Participant Guide

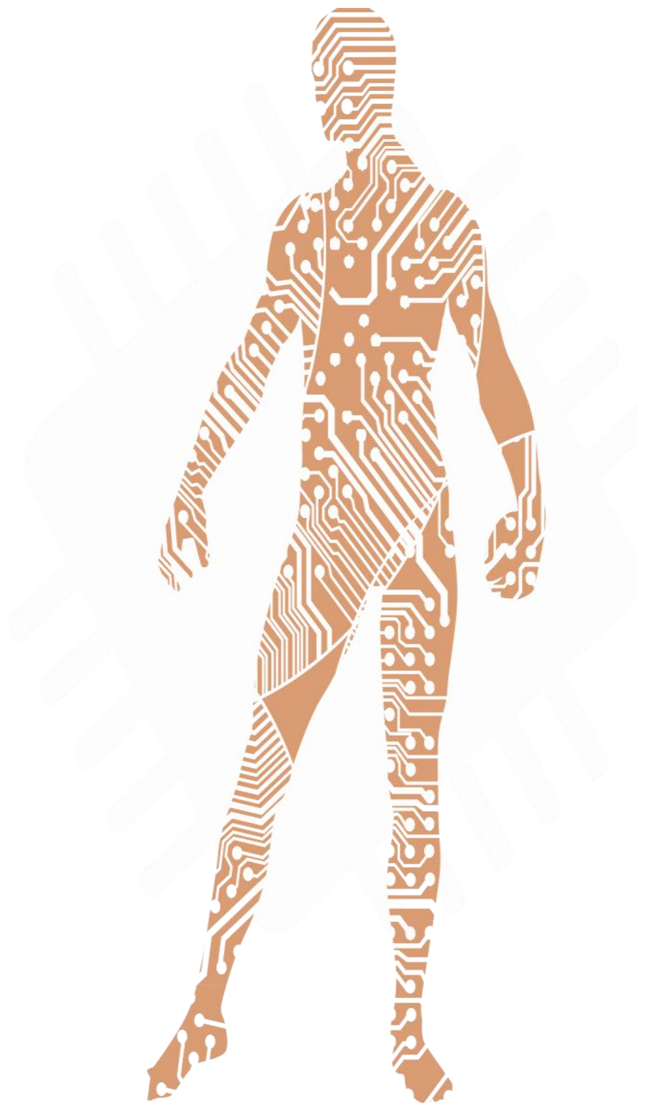
AI – ML Part II

AI and Machine Learning Basics
Part - II

We will be starting soon

Agenda

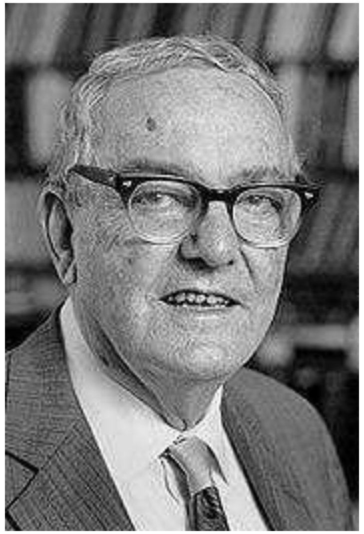
- Machine Learning - Intro
- Supervised Learning
- Classification
- Regression
- Unsupervised Learning
- Clustering
- Reinforcement Learning
- ML Use Cases
- Hands On Exercises



Machine Learning

Machine Learning

- **Herbert Alexander Simon**: “Learning is any process by which a system improves performance from experience.”
- “Machine Learning is concerned with computer programs that automatically improve their performance through experience. “



Herbert Simon
Turing Award 1975
Nobel Prize in Economics 1978

Why Machine Learning?

- Develop systems that can automatically adapt and customize themselves to individual users.
  - Personalized news or mail filter
- Discover new knowledge from large databases (***data mining***).
  - Market basket analysis (e.g. diapers and beer)
- Ability to mimic human and replace certain monotonous tasks - which require some intelligence.
  - like recognizing handwritten characters
- Develop systems that are too difficult/expensive to construct manually because they require specific detailed skills or knowledge tuned to a specific task

Why now?

- Flood of available data (especially with the advent of the Internet)
- Increasing computational power
- Growing progress in available algorithms and theory developed by researchers
- Increasing support from industries

The concept of learning in a ML system

- Learning = Improving with experience at some task
  - Improve over task *T*,
  - With respect to performance measure, *P*
  - Based on experience, *E*.

Example: Learning to Filter Spam

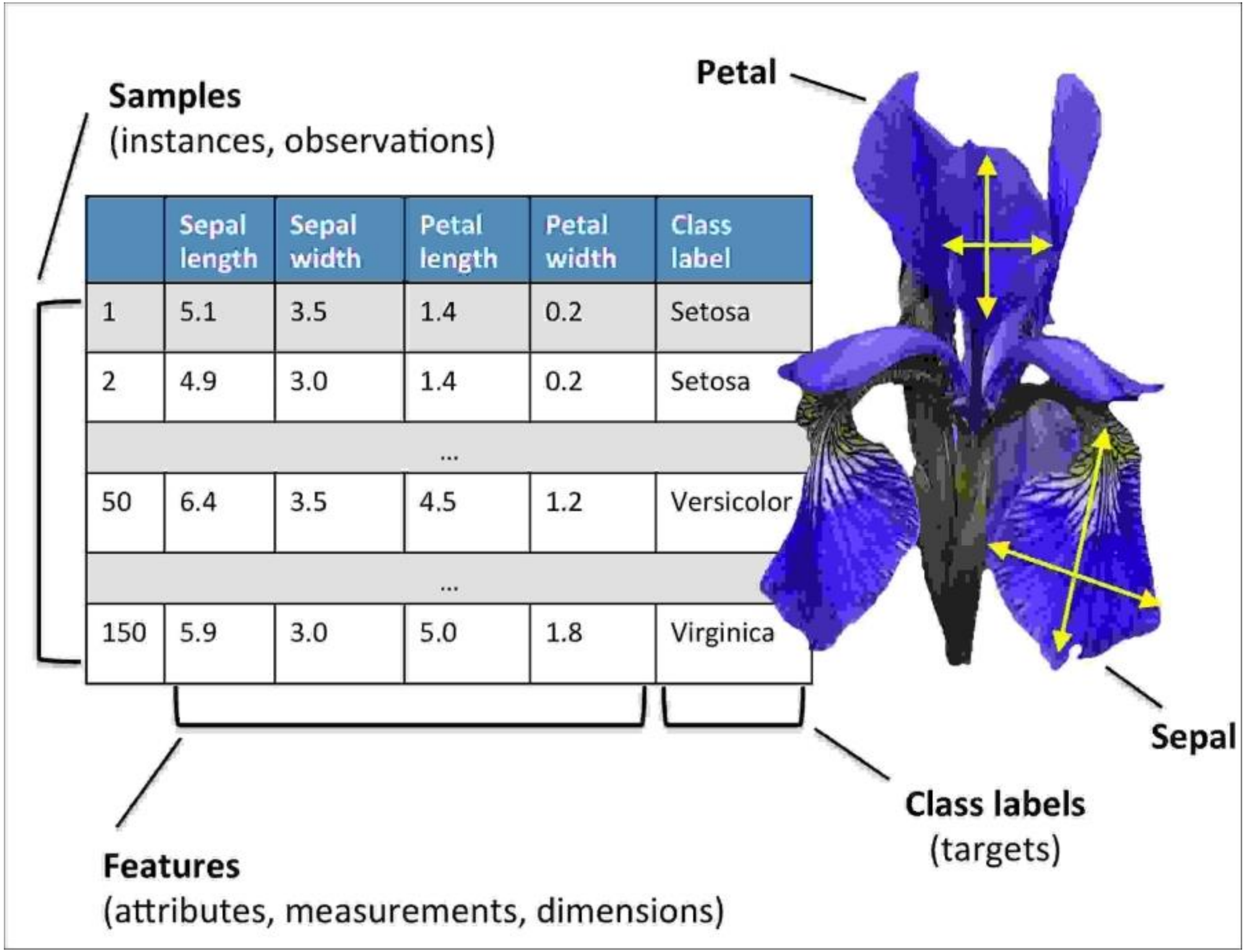
Example: Spam Filtering

Spam - is all email the user does not want to receive and has not asked to receive

T: Identify Spam Emails
P: % of spam emails that were filtered
E: a database of emails that were labelled by users

DATA

Dataset - Terminology



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
…					
50	6.4	3.5	4.5	1.2	Versicolor
…					
150	5.9	3.0	5.0	1.8	Virginica

Data Set

Input Attributes

Target Attribute

Instances

Number of new Recipients	Email Length (K)	Country (IP)	Customer Type	Email Type
0	2	Germany	Gold	Ham
1	4	Germany	Silver	Ham
5	2	Nigeria	Bronze	Spam
2	4	Russia	Bronze	Spam
3	4	Germany	Bronze	Ham
0	1	USA	Silver	Ham
4	2	USA	Silver	Spam

Numeric

Nominal

Ordinal

Big Data Trunk

Labelled Dataset – Numerical Outputs

Input Features

Output Label

Size	Color	Shape	Taste	Price
L	Red	Square	Good	$15
M	Blue	Circle	Bad	$10
S	Red	Triangle	Good	$7
M	Yellow	Square	Bad	$19

Labelled Dataset – Class Labels

BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
72	35	0	33.6	0.627	50	1
66	29	0	26.6	0.351	31	0
64	0	0	23.3	0.672	32	1
66	23	94	28.1	0.167	21	0
40	35	168	43.1	2.288	33	1
74	0	0	25.6	0.201	30	0
50	32	88	31	0.248	26	1
0	0	0	35.3	0.134	29	0
70	45	543	30.5	0.158	53	1
96	0	0	0	0.232	54	1
92	0	0	37.6	0.191	30	0
74	0	0	38	0.537	34	1
80	0	0	27.1	1.441	57	0
60	23	846	30.1	0.398	59	1
72	19	175	25.8	0.587	51	1
0	0	0	30	0.484	32	1

Unlabelled Dataset – Class Labels

Values	Attribute	Clump Thickness		Uniformity of Cell Size		Uniformity of Cell Shape		Marginal Adhesion		Single Epithelial Cell Size		Bare Nuclei		Bland Chromatin		Normal Nucleoli		Mitoses	
		M	B	M	B	M	B	M	B	M	B	M	B	M	B	M	B	M	B
1	Malignant=241, Benign = 458	2%	98%	1%	99%	0.6%	99.4	8%	92%	2%	98%	30%	70%	1%	99%	9%	91%	23%	77%
2		8%	92%	18%	82%	12%	88%	36%	64%	6%	94%	35%	65%	4%	96%	17%	83%	77%	23%
3		11%	89%	48%	52%	41%	59%	47%	53%	60%	40%	45%	55%	22%	78%	73%	27%	94%	6%
4		15%	85%	77%	23%	70%	30%	85%	15%	85%	15%	30%	70%	80%	20%	94%	6%	100%	0%
5		35%	65%	100%	0%	91%	9%	83%	17%	87%	13%	50%	50%	88%	12%	89%	11%	83%	17%
6		53%	47%	93%	7%	90%	10%	82%	18%	95%	5%	75%	25%	90%	10%	82%	18%	100%	0%
7		96%	4%	95%	5%	93%	7%	100%	0%	75%	25%	75%	25%	90.%	10%	87%	13%	89%	11%
8		91%	9%	97%	3%	96%	4%	100%	0%	90%	10%	62%	38%	100%	0%	83%	17%	88%	12%
9		100%	0%	83%	17%	100%	0%	80%	20%	100%	0%	67%	33%	100%	0%	94%	6%	50%	50%
10		100%	0%	100%	0%	100%	0%	98%	2%	97%	3%	33%	67%	100%	0%	100%	0%	100%	0%

The concept of learning in a ML system

- Learning = Improving with experience at some task
  - Improve over task *T*,
  - With respect to performance measure, *P*
  - Based on experience, *E*.

