P: % of emails correctly classified
- E: Labeled emails from your friend Abdullah

Examples of machine learning problems

- Medical diagnoses: ML is trained to recognize cancerous tissues
  - “Is this cancer?”
- Graph Processing:
  - “Which of these people are good friends with each other?”
- Recommender Systems:
  - “Will person X likes movie Y?”
- Financial industry and trading —fraud investigations and loan sanction
- Speech Recognition:
  - “Is this his/her voice?” (voice searches, voice dialing, call routing, and appliance control)

Such problems are excellent targets for Machine Learning, and in fact machine learning has been applied to such problems with great success.

What is a Model



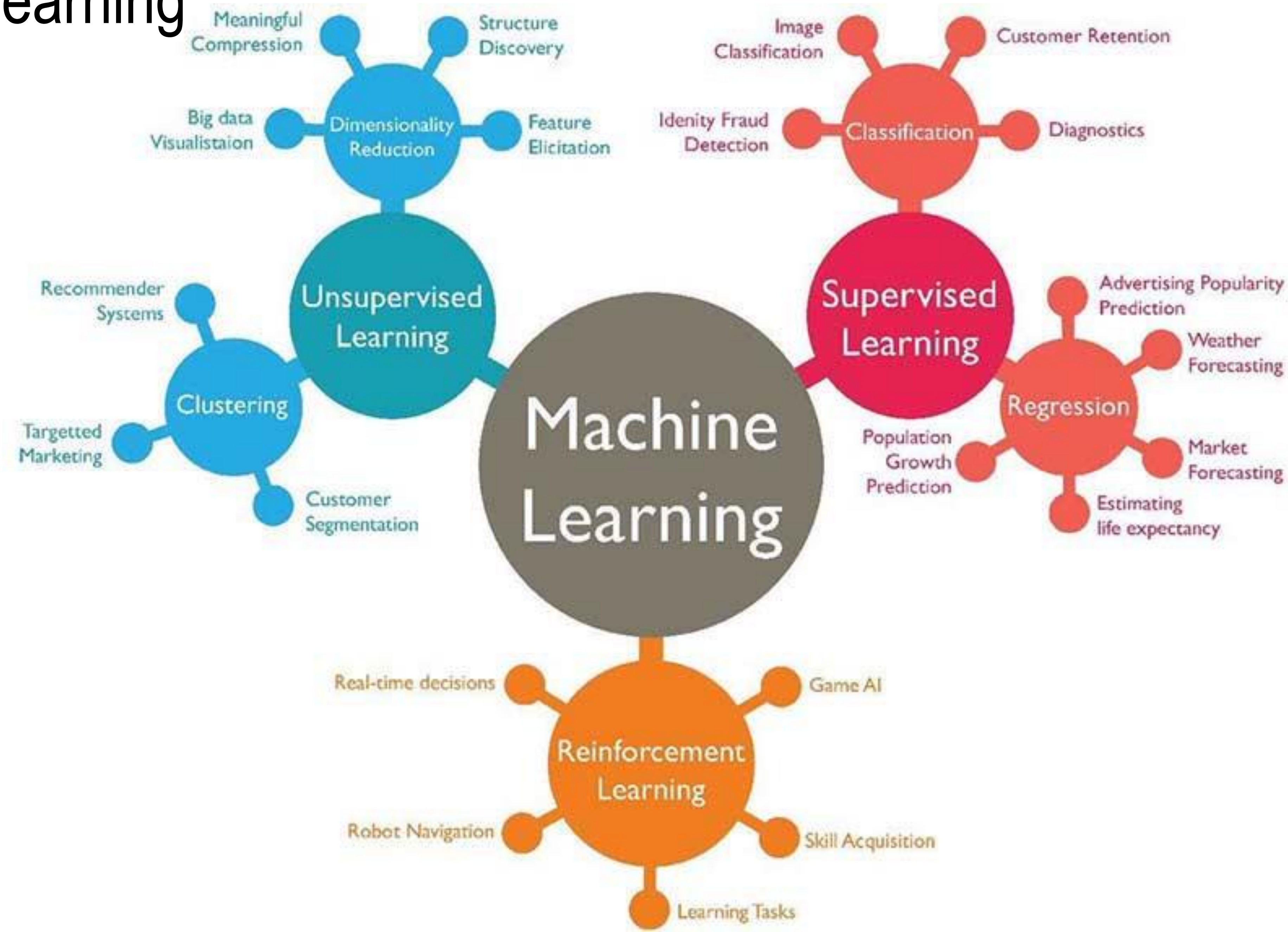
- A **model** is a mathematical formula with a number of parameters that need to be learned from the data
  - Fitting a model to the data is a process known as model training

- Example: Consider a one feature/variable linear regression, where the goal is to fit a line (described by the equation y = ax + b) to a set of distributed data points.

- Once the model training is completed we get a model equation y = 2x + 5.
  - Then for a set of inputs [1, 0, 7, 2, …] we would get a set of outputs [7, 5, 19, 9, …].

Types of Machine Learning

Machine learning can be classified into 3 types of algorithms.



Supervised Learning

- Supervised learning is a learning model built to make prediction, given an unforeseen input instance.
- A supervised learning algorithm takes a known set of input dataset and its known responses to the data (output) to learn the model.
- A learning algorithm then trains a model to generate a prediction for the response to new data or the test dataset.

- Supervised learning uses classification algorithms and regression techniques to develop predictive models.
- The algorithms include linear regression, logistic regression, neural networks, decision tree, Support Vector Machine (SVM), random forest, naive Bayes, and k-nearest neighbor.

