AI 330: Machine Learning

Fall 2023

Dr. Wessam EL-Behaidy
Associate Professor, Computer Science Department,
Faculty of Computers and Artificial Intelligence,
Helwan University.

Dr. Ensaf Hussein
Associate Professor, Computer Science Department,
Faculty of Computers and Artificial Intelligence,
Helwan University.

Agenda

- **Course Objectives**
- **Course Map**
- **Course topics**
- **References**
- **Grading Distribution**
- **Introduction to machine learning**
  - **What is machine learning (ML)?**
  - **ML Applications**
  - **ML Life Cycle**
  - **Types of ML**

Course Objectives

Fundamental Understanding: Provide students with a solid foundation in the basic principles, algorithms, and theories of Machine Learning, ensuring they grasp the core concepts.

→ How to implement this models?

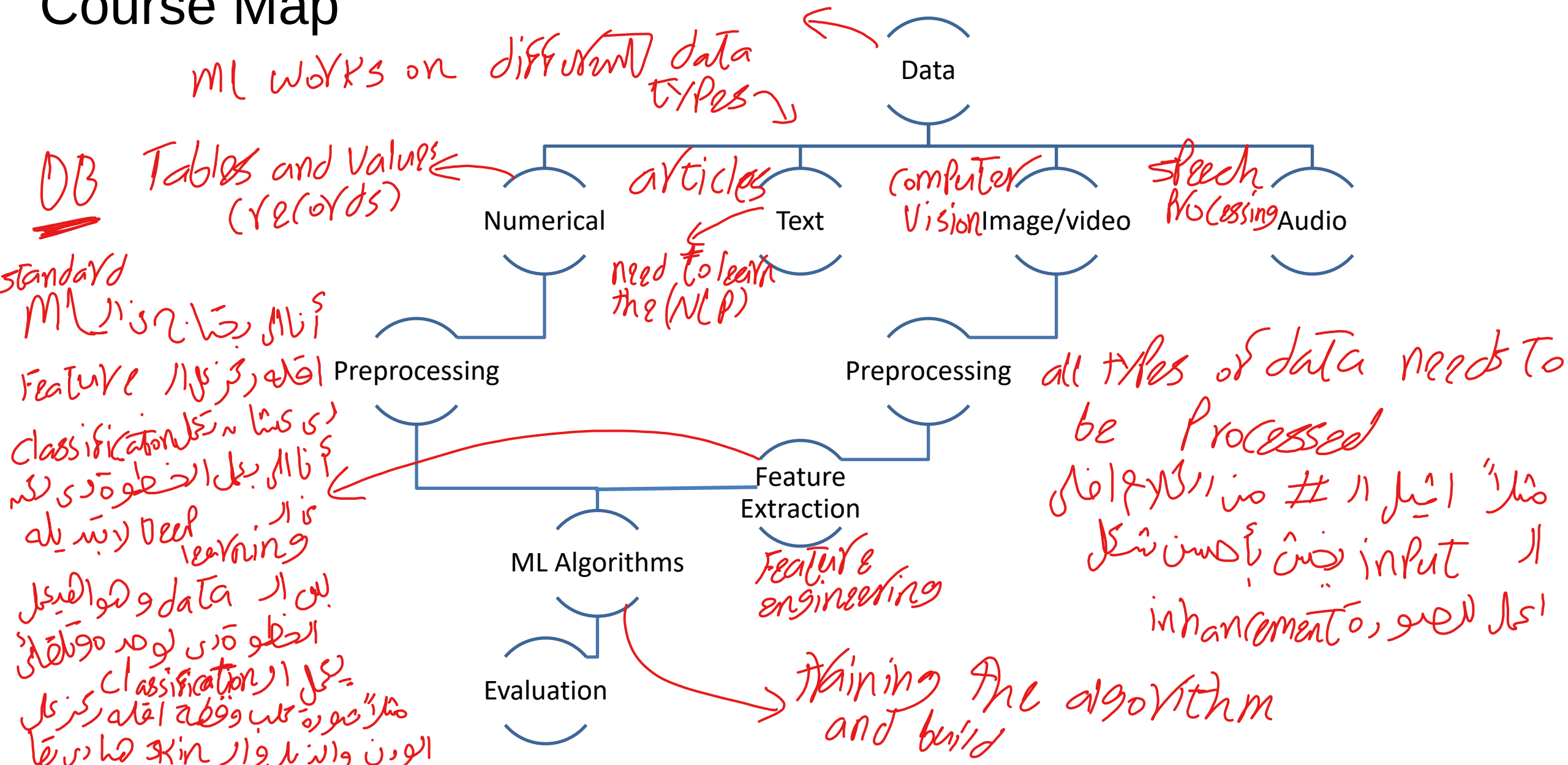
Practical Application: Enable students to apply machine learning techniques to real-world problems, fostering their ability to analyze data, build models, and make data-driven decisions.

Hands-On Experience: Offer practical, hands-on experience with machine learning tools and libraries, ensuring students can implement and experiment with various algorithms.

→ what is the best model for this data to implement?

Critical Thinking: Cultivate critical thinking skills by encouraging students to evaluate the suitability of machine learning solutions for different problems, understand ethical considerations, and explore the limitations of the technology.