Traditional Software Development

- Convert inches to cm

- Input:
- Output:

Traditional Software Development

- Input: **inches**
- Relationship: **cm** = **inches** * 2.54
- Output: cm

Traditional Software Development

- Convert a number to its absolute value

Input:

Traditional Software Development

Convert a to its **absolute value**

Input: **number**

Output:

Traditional Software Development

Input: **number**

Rules:

if **number** >= 0: **abs. value** = **number**

Output:

Big Data Trunk

Traditional Software Development

Input: **number**

Rules:

if **number** >= 0: **abs. value** = **number**

else: **abs. value** = **number** * -1

Output:

Traditional Software Development

Input: number

Rules:

if number >= 0: abs. value = number

else: abs. value = number * -1

Output: abs. value

Machine Learning

Input	144	181	200	317	800
Output					

Machine Learning

Input	144	181	200	317	800
Output	256	219	200	?	-400

Machine Learning

Input	144	181	200	317	800
Output	256	219	200	83	-400

Output = 400 - Input

Input	144	181	200	317	800
Output	256	219	200	83	-400

Machine Learning

Input: [144, 181, 200, 800]

Machine Learning

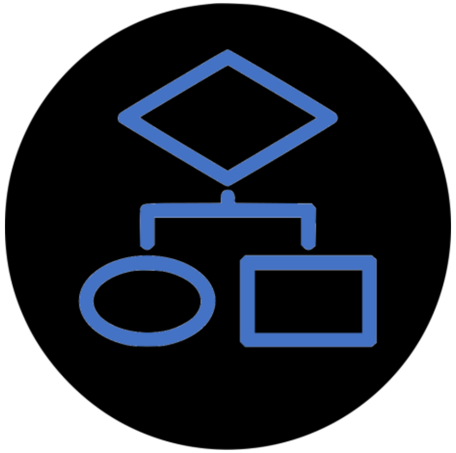
Input: [144, 181, 200, 800]

Output: [256, 219, 200, -400]

Machine Learning

- Input: [144, 181, 200, 800]
- Relationship: ?
- Output: [256, 219, 200, -400]

Common ML Algorithms



=

Linear Regression

Logistic Regression

Naïve Bayes

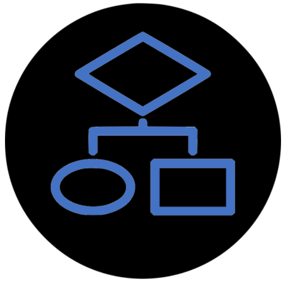
Support Vector Machine

Decision Tree

K-Nearest Neighbor

Machine Learning

- Input: [144, 181, 200 800]
- Relationship:



- Output: [256, 219, 200, -400]

Machine Learning

- Input: [144, 181, 200 800]

- Relationship: 400 - **input**

- Output: [256, 219, 200, -400]

Machine Learning

- Input: [144, 181, 200 800]

- Relationship: **400 - input** ⟵ **Model**

- Output: [256, 219, 200, -400]

Machine Learning

ML Model

400 - Input

Machine Learning

ML Model

New input: **317** → **400 - Input**

Machine Learning

ML Model

New input: **317** → **400 - Input** → output: **83**

Models the relationship between input and output

The Prediction

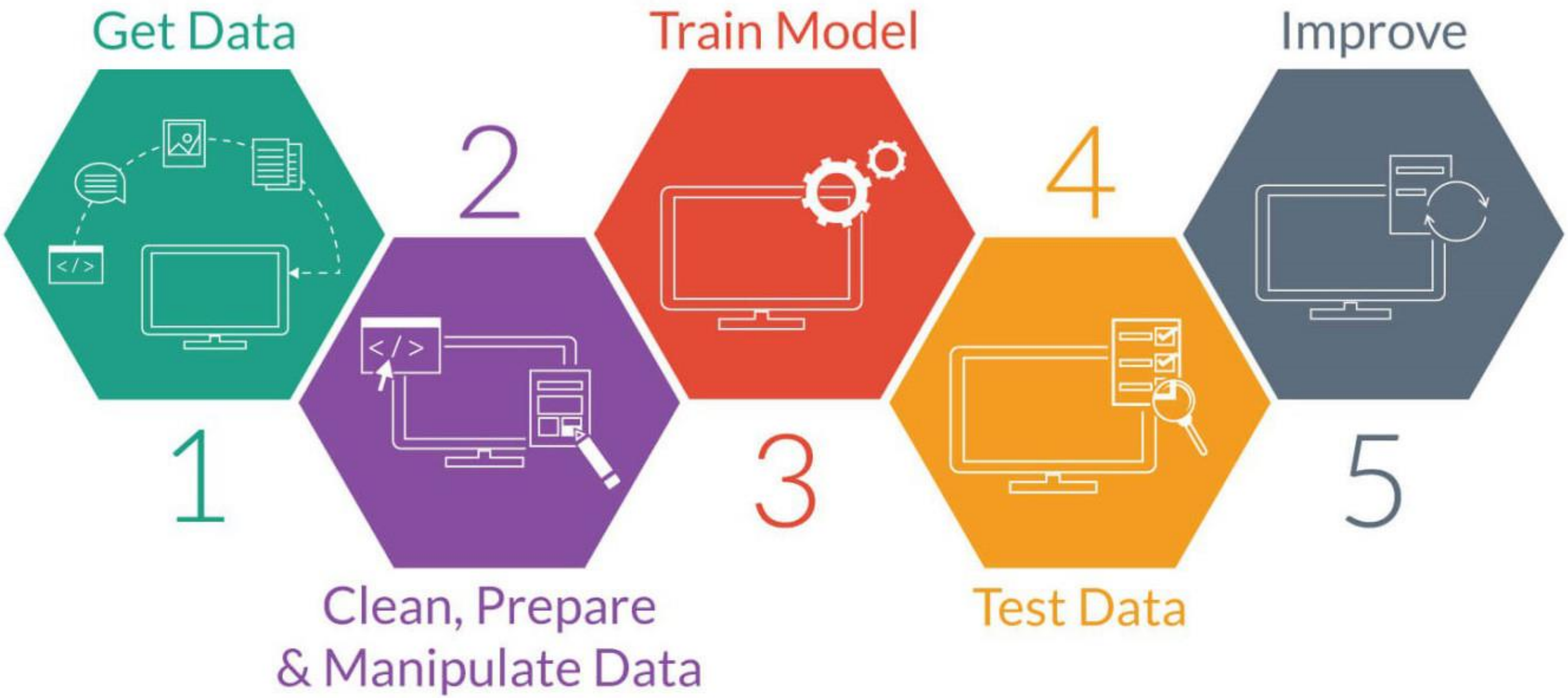
$$\underset{\text{output}}{\hat{y}} = \underset{}{\hat{f}}\underset{\text{input}}{(X)}$$