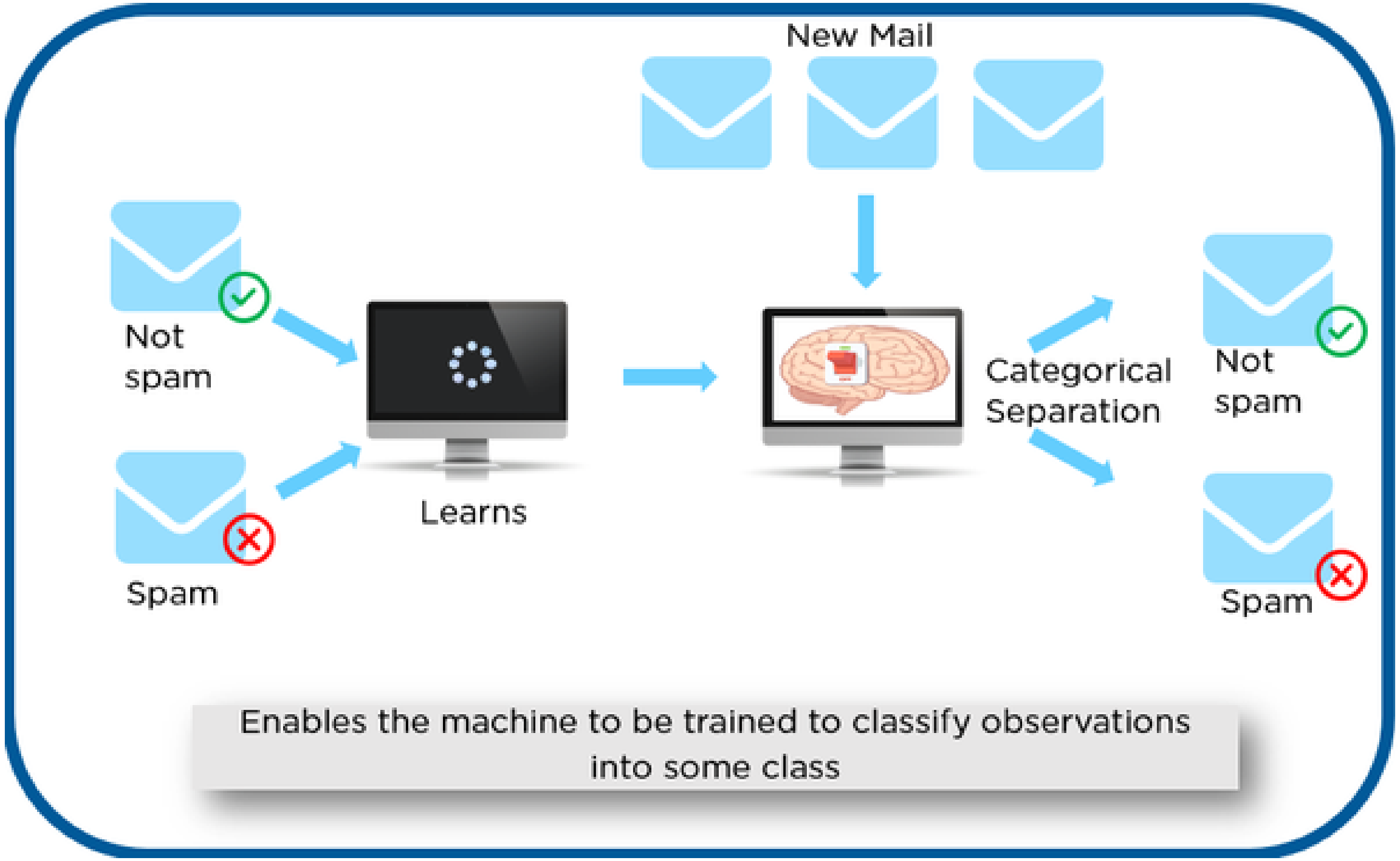
Supervised Learning

For example:

Spam filtering where large number of email messages are labelled as either:

- spam
- non-spam

- New email message will then be classified as spam or non-spam



Supervised Learning: learn from examples

Target function

Feature Matrix

Patient Age	Tumor Size	Clump Thickness	…	Malignant?
55	5	3		TRUE
70	4	7		TRUE
85	4	6		FALSE
35	2	1		FALSE
…	…	…	…	FALSE
				TRUE

Labeled dataset

Cancer model
F(k1, k2, k3, k4)

Feature Vector

$f(V1, V2, V3, \ldots) = ?$

Patient age	**Tumor size**	**Clump**	**…**
72	3	3	
66	4	4	

Malignant
?
?

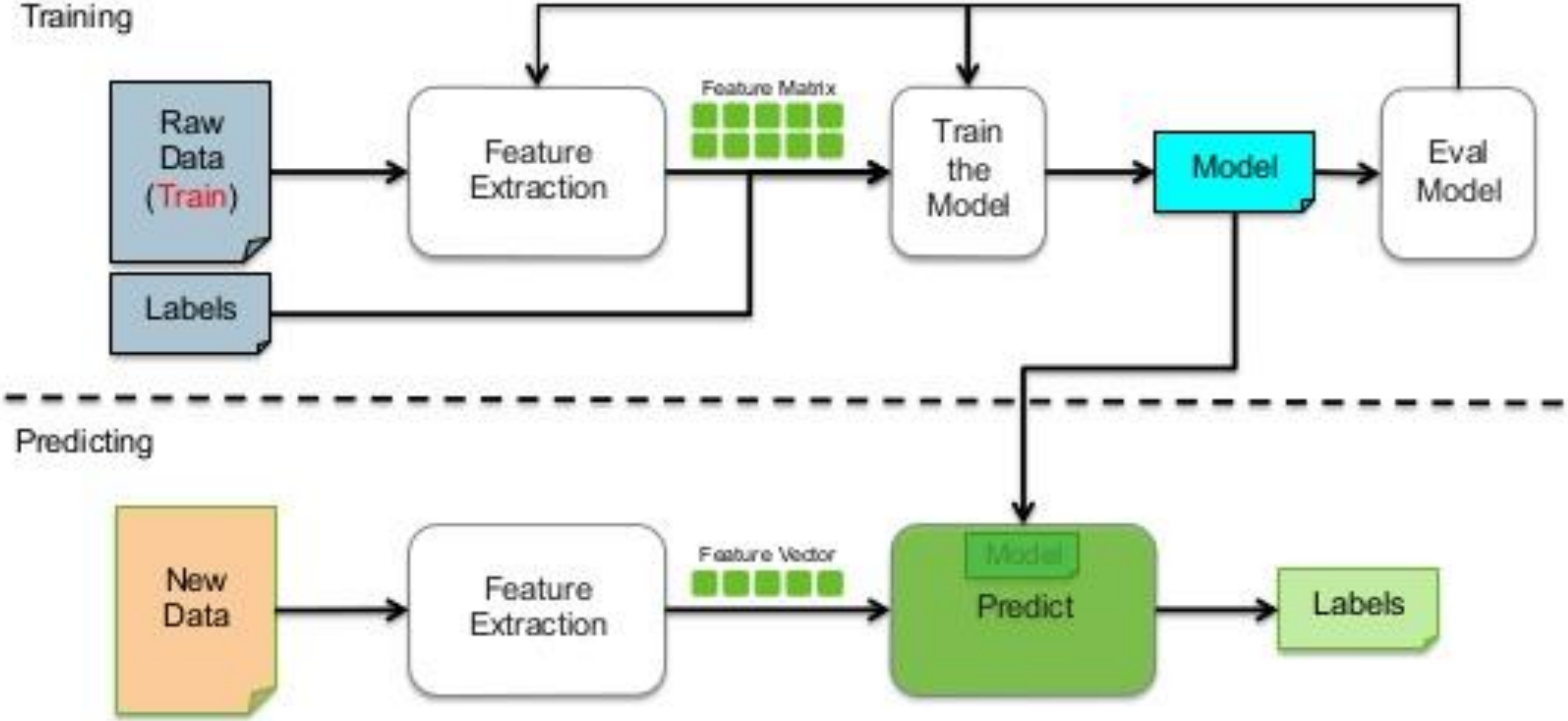
Test data

Supervised Learning

Training Images

Test Image

Supervised Learning Workflow



Classification

Classification learns from existing categorizations and then assigns unclassified items to the best category.

- Classification models classify input data into categories and predicts discrete responses
- Classification is recommended if the data can be categorized, tagged, or separated into specific groups or classes
- **Classification Examples:**
  - Bank credit scoring
  - Medical imaging
  - Speech recognition
  - To recognize letters and numbers in Handwriting
  - To check whether an email is genuine or spam
  - To detect whether a tumor is benign or cancerous
- Classification Algorithms:
  - k-nn, Decision Trees, Random Forest, SVM, Neural Network…

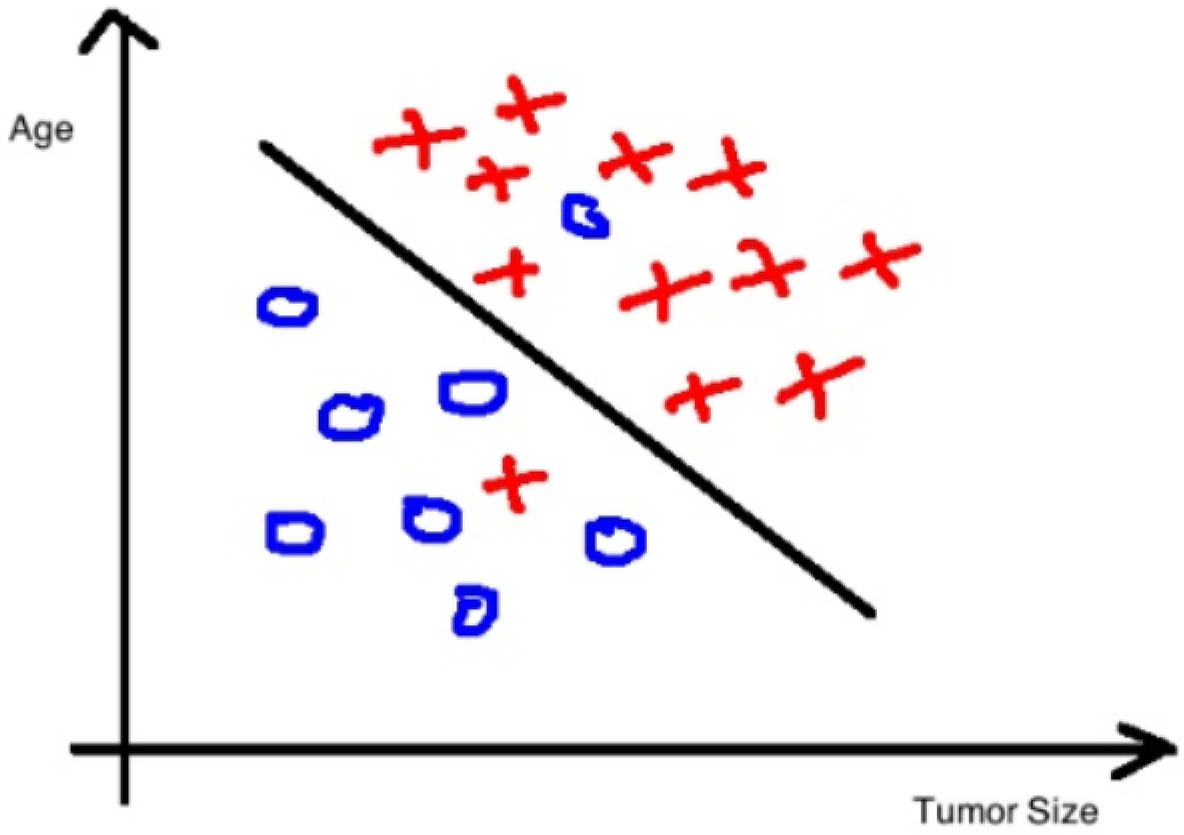
Classification Algorithms

- Classification algorithms attempt to estimate the mapping function (f) from the input variables (x) to discrete or categorical output variables (y).
  - In this case, y is a category that the mapping function predicts.

- **For example**, when provided with a dataset about houses, a classification algorithm predict whether the prices for the houses “sell more or less than the recommended retail price”

- **For example**, in a banking application, the customer who applies for a loan may be classified as a safe and risky according to his/her age and salary. The constructed model can be used to classify new data

Classification: predicting a category



Some techniques:

- Naïve Bayes
- Decision Tree
- Logistic Regression
- SGD
- Support Vector Machines
- Neural Network
- Ensembles

Basics: Regression Algorithms

Regression techniques predict continuous responses

Regression techniques predict a continuous-valued attribute associated with an object

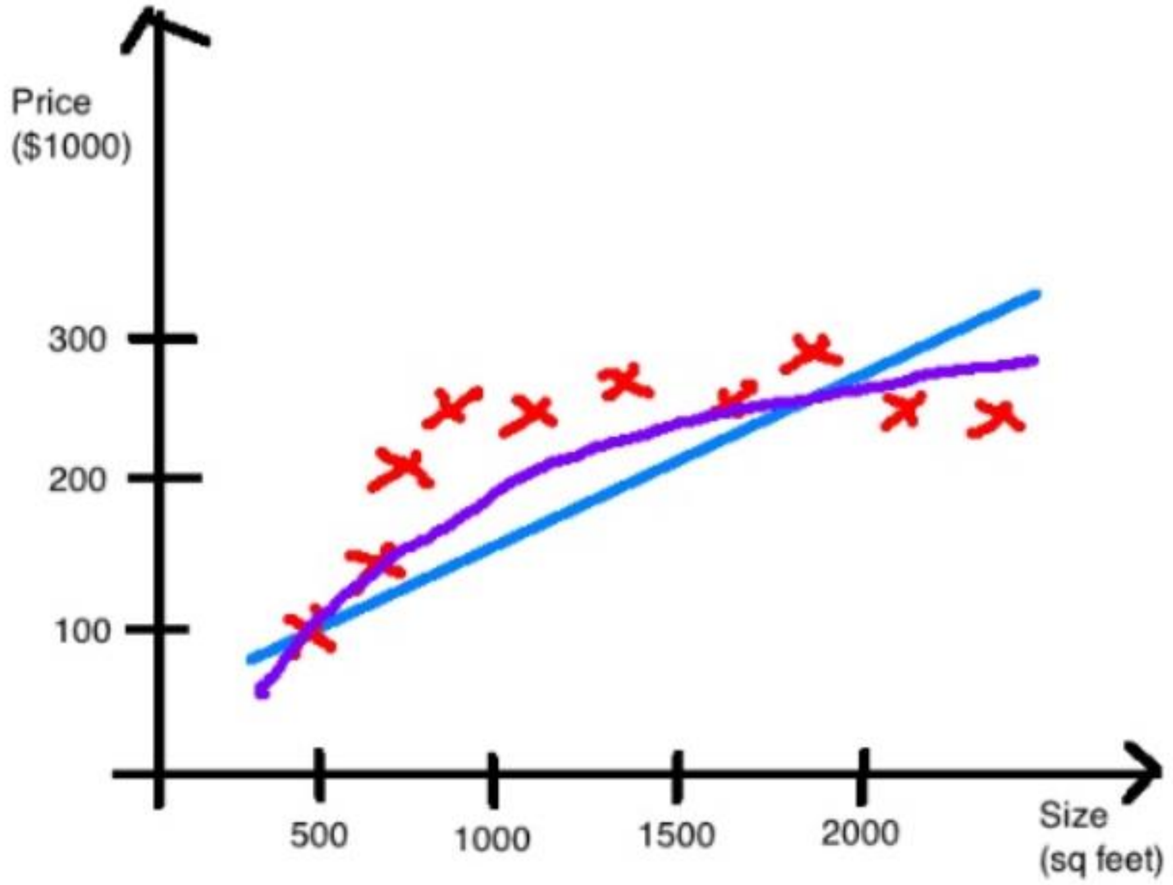
- Regression algorithms attempt to estimate the mapping function (f) from the input variables (x) to numerical or continuous output variables (y).
  - In this case, y is a real value, which can be an integer or a floating point value.
  - Therefore, regression prediction problems are usually quantities or sizes.
- For example, when provided with a dataset about houses, and you are asked to predict their prices, that is a regression task because price will be a continuous output.
- Regression algorithms include linear regression, Ensembles, Support Vector Regression (SVR), and regression trees.

Regression Examples:

- A linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data
- For example, a data is collected about how happy people are after getting so many hours of sleep.
  - In this dataset, sleep and happy people are the variables.