Models the relationship between X and y

The Prediction

$$\overset{\text{output}}{\hat{y}} = \hat{f}(\overset{\text{input}}{X})$$

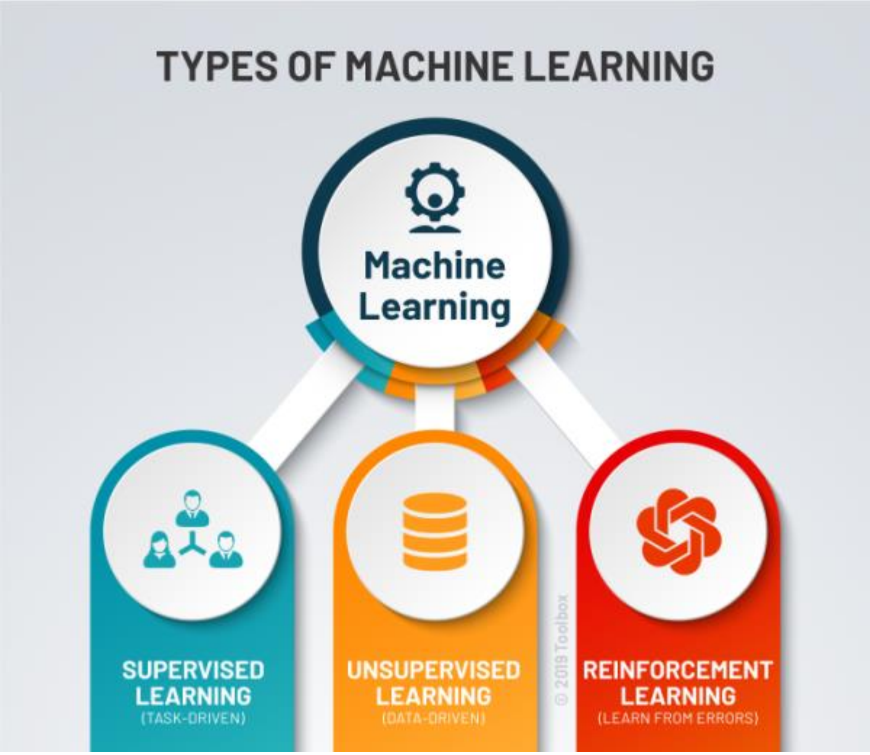
The ML Process



Machine Learning - Types

ML – Types

- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
- Reinforcement Learning

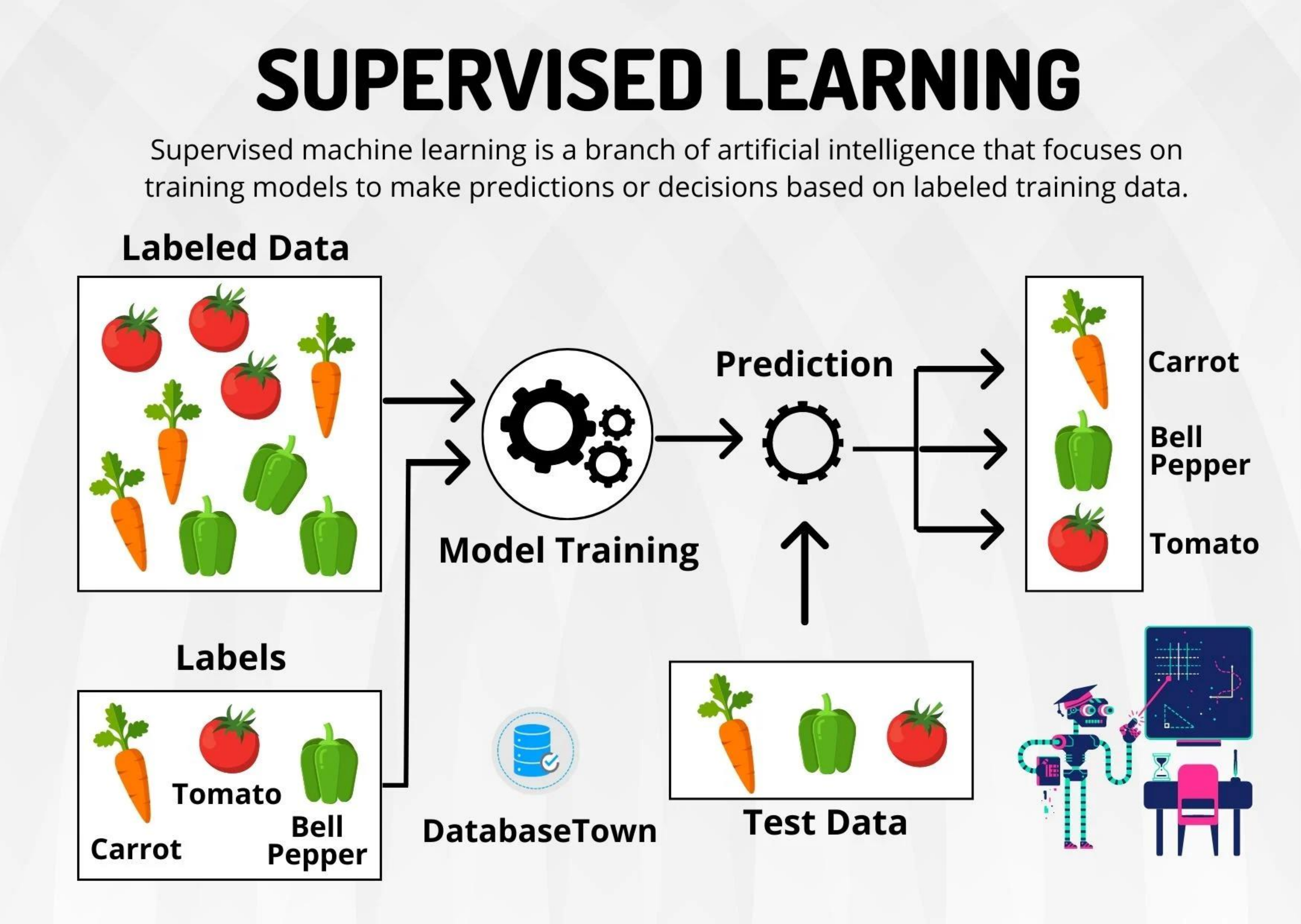


Supervised, unsupervised, and reinforcement learning.

Big Data Trunk

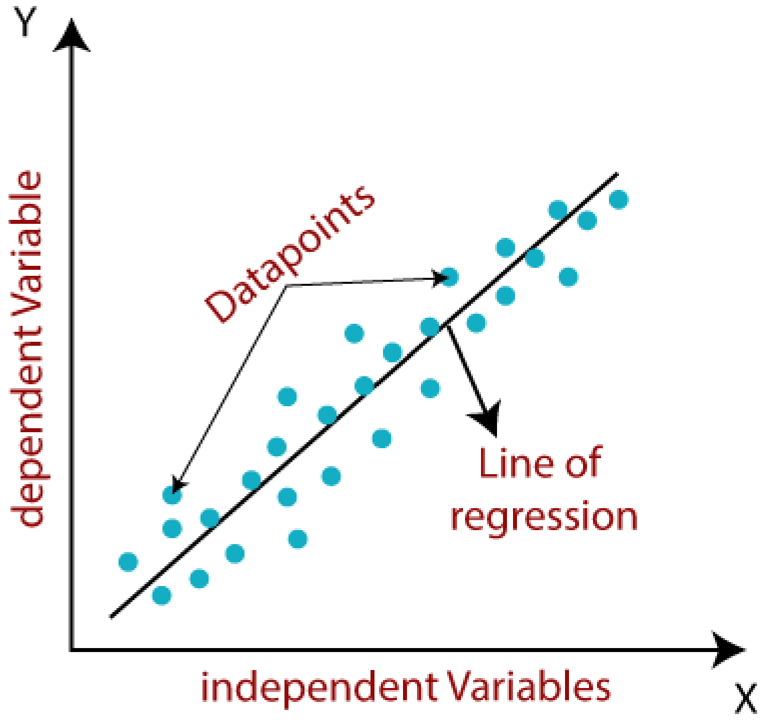
Supervised Learning

- Supervised learning is a type of machine learning where the algorithm learns from labeled data to make predictions or decisions based on the data inputs.
- Labeled data means that some input data is already tagged with the correct output or the desired outcome.
- The algorithm tries to learn the relationship between the input and output data so that it can make accurate predictions on new, unseen data.
- Supervised learning can be used for various applications such as image classification, spam filtering, fraud detection, risk assessment, etc


SUPERVISED LEARNING
Supervised machine learning is a branch of artificial intelligence that focuses on training models to make predictions or decisions based on labeled training data.
Labeled Data
Model Training
Prediction
Carrot
Bell Pepper
Tomato
Labels
Tomato
Carrot
Bell Pepper
DatabaseTown
Test Data

Types of Supervised Learning - Regression

- Two main categories: Regression and Classification.
- Regression is a type of supervised learning where the output is a continuous numerical value, such as the price of a house, the wind speed, the temperature, etc.
- Examples of regression algorithms: Linear regression, Polynomial regression etc.



Regression

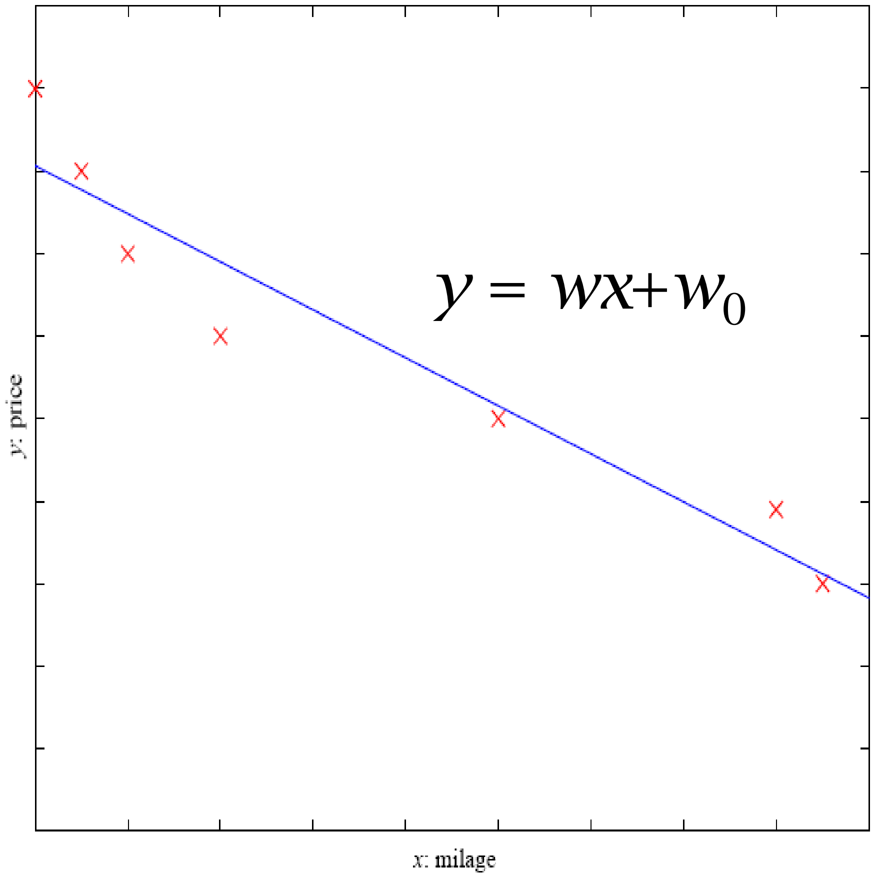
- Example: Price of a used car
- x : car attributes

  y : price

  $y = g(x \mid \theta)$

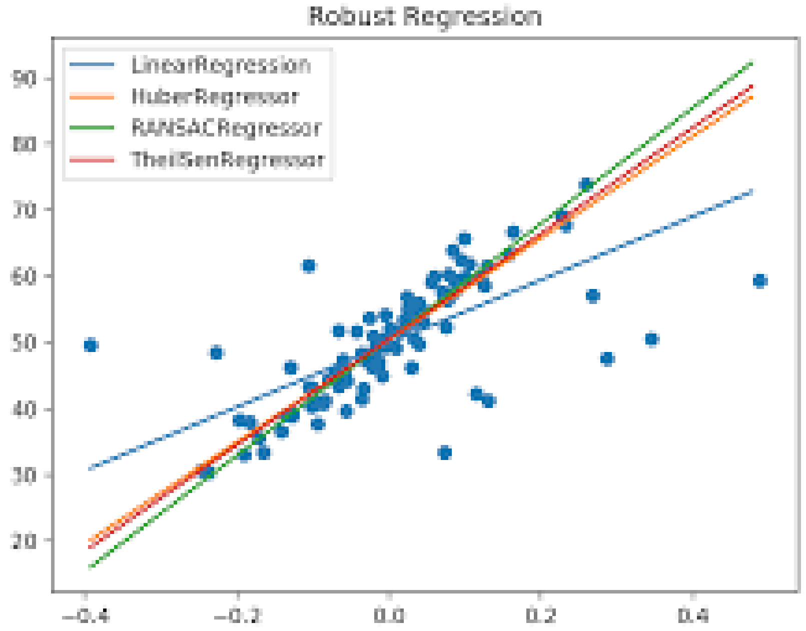
  $g(\ )$ model,

  θ parameters



Regression Applications

- Navigating a car: Angle of the steering wheel (CMU NavLab)
- Kinematics of a robot arm