- Other Examples:
  - Drug Response, Stock Price, …

Regression: predict a continuous value



Some techniques:

- Linear Regression / GLM
- Decision Trees
- Support vector regression
- SGD
- Ensembles

Unsupervised Learning

In unsupervised learning the training data comprises examples of input vectors WITHOUT any corresponding target variables.

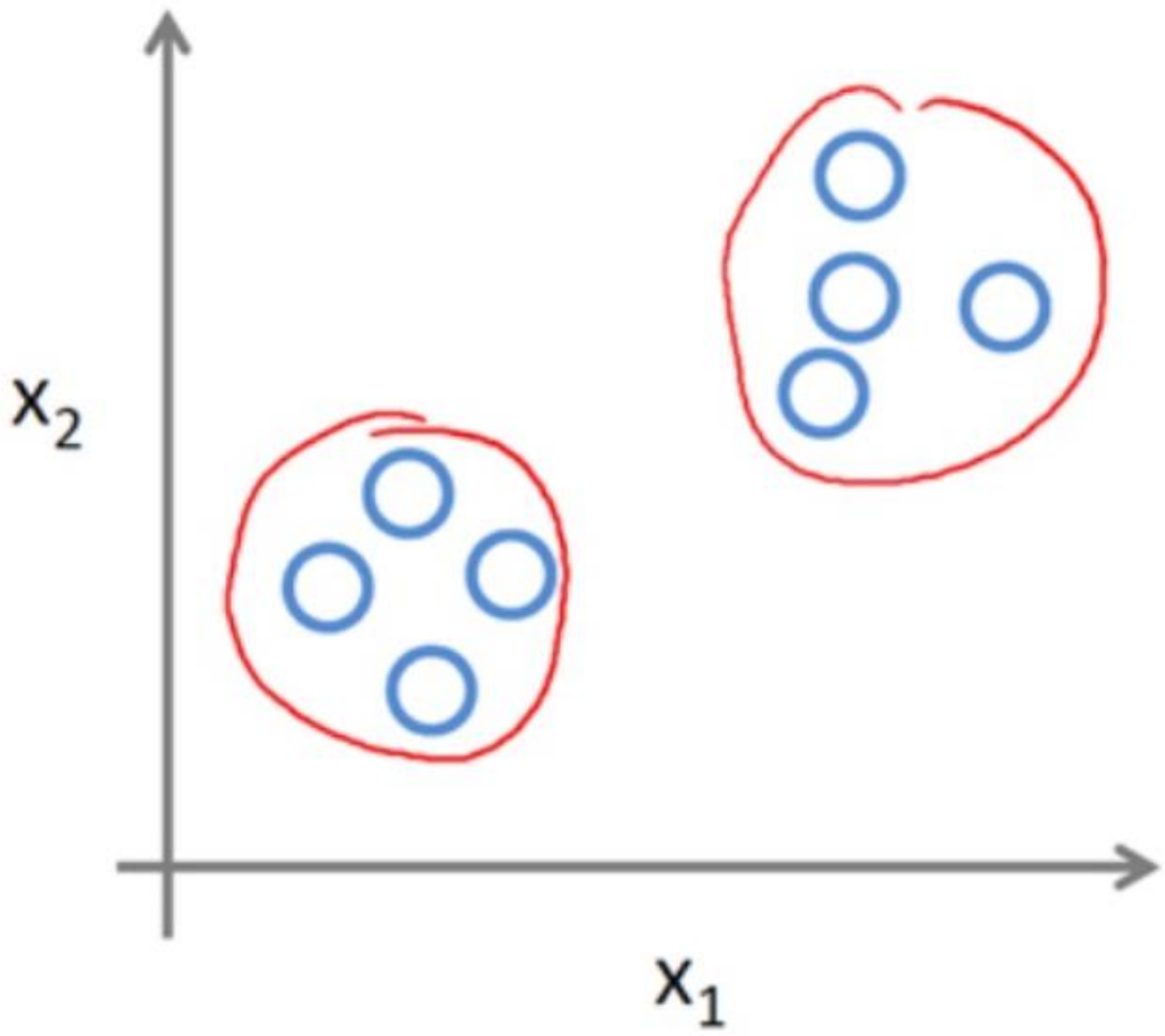
- In unsupervised learning, the algorithm learns from plain examples without any associated response, leaving to the algorithm to determine the data patterns on its own.
  - In an unsupervised learning you only have input data (X) and no corresponding output variables.
- The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

Unsupervised Learning

Examples: Speech recognition, document clustering, and image compression.

- In document clustering, the aim is to group documents into various reports of politics, entertainment, sports, culture, heritage, art, and so on.
- Fraud Detection: Identify groups of motor insurance policy holders with a high average claim cost
- Social Networks: Recognize communities within large groups of people

Clustering: detect similar instance groupings



Some techniques:

- k-means
- Spectral clustering
- DB-scan
- Hierarchical clustering

Unsupervised Learning: detect natural patterns

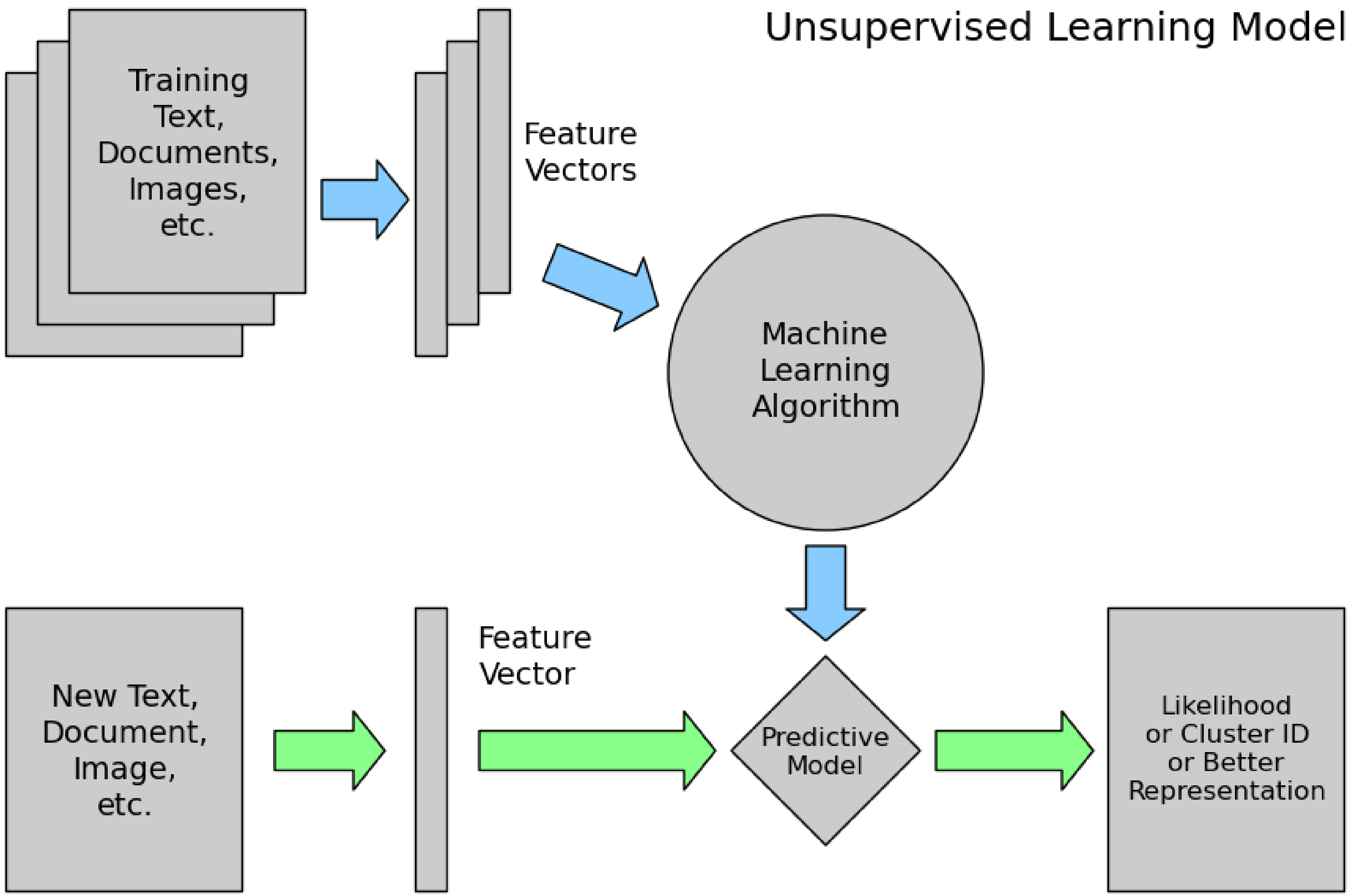
Age	State	Annual Income	Marital status
25	CA	$80,000	M
45	NY	$150,000	D
55	WA	$100,500	M
18	TX	$85,000	S
…	…	…	…