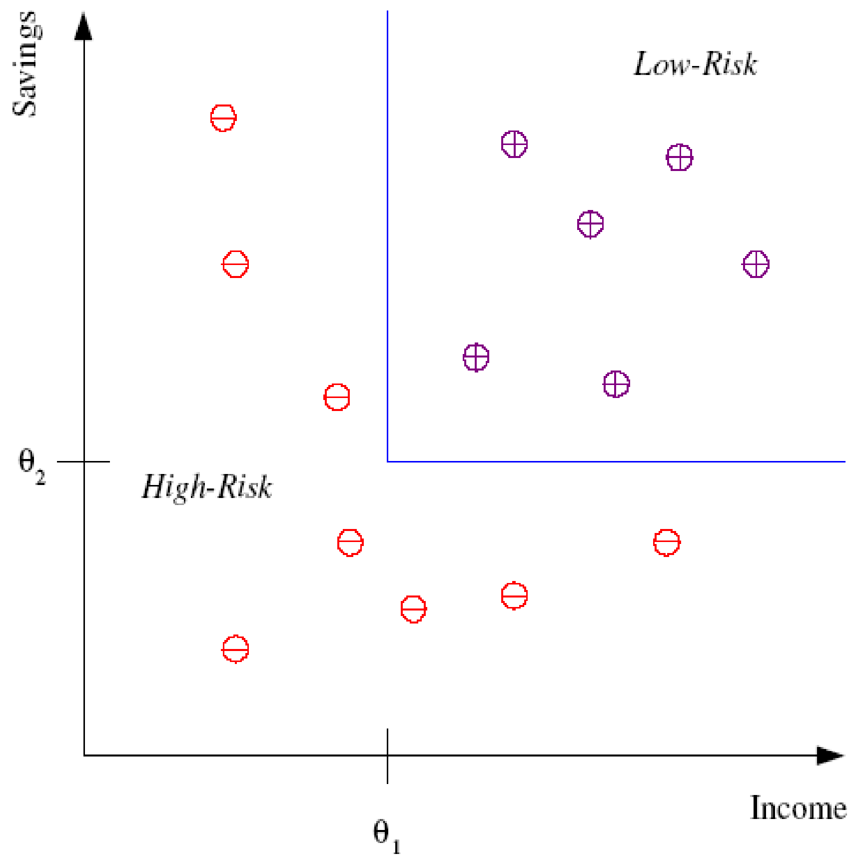
Big Data Trunk

Types of Supervised Learning - Classification

- Classification is a type of supervised learning where the output is a discrete categorical value, such as the shape of an object, the sentiment of a text, the type of a flower, etc.
- Examples of classification algorithms: Logistic regression, K-nearest neighbors, Support Vector Machine, Decision Trees etc.

Classification

- Example: Credit scoring
- Differentiating between low-risk and high-risk customers from their *income* and *savings*

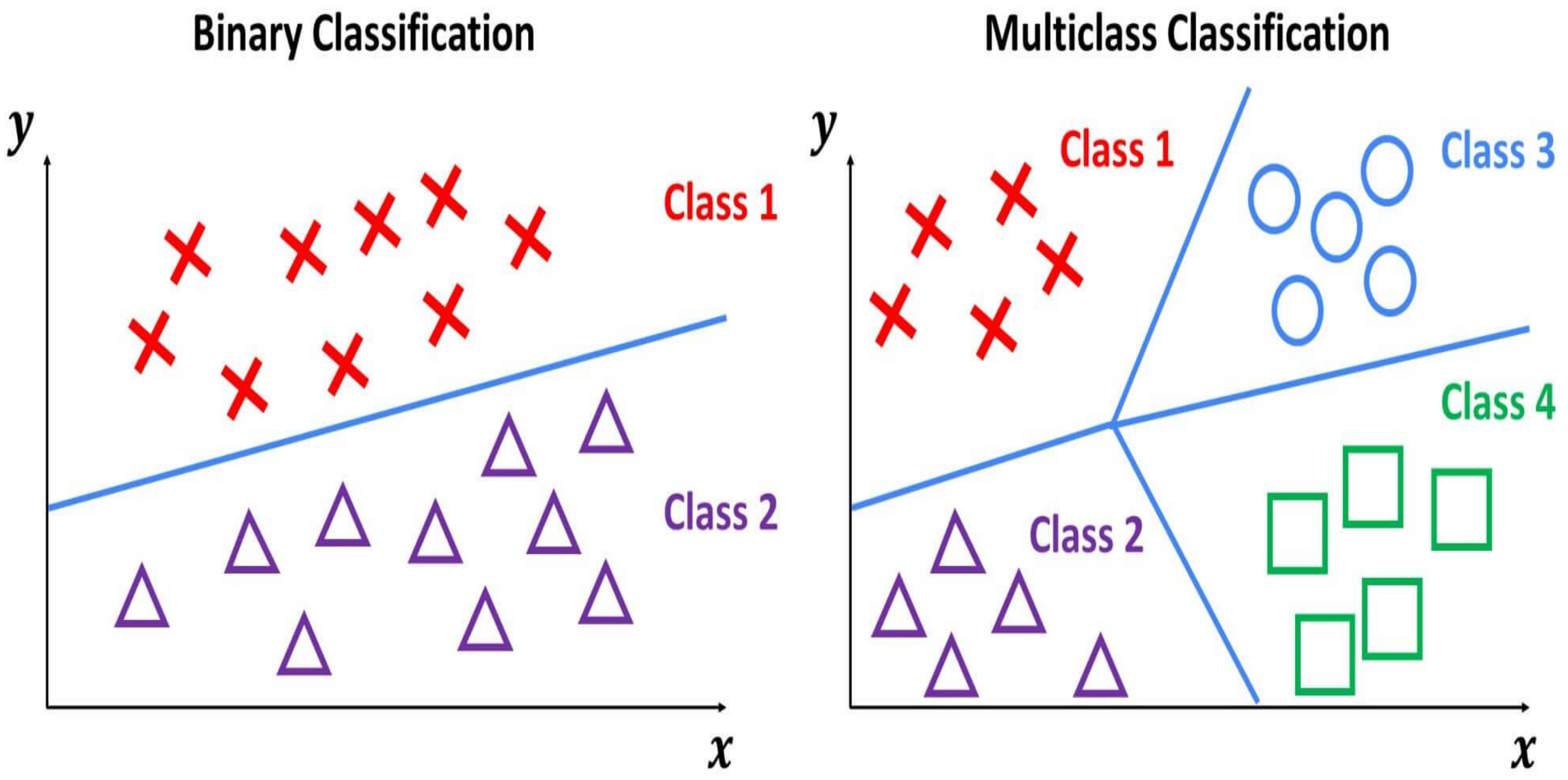


Discriminant: IF *income* > θ_1 AND *savings* > θ_2

THEN low-risk ELSE high-risk

Classification - Types

- Two main types: Binary and Multiclass classification.
- Binary classification is when the model has to predict one of two possible outcomes, such as yes or no, true or false, positive or negative, etc.
- Multiclass classification is when the model has to predict one of more than two possible outcomes, such as red, green, or blue, dog, cat, or bird, etc.
- Binary classification algorithms: logistic regression, support vector machine, decision tree, etc.
- Multiclass classification algorithms: K-nearest neighbors, Naive Bayes, Random forest, etc



Big Data Trunk

Classification

- Classification: Supervised machine learning where the model tries to predict the correct label or category of a given input data.
- The model is fully trained using the training data, and then it is evaluated on test data before being used to perform prediction on new unseen data.
- The main objective of classification is to build a model that can accurately assign a label or category to a new observation based on its features.
- Classification can be used for various applications such as image recognition, spam detection, sentiment analysis, medical diagnosis, etc.