No labels

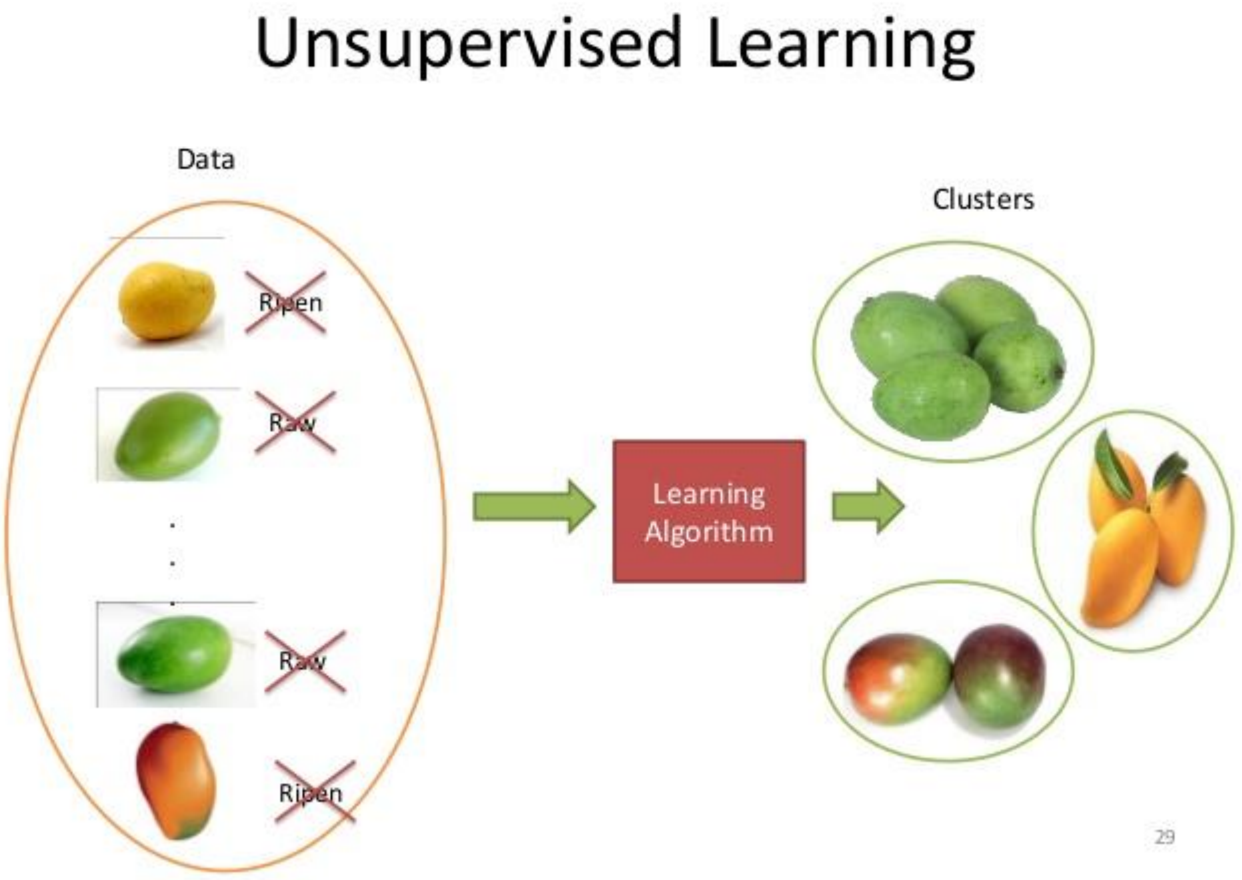
Model → Naturally occurring (hidden) structure

Hortonworks

Unsupervised Learning



Unsupervised Learning

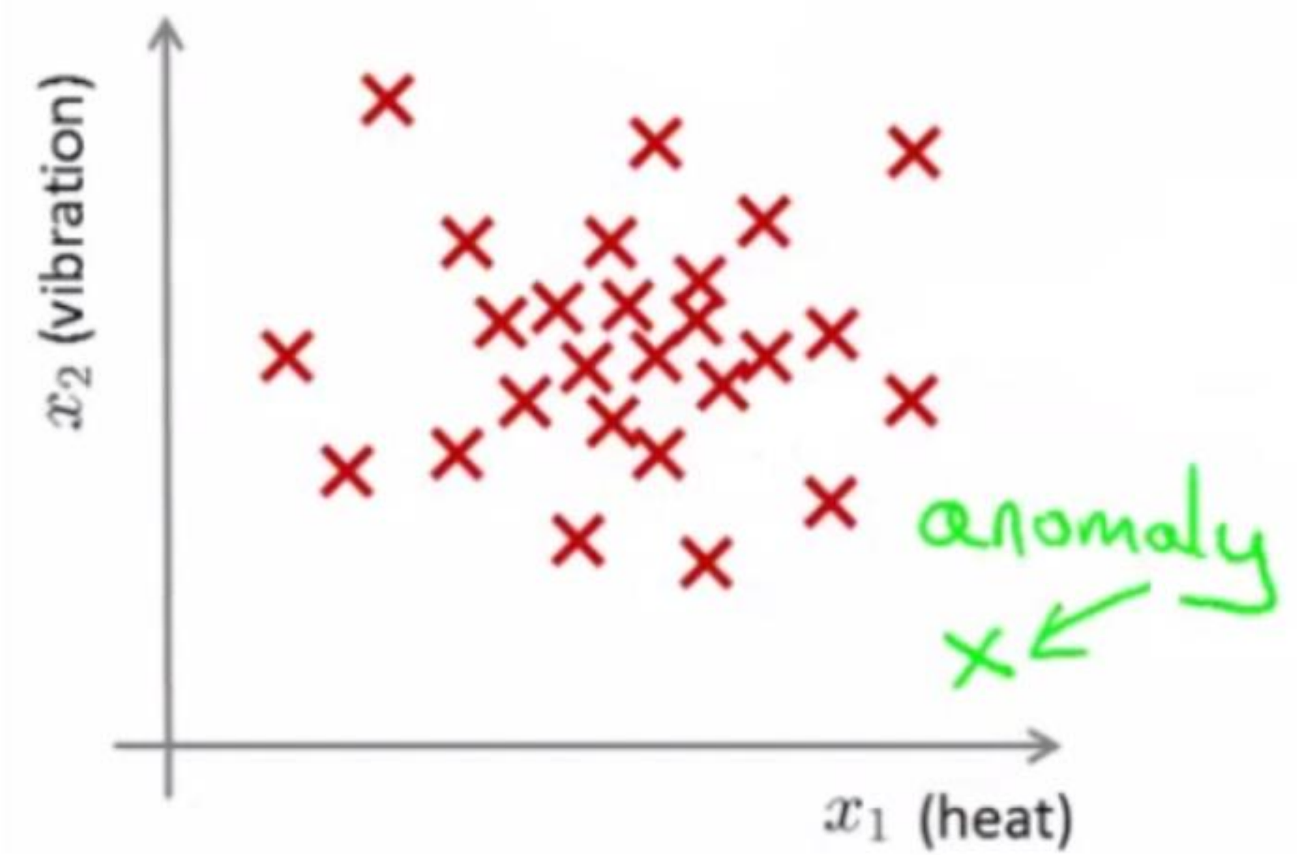


Unsupervised Learning

Objective is simply to divide above Images into N groups
Here ideally N = 2

The method can also assume other groups e.g. Group with one object or group with multiple objects

Outlier Detection: identify abnormal patterns



Example: identify engine anomalies
Features:

- Heat generated
- Vibration of engine

Hortonworks

Affinity Analysis: identifying frequent item sets

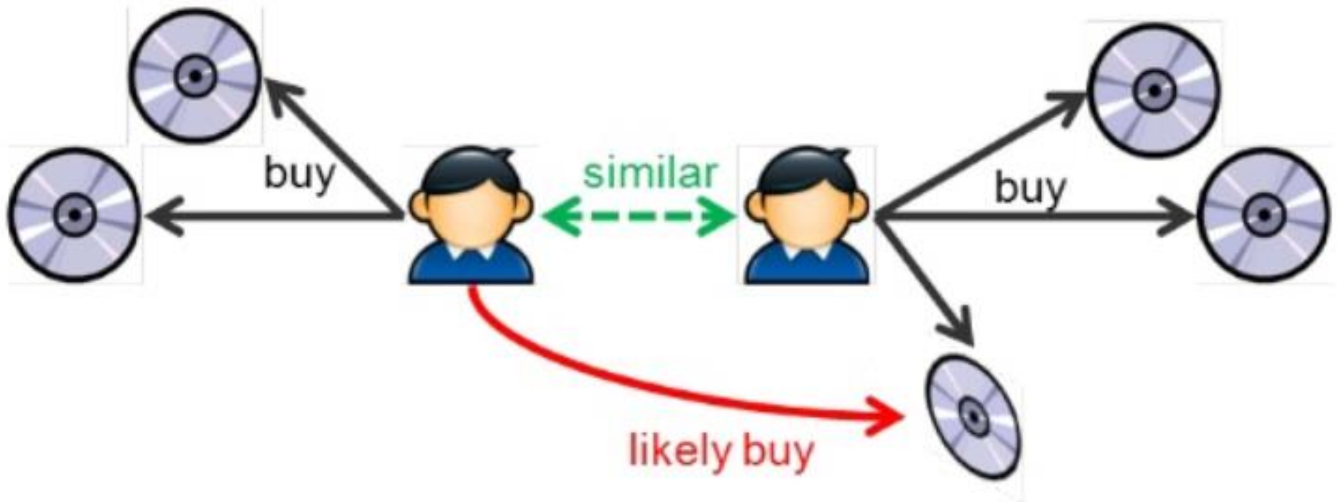
	Item 1	Item 2	Item 3	Item 4	Item 5	…
Tx 1	Y	N	N	Y	N	
Tx 2	Y	N	N	Y	N	
Tx 3	Y	Y	N	Y	N	
Tx 4	N	N	Y	Y	Y	
Tx 5						
…						

	Item 1	Item 2	Item 3	Item 4	Item 5	…
Tx 1	Y	N	N	Y	N	
Tx 2	Y	N	N	Y	N	
Tx 3	Y	Y	N	Y	N	
Tx 4	N	N	Y	Y	Y	
Tx 5						
…						

Goal: identify frequent item set
Techniques: FP Growth, a priori

Product recommendation: predicting “preference”

Clip slide



Collaborative Filtering
Identify users with similar “taste”

Collaborative filtering -> matrix completion

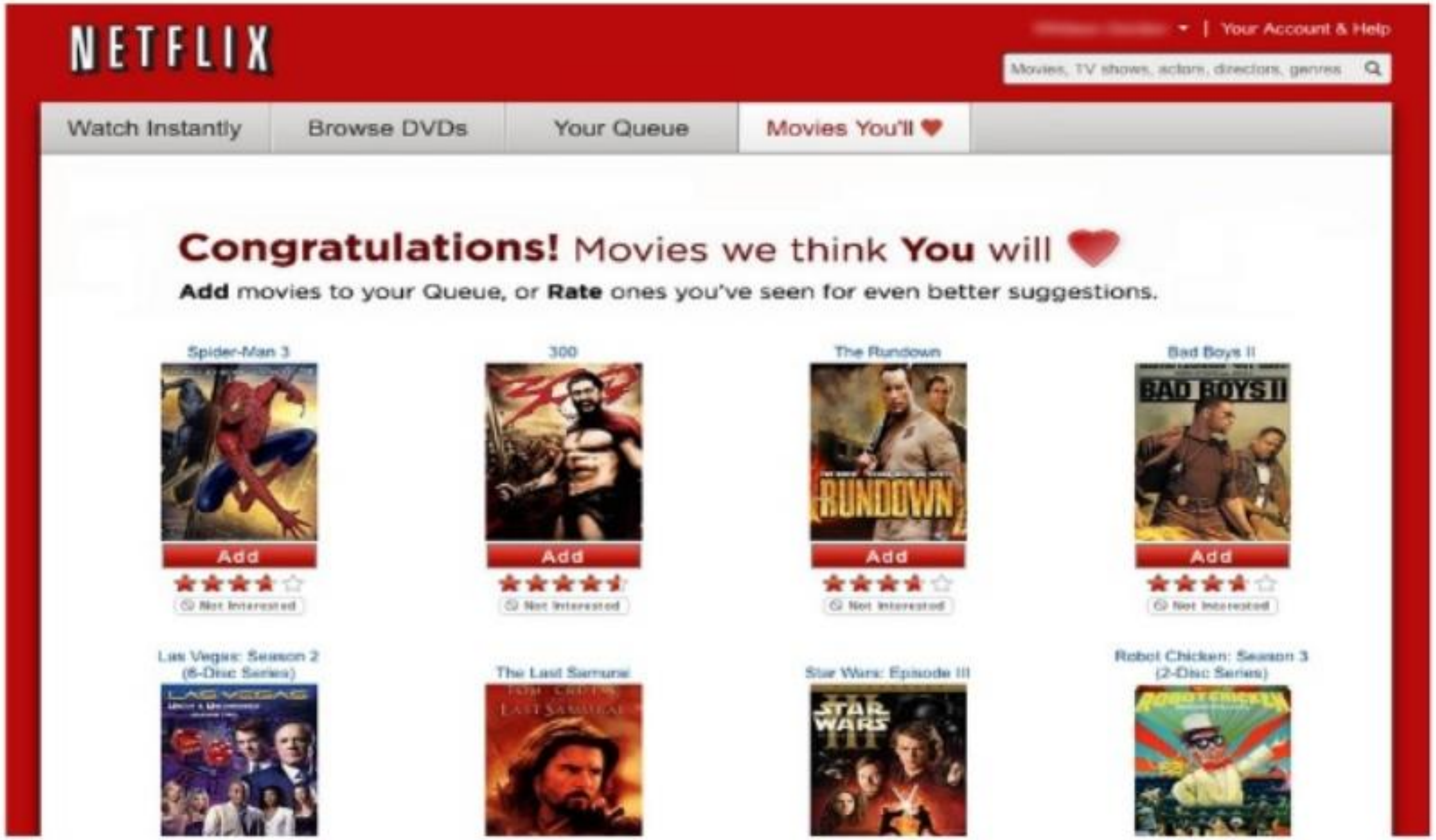
	Harry potter	X-Men	Hobbit	Argo	Pirates
101	5	2	4	?	?
102	?	?	5	2	?
103	1	2	?	?	3
104					
105					
...					