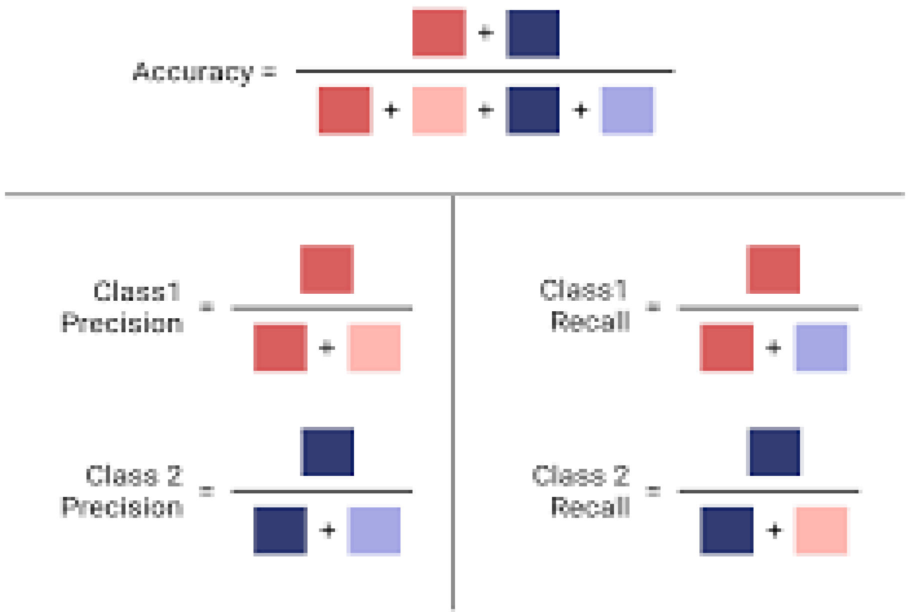
Classification: Applications

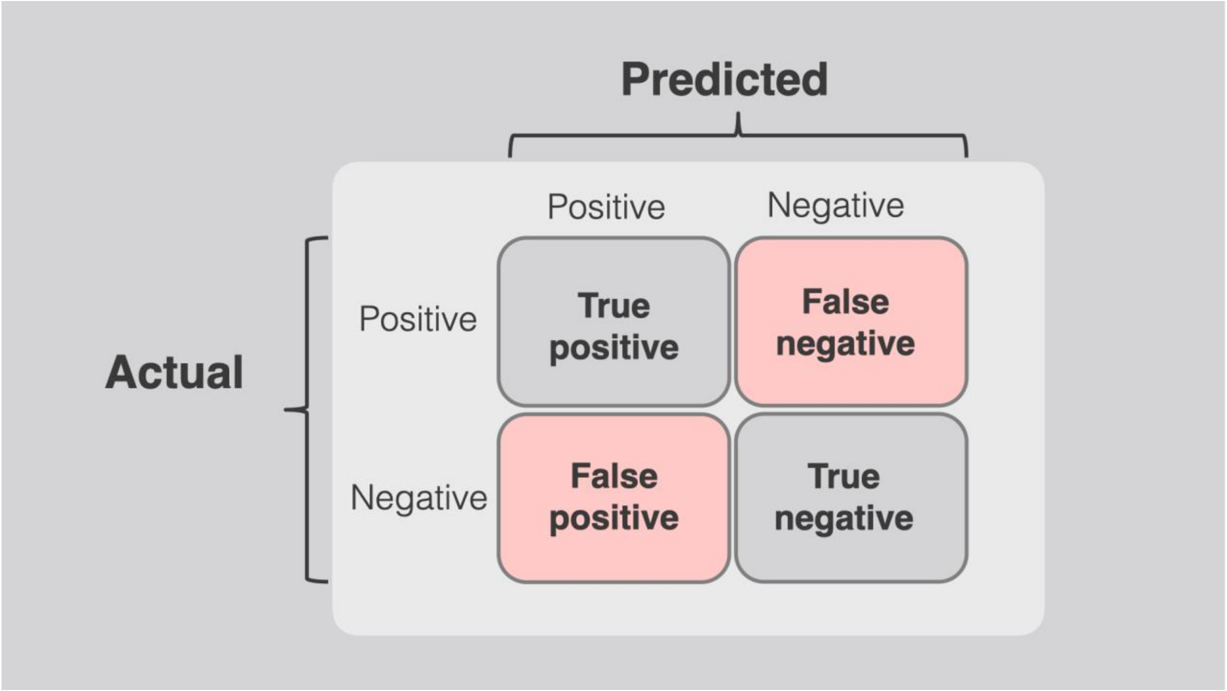
- Face recognition: Pose, lighting, occlusion (glasses, beard), make-up, hair style
- Character recognition: Different handwriting styles.
- Speech recognition: Temporal dependency.
- Use of a dictionary or the syntax of the language.
- Sensor fusion: Combine multiple modalities; eg, visual (lip image) and acoustic for speech
- Medical diagnosis: From symptoms to illnesses

Classification - Evaluation

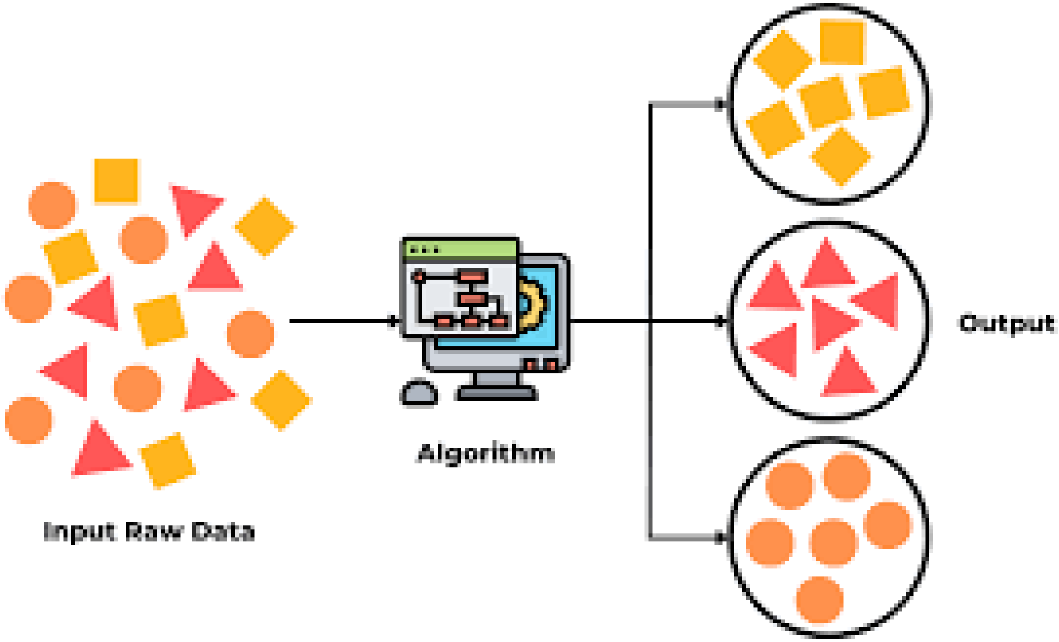
- Compare the predicted labels with the actual labels and measure how well the model can classify the data.
- Metrics that can be used: Accuracy, Precision, Recall, F1-score, Confusion Matrix, ROC curve, etc.
- Accuracy is the ratio of correctly predicted observations to the total number of observations. It measures how often the model predicts the correct label.
- Precision is the ratio of correctly predicted positive observations to the total number of predicted positive observations. It measures how precise the model is when it predicts a positive label.

Metrics Explained..





Unsupervised Learning



Unsupervised Learning

- Unsupervised learning is a type of machine learning where the algorithm learns from unlabeled data to find hidden patterns or structure in the data without any supervision.
- Unlabeled data means that the input data is not tagged with the correct output or the desired outcome.
- The algorithm tries to discover the underlying features or characteristics of the data that can help to group or cluster the data into meaningful categories or associations.
- Unsupervised learning can be used for various applications such as anomaly detection, customer segmentation, dimensionality reduction, recommendation systems, etc.

Unsupervised Learning - Types

- Two main categories: Clustering and Association.
- Clustering: The algorithm groups the data objects into clusters based on their similarities and differences.
- Association: The algorithm finds the rules or patterns that describe the relationship between the data items or variables.
- Examples of clustering algorithms: K-means, Hierarchical clustering, DBSCAN, etc.
- Examples of Association algorithms: Apriori, Eclat, FP-Growth, etc.

Clustering

- Clustering is a technique used to group similar objects or data points together based on their characteristics or attributes
- It is a fundamental unsupervised learning task that aims to identify hidden structures or patterns within a dataset without any prior knowledge or labels
- Clustering helps to identify patterns and structure in data, making it easier to understand and analyze
- Clustering can be used for various applications such as anomaly detection, customer segmentation, image segmentation, recommendation systems, etc

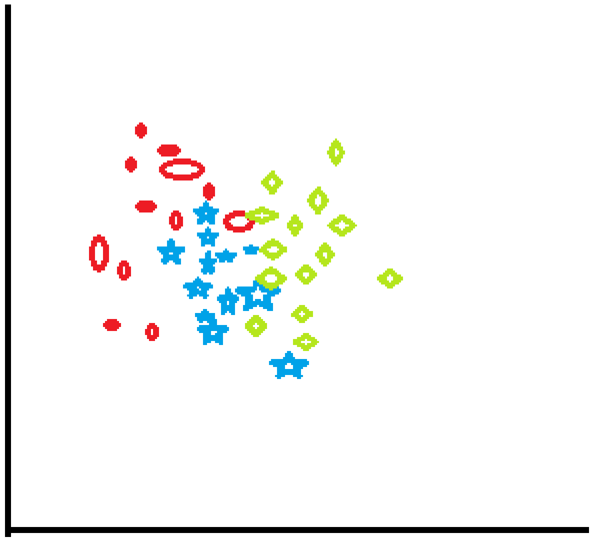


fig 1: before applying
k-means clustering

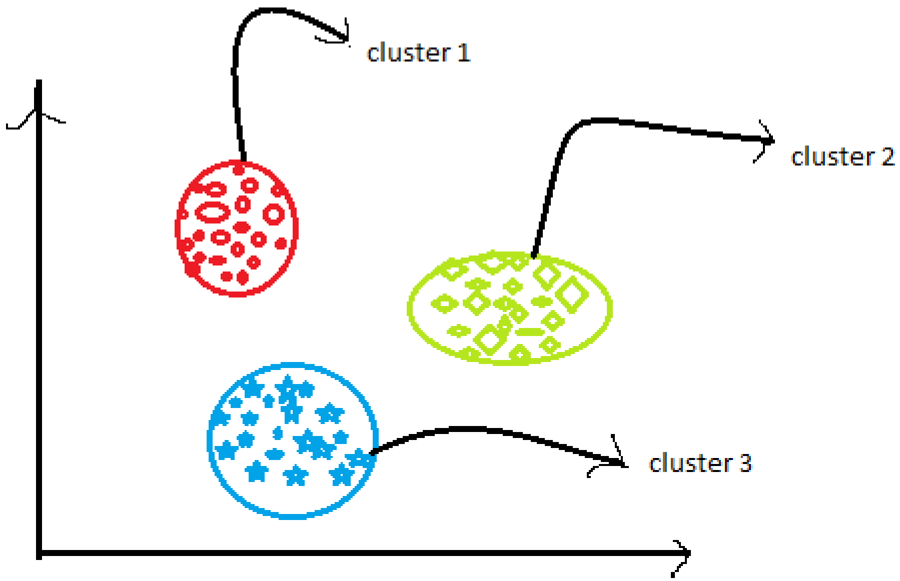
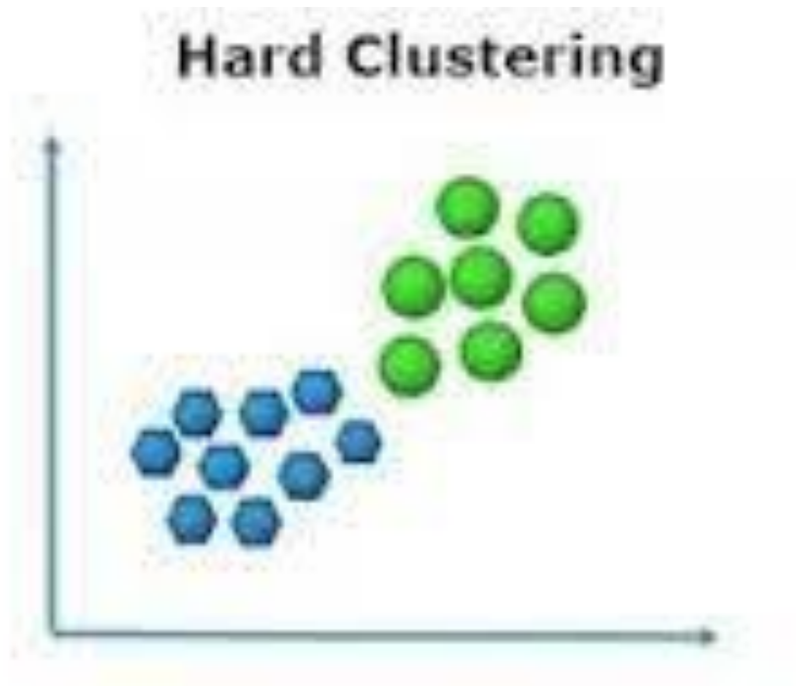
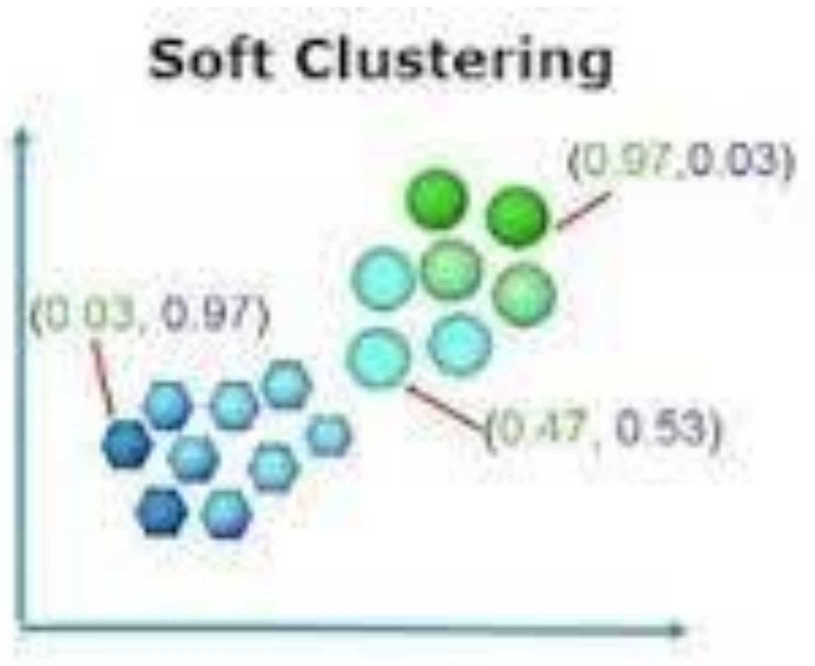


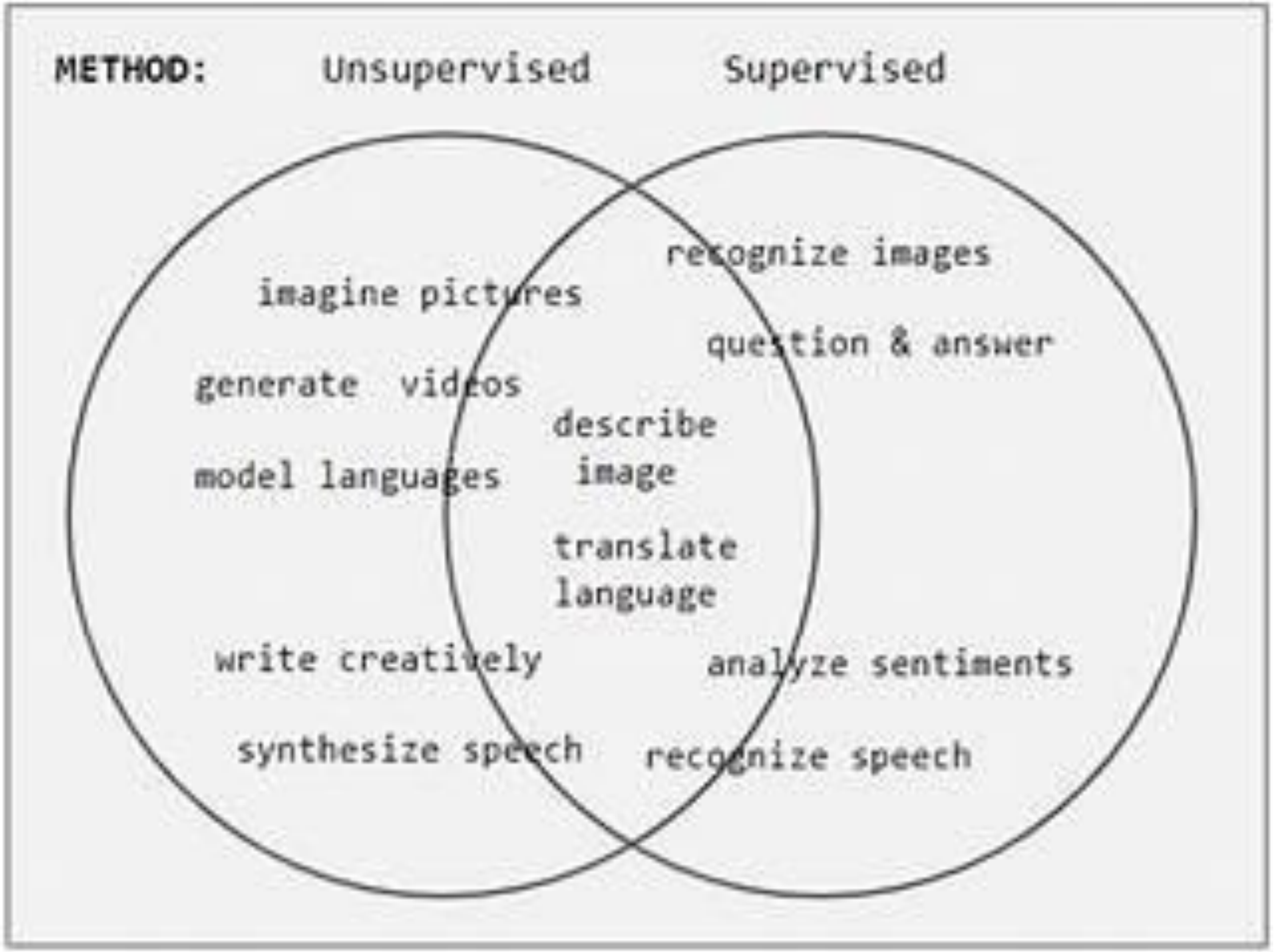
fig 2: After applying K-
means clustering

Types of Clustering

- Two types of Clustering: Hard clustering and soft clustering
- Hard clustering is when the model assigns each data point to one and only one cluster, such as k-means, hierarchical clustering, DBSCAN, etc
- Soft clustering is when the model assigns each data point to one or more clusters with some degree of membership, such as fuzzy clustering, Gaussian mixture model, etc.
- Clustering can also be categorized based on the criteria or assumptions used to form the clusters, such as partitioning, density-based, distribution-based etc.


Hard Clustering

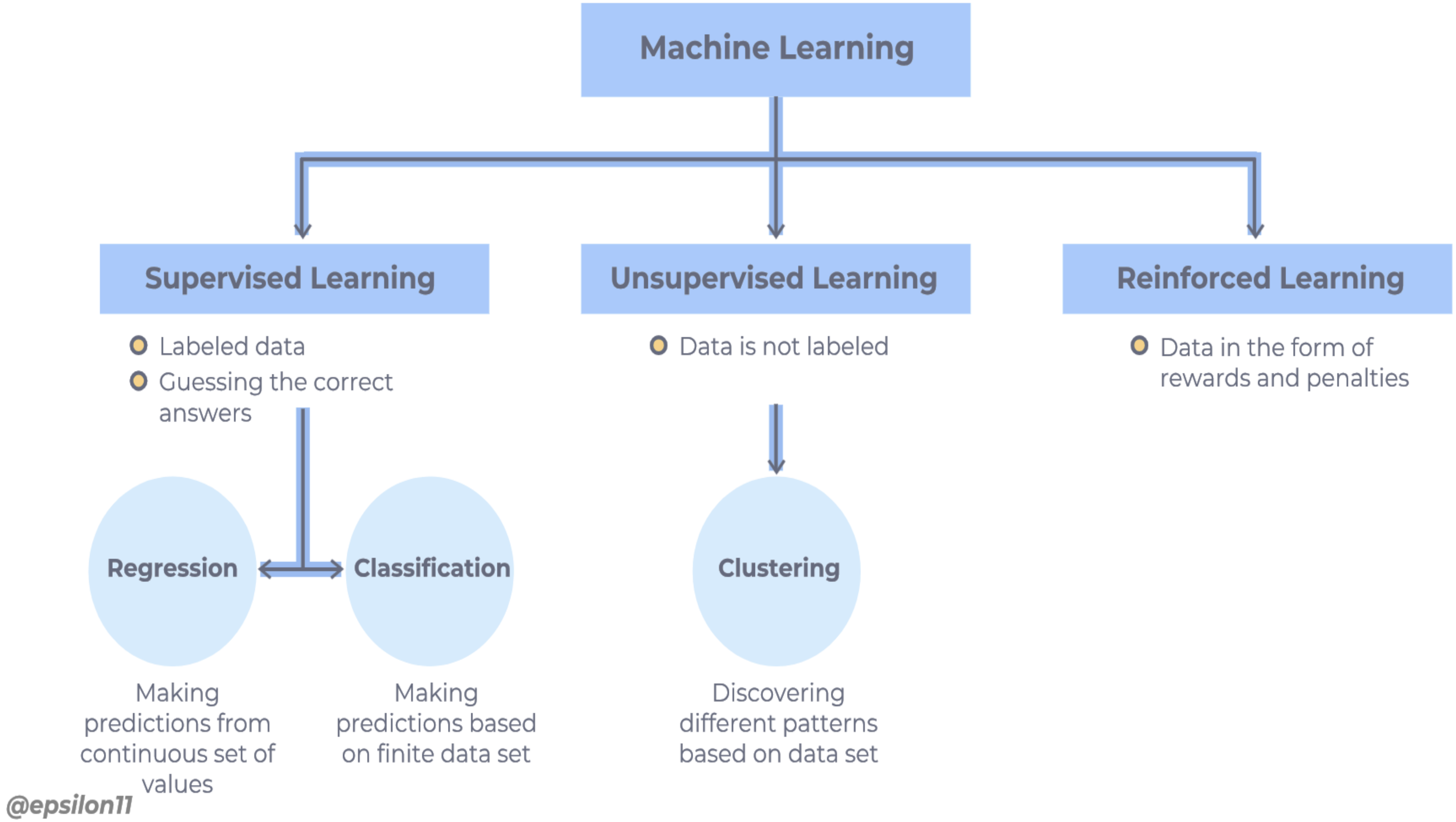

Soft Clustering
(0.97,0.03)
(0.03, 0.97)
(0.47, 0.53)