→

	Harry potter	X-Men	Hobbit	Argo	Pirates
101	5	2	4	1	3
102	4	1	5	2	3
103	1	2	4	1	3
104					
105					
...					

Hortonworks

Example: Netflix



Hortonworks

Example: market segmentation

Types of Unsupervised Learning

- In Clustering similar instances are grouped, based on their features or properties.

- **Association**: Association rules find associations amongst items within large commercial databases (e.g. Collaborative filtering)
  - Discover rules that describe large portions of data, such as people that buy X also tend to buy Y

- The similarity between two objects is measured by the **similarity function**
  - The distance between those two object is measured.
  - Shorter the distance higher the similarity, conversely longer the distance higher the dissimilarity.

Reinforcement learning

- Reinforcement learning is a type of dynamic programming that trains algorithms using a system of reward and punishment.

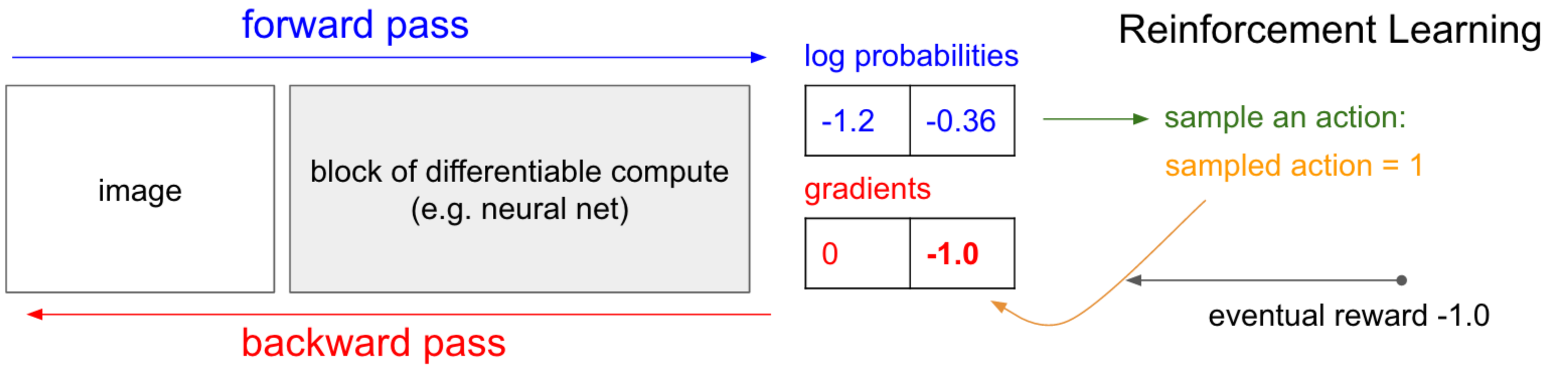
- A reinforcement learning algorithm, or agent, learns by interacting with its environment without intervention from a human by maximizing its reward and minimizing its penalty.

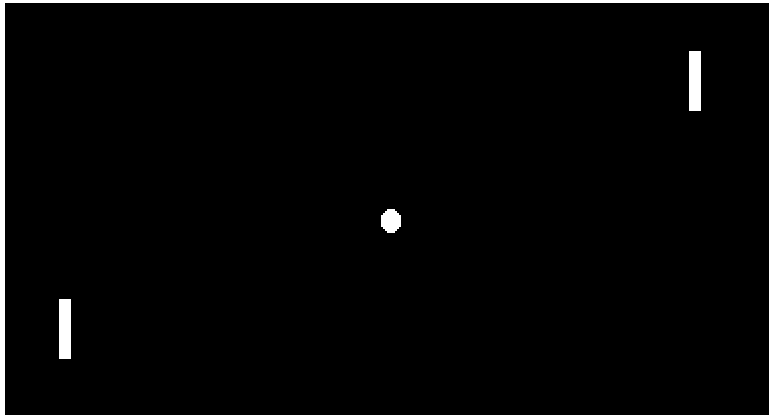
    - The agent receives rewards by performing correctly and penalties for performing incorrectly.

- Reinforcement learning contrasts with other machine learning approaches in that the algorithm is not explicitly told how to perform a task, but works through the problem on its own.

Reinforcement learning

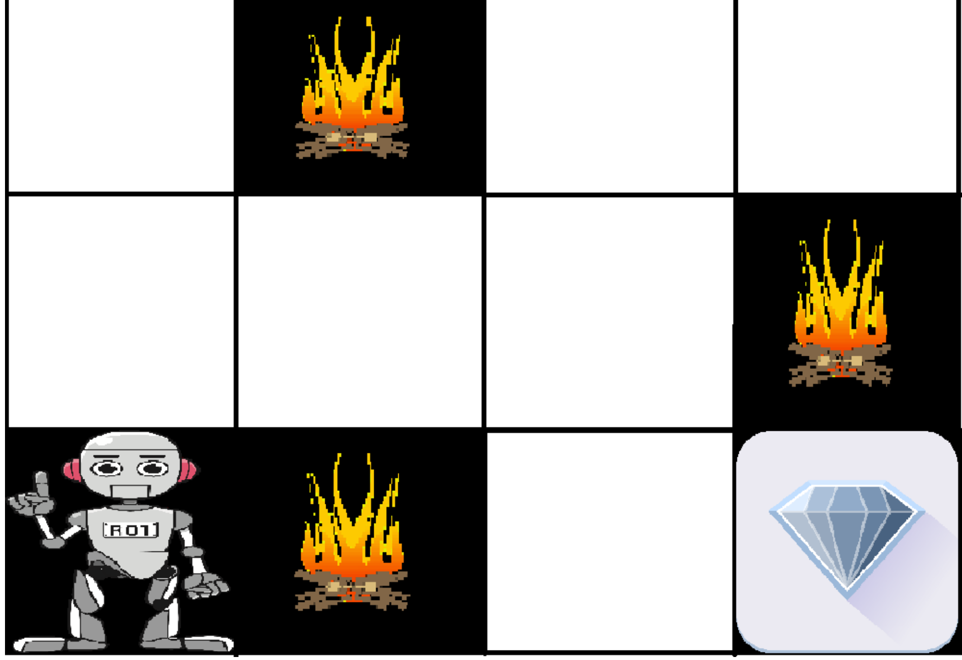
- As an agent, which could be a self-driving car or a program playing chess, interacts with its environment, receives a reward state depending on how it performs, such as driving to destination safely or winning a game.
- Conversely, the agent receives a penalty for performing incorrectly, such as going off the road or being checkmated.
    - The agent over time makes decisions to maximize its reward and minimize its penalty using dynamic programming.
- The advantage of this approach to artificial intelligence is that it allows an AI program to learn without a programmer spelling out how an agent should perform the task.