METHOD:
Unsupervised
Supervised
imagine pictures
generate videos
model languages
write creatively
synthesize speech
describe image
translate language
recognize images
question & answer
analyze sentiments
recognize speech

Classification - Evaluation

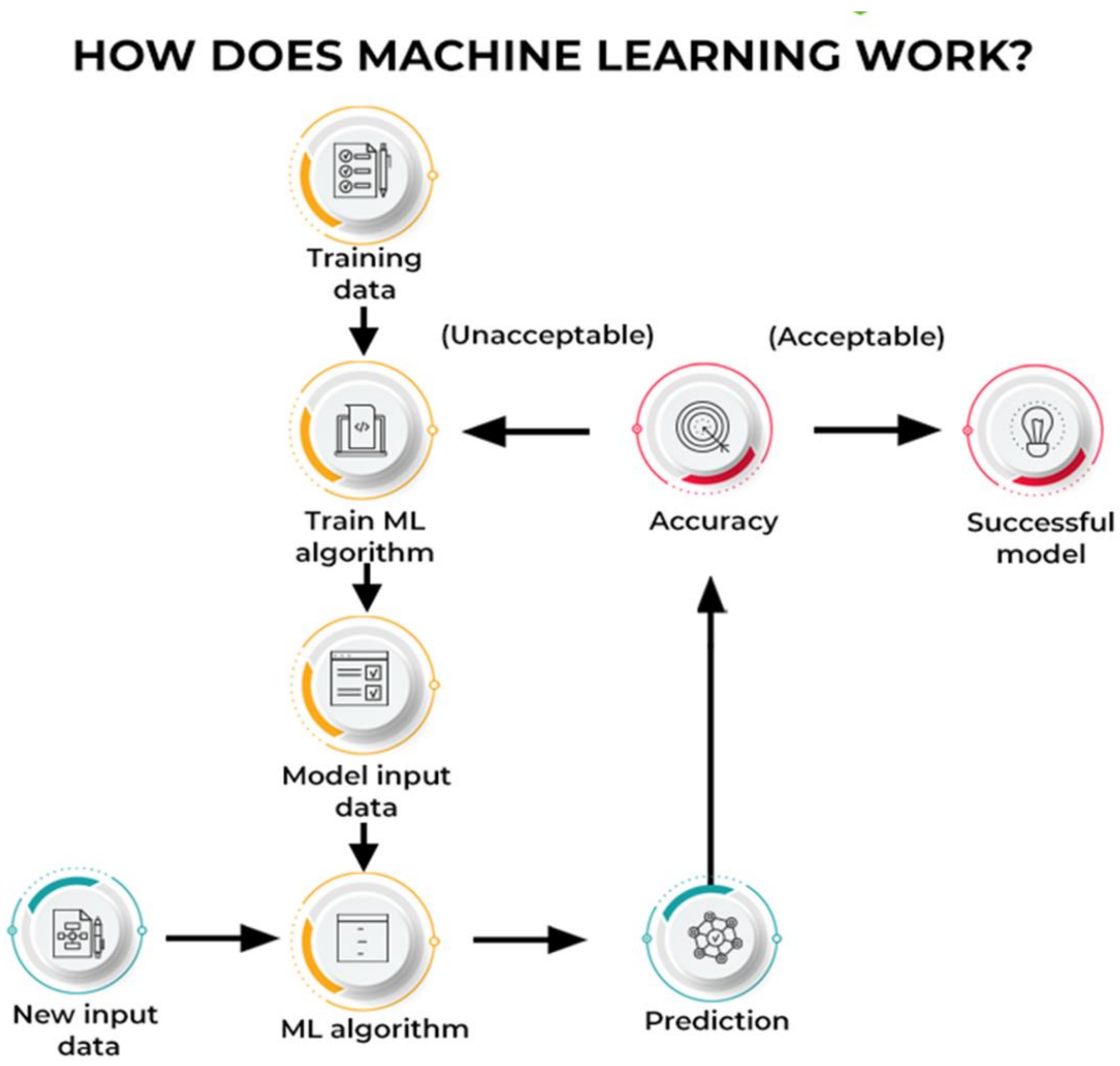
- Recall is the ratio of correctly predicted positive observations to the total number of actual positive observations. It measures how well the model can find all the positive labels.
- F1-score is the harmonic mean of precision and recall. It measures the balance between precision and recall.
- Confusion matrix is a table that shows the number of true positives, false positives, true negatives, and false negatives for each class. It helps to understand the errors made by the model.

Summary


Machine Learning
Supervised Learning
Unsupervised Learning
Reinforced Learning
Labeled data
Guessing the correct answers
Data is not labeled
Data in the form of rewards and penalties
Regression
Classification
Clustering
Making predictions from continuous set of values
Making predictions based on finite data set
Discovering different patterns based on data set
@epsilon11

Machine Learning

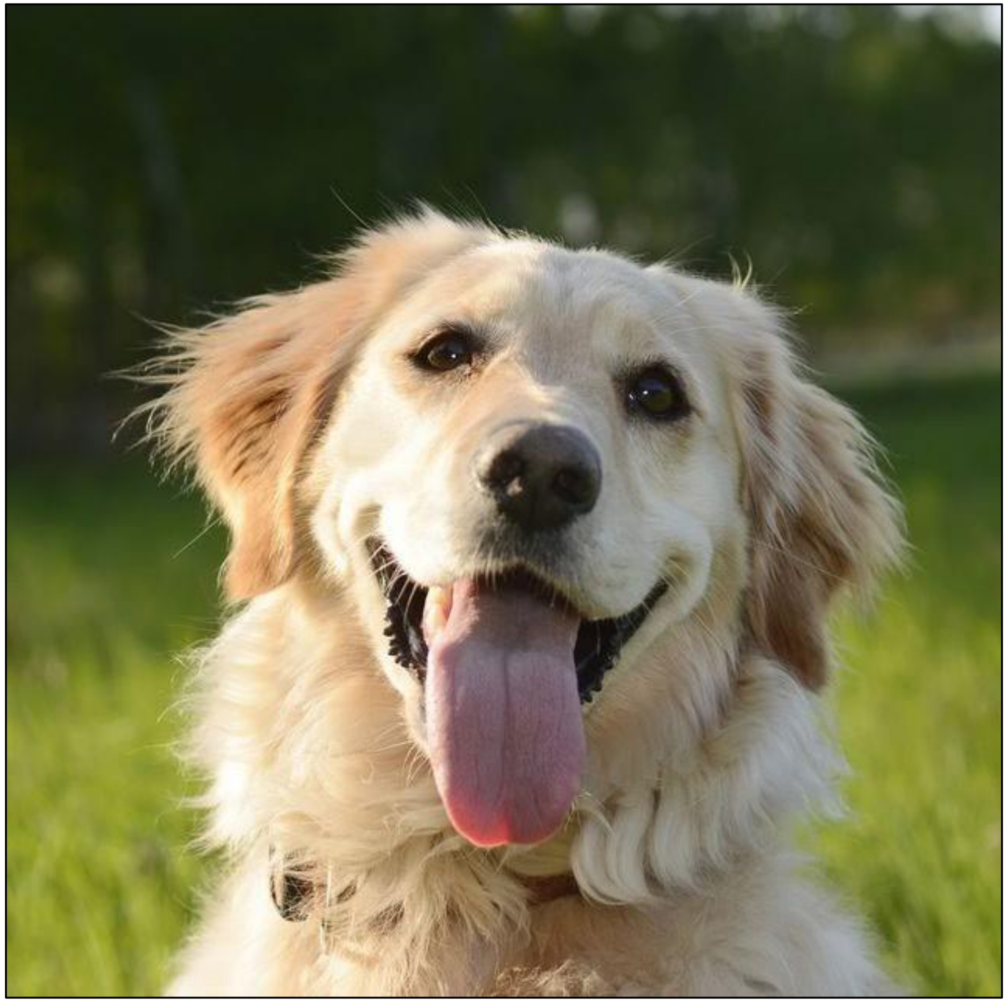
When & How?



With the right data and the right model, machine learning can solve many problems.

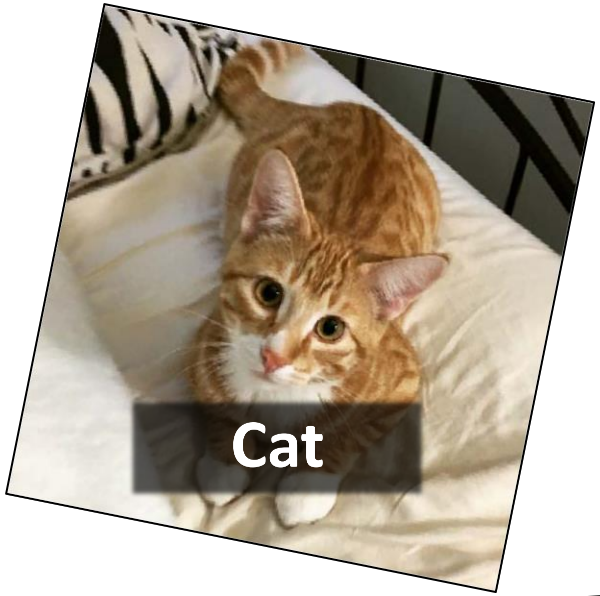
But finding the right data and training the right model can be difficult.

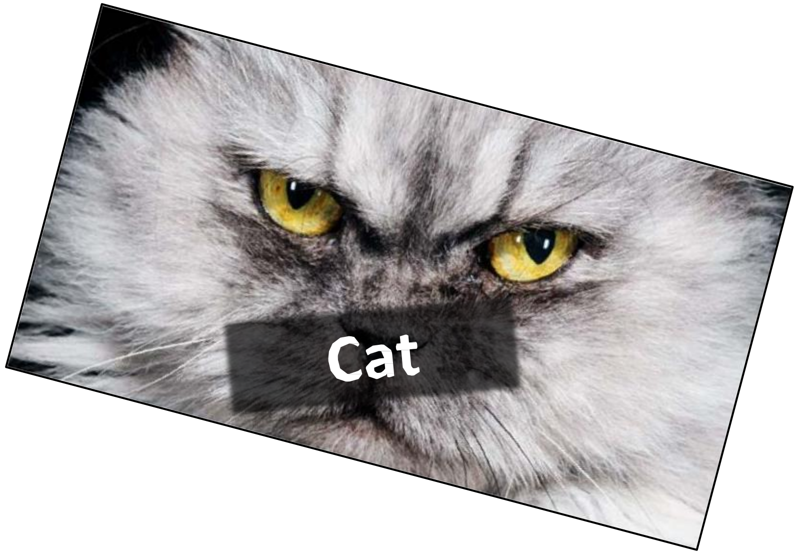
1. Define a problem

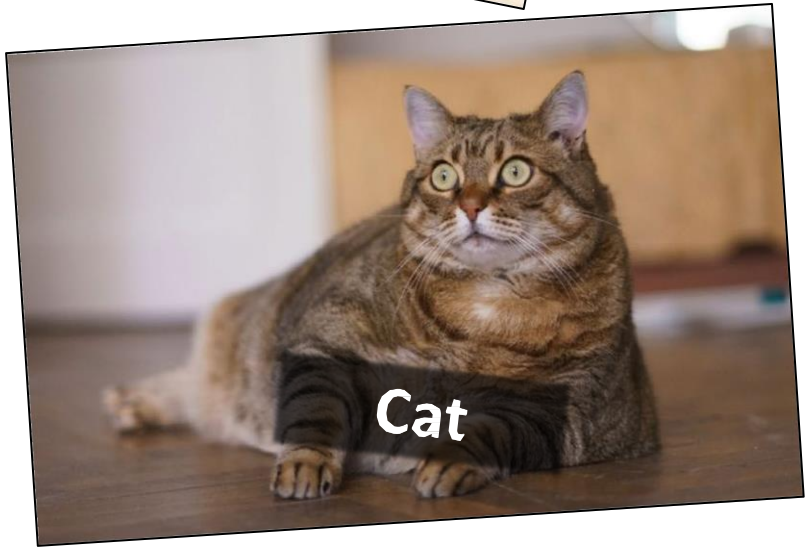


Big Data Trunk

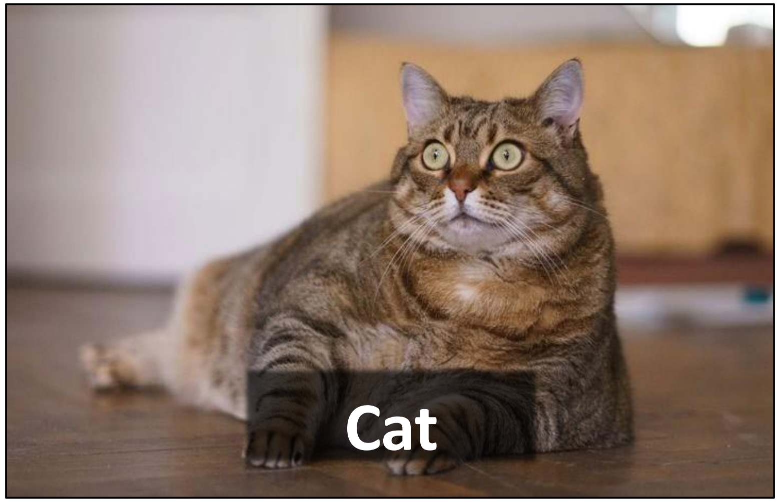
2.Get Data







3. Clean data.



4. Choose a model.

Dogs	Cats
<u>Always</u>	<u>Always</u>
<u>Sometimes</u>	<u>Sometimes</u>

5. Train the model.

5. Train the model.

Cat

5. Train the model.

Cat

5. Train the model.

Dog

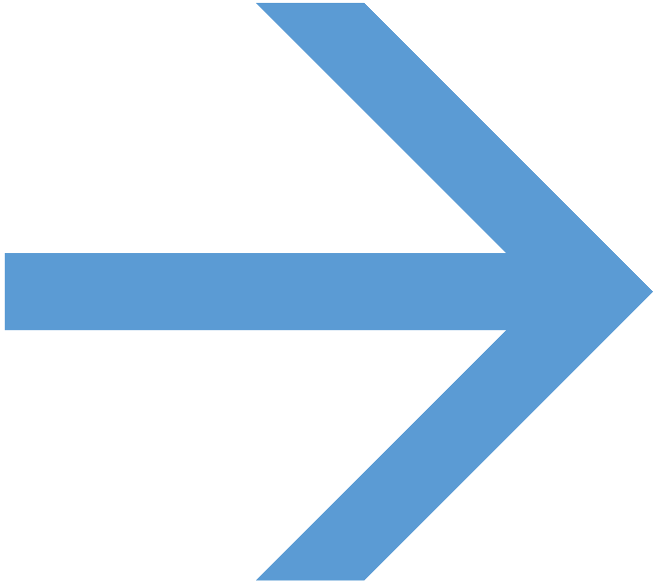
5. Train the model.

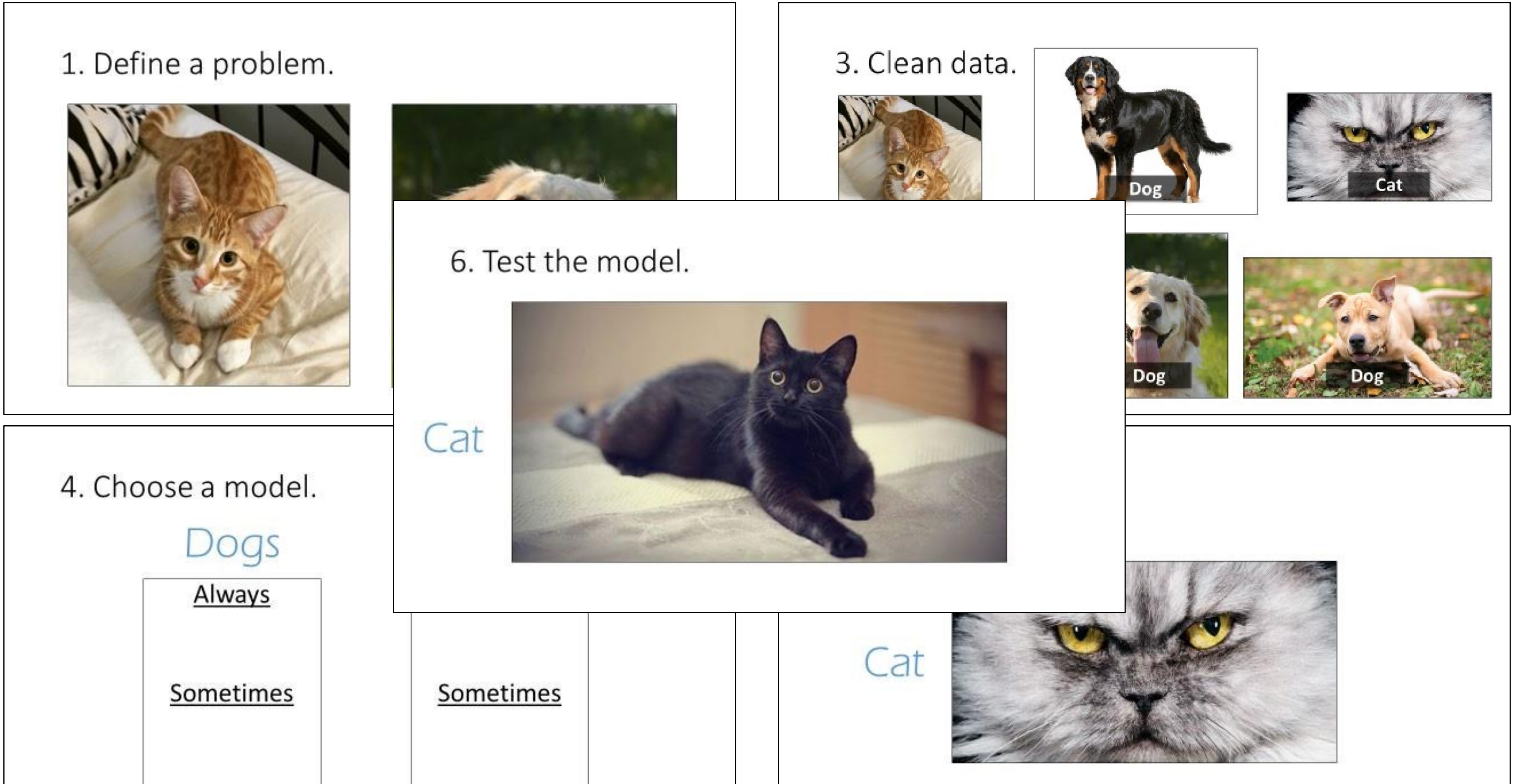
Dog

6. Test the model.

Cat

7. Deploy the model.




1. Define a problem.
3. Clean data.
Dog
Cat
Dog
Dog
6. Test the model.
Cat
4. Choose a model.
Dogs
Always
Sometimes
Sometimes
Cat

Building an ML Model

1. Goal?
2. Training data?
3. Model?
4. Accuracy?

Feedback Survey

https://app.performitiv.com/fv2/cisco/ceoevt/VC00508164

Quiz time

- https://create.kahoot.it/share/ai-ml-quiz-2/3ebe70f2-cbd1-4278-abcf-33661cdb70e2

Demos

- Supervised - Classification
  - Binary Classification - Decision tree
  - Multi Class Classification – K Nearest Neighbours

- Supervised – Regression
  - Linear Regression

- Unsupervised
  - Clustering

1. Decision Tree – Iris Data Set

https://colab.research.google.com/drive/1z_FSgk1QISJG55fTP6Z661ahZyTQJoRc

2. Naïve Bayes – Iris Data set

https://colab.research.google.com/drive/1ib8RH4R7K28PS9TxPqConnlSbCeiQBnj

- Linear Regression

https://colab.research.google.com/drive/1EMEuHxQj3Wz3sivut3tOqfq5cfuXbPmC#scrollTo=EZOJ--G_0z9H

- Clustering – Iris

https://colab.research.google.com/drive/14NYddEm1Wrqe61YT0C-jeY_iBXCBanYm?usp=sharing

Clustering – Breast cancer

https://colab.research.google.com/drive/1R0acguX2bIq7hP7RyUwrDCwsX7TzG_0U#scrollTo=y9-ry9VzzEWp

ML Use Cases