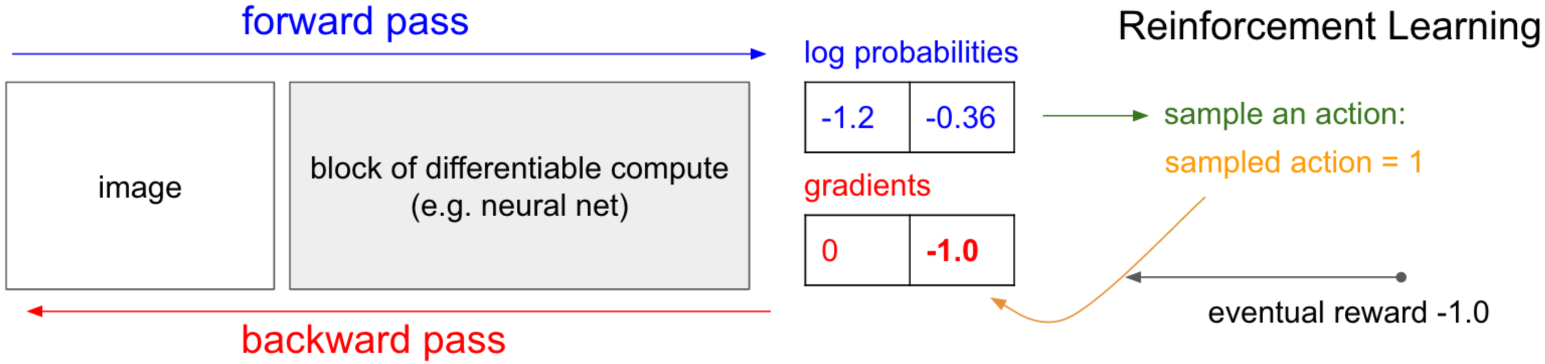




Reinforcement learning

- The agent is supposed to find the best possible path to reach the reward.
- The goal of the robot is to get the reward that is the diamond and avoid the hurdles that is 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 diamond.





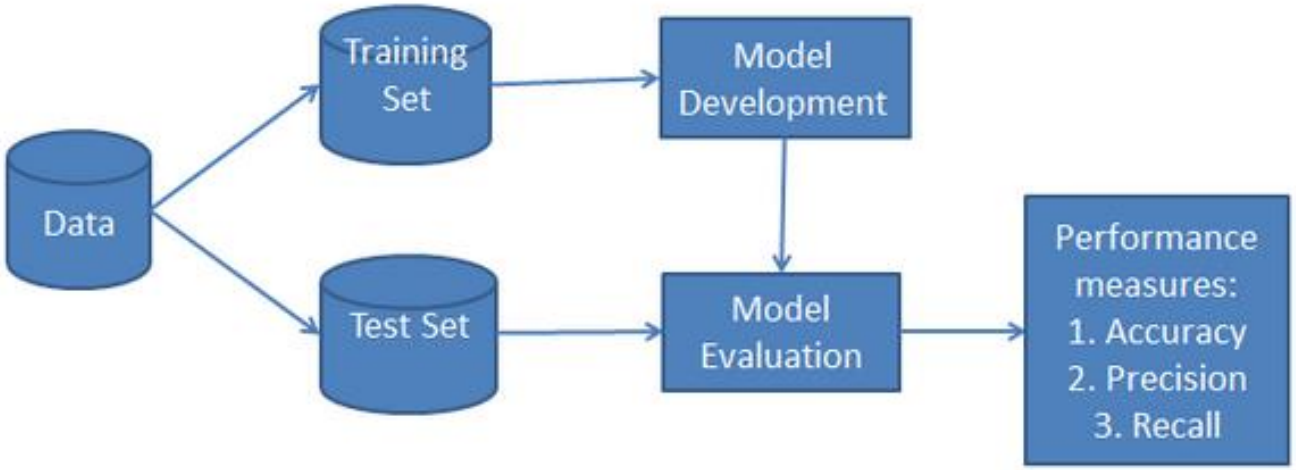
Classification Workflow

What is a classifier: A classifier is a machine learning model that is used to discriminate different objects based on certain features.

- Whenever you perform classification, the first step is to understand the problem and identify potential features and label.
- Features are those characteristics or attributes which affect the results of the label.
- For example, in the case of a loan distribution, bank manager's identify customer’s occupation, income, age, location, previous loan history, transaction history, and credit score.
- These characteristics are known as features which help the model classify customers.

Classification Workflow

- The classification has two phases, a learning phase, and the evaluation phase
  - In the learning phase, classifier trains its model on a given dataset and
  - In the evaluation phase, it tests the classifier performance
- Performance is evaluated on the basis of various parameters such as accuracy, error, precision, and recall



K Nearest Neighbors - Classification

- K nearest neighbors algorithm stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions).
- KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique.
- A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its k nearest neighbors measured by a distance function.
- If k = 1, then the case is simply assigned to the class of its nearest neighbor.

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

All three distance measures are only valid for continuous variables

K Nearest Neighbors - Classification

- In the instance of categorical variables the Hamming distance must be used.
- It also brings up the issue of standardization of the numerical variables between 0 and 1 when there is a mixture of numerical and categorical variables in the dataset.

Hamming Distance

$$D_H = \sum_{i=1}^{k} |x_i - y_i|$$

$$x = y \Rightarrow D = 0$$
$$x \neq y \Rightarrow D = 1$$

X	Y	Distance
Male	Male	0
Male	Female	1

- Choosing the optimal value for K is best done by first inspecting the data.
- In general, a large K value is more precise as it reduces the overall noise but there is no guarantee.
- Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value.
- Historically, the optimal K for most datasets has been between 3-10. That produces much better results than 1NN.

k Nearest Neighbors – Classification

- k-NN is a non-parametric method used for classification
- Prediction for the test data is done on the basis of its neighbor
- k is an integer(small), if k=1, k is assigned to the class of single nearest neighbor

Example:

Name	Acid Durability	Strength	Class
Type-1	7	7	Bad
Type-2	7	4	Bad
Type-3	3	4	Good
Type-4	1	4	Good

Assume the Test Data is: Acid Durability=3, and Strength=7. What is the class?

k Nearest Neighbors – Classification

The similarity is calculated using distance measure like Euclidean

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2 + (q_2-p_2)^2 + \cdots + (q_n-p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

Name	Acid Durability	Strength	Class	Distance
Type-1	7	7	Bad	Sqrt((7-3)^2 + (7-7)^2)=4
Type-2	7	4	Bad	5
Type-3	3	4	Good	3
Type-4	1	4	Good	3.6

k Nearest Neighbors – Classification

Rank these Attributes

Name	Acid Durability	Strength	Class	Distance	Rank
Type-1	7	7	Bad	4	3
Type-2	7	4	Bad	5	4
Type-3	3	4	Good	3	1
Type-4	1	4	Good	3.6	2

k Nearest Neighbors - Classification

k= 1

Name	Acid Durability	Strength	Class	Distance	Rank
Type-1	7	7	Bad	4	3
Type-2	7	4	Bad	5	4
Type-3	3	4	Good	3	1
Type-4	1	4	Good	3.6	2

Acceptance level is good in the two neighbors