Course Map


ML works on different data types
Data
DB Tables and values (records)
articles
computer Vision
Speech Processing
Numerical
Text
Image/video
Audio
need to learn the (NLP)
standard ML
Preprocessing
Preprocessing
Feature Extraction
ML Algorithms
Evaluation
Feature engineering
Training the algorithm and build
all types of data need to be Processed

Course Topics

Date	Lecture	Description
Week1	Introduction to Machine Learning	• Understanding the fundamentals of machine learning. • Types of machine learning: supervised, unsupervised, and more. • Applications and importance of machine learning.
Week2	Data Preprocessing & Feature Extraction	• Data cleaning and preprocessing techniques. • Feature selection and engineering.
Week3	Supervised Regressors	• Linear Regression with one variable. • Gradient Descent optimization
Week4	Supervised Regressors (cont.)	• Linear Regression with Multiple Variables. • Handling categorical data.
Week5	Supervised classifiers	• Logistic Regression. • Regularization techniques (L1 and L2).
Week6	k-Nearest Neighbors (k-NN) algorithm	• Understanding the k-NN algorithm. • Choosing the optimal k value.

Course Topics

Date	Lecture	Description
Week8	Model Evaluation and Hyperparameter Tuning	• Evaluation metrics for regression and classification. • Cross-validation and hyperparameter tuning.
Week9	Ensemble Methods	• Bagging and Random Forests. • Boosting: AdaBoost and Gradient Boosting.
Week10	Introduction to unsupervised learning.	• Introduction to unsupervised learning. • Clustering techniques: k-Means.
Week11	Dimensionality Reduction	• Principal Component Analysis (PCA). • Applications and interpretation of PCA
Week12		Project Discussion

This course focus on the standard ML preprocessing, feature abstraction, supervised and unsupervised ML

References

Advanced ML → Neural networks, CNN, RNN, transformal.

Lectures (Course slides) are based on :

- Machine Learning Specialization https://www.coursera.org/specializations/machine-learning-introduction at Stanford University (**Andrew Ng**)
- **A Course in Machine Learning by Hal Daumé III**

Practical Labs based on:

Machine learning A to Z: **Kirill Eremenko ©superdatascience**

Grading Distribution

- Project: (15%)

  Code “github”, pitching video, presentation.

- Programming Assignment (DataCamp)(10%)
- Practical Assessment (10%)
- Midterm exam (15%)
- Final exam (50%)

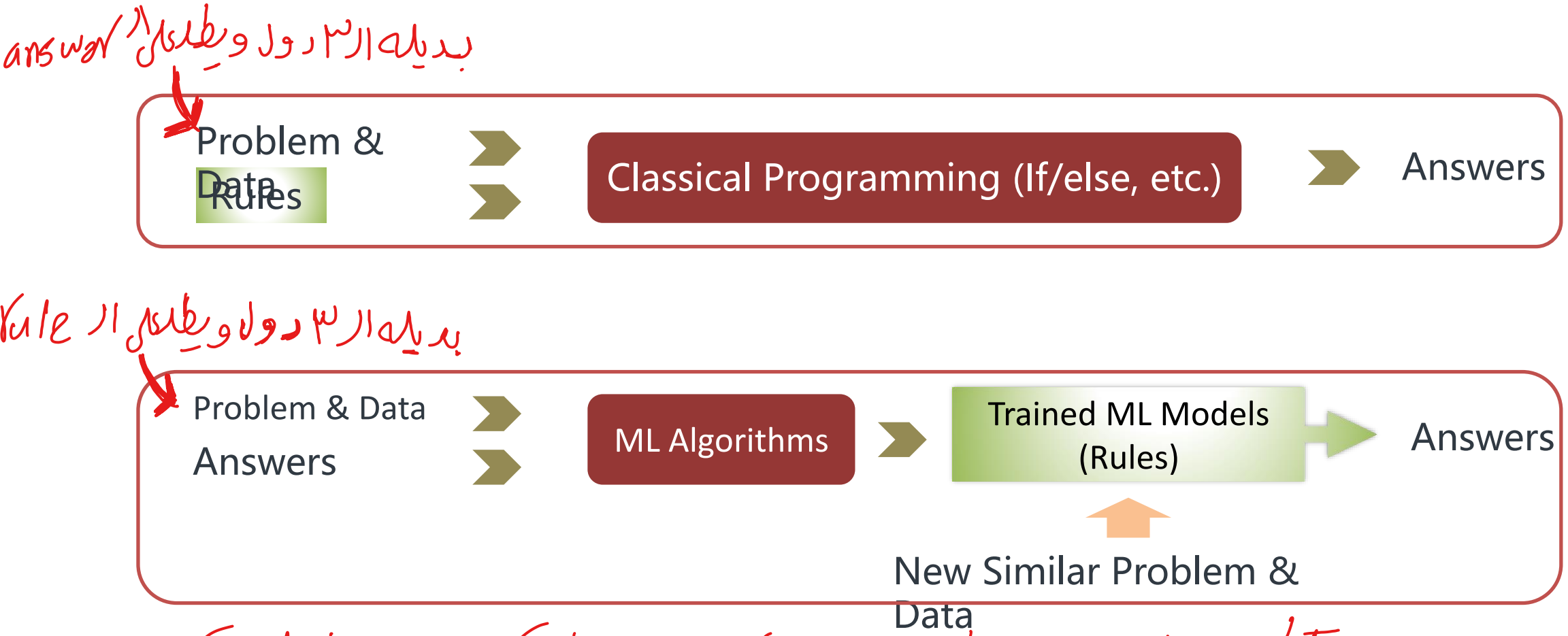
Join **Machine Learning(Fall2023)** Team Class

To communicate with us and get course materials

Team class code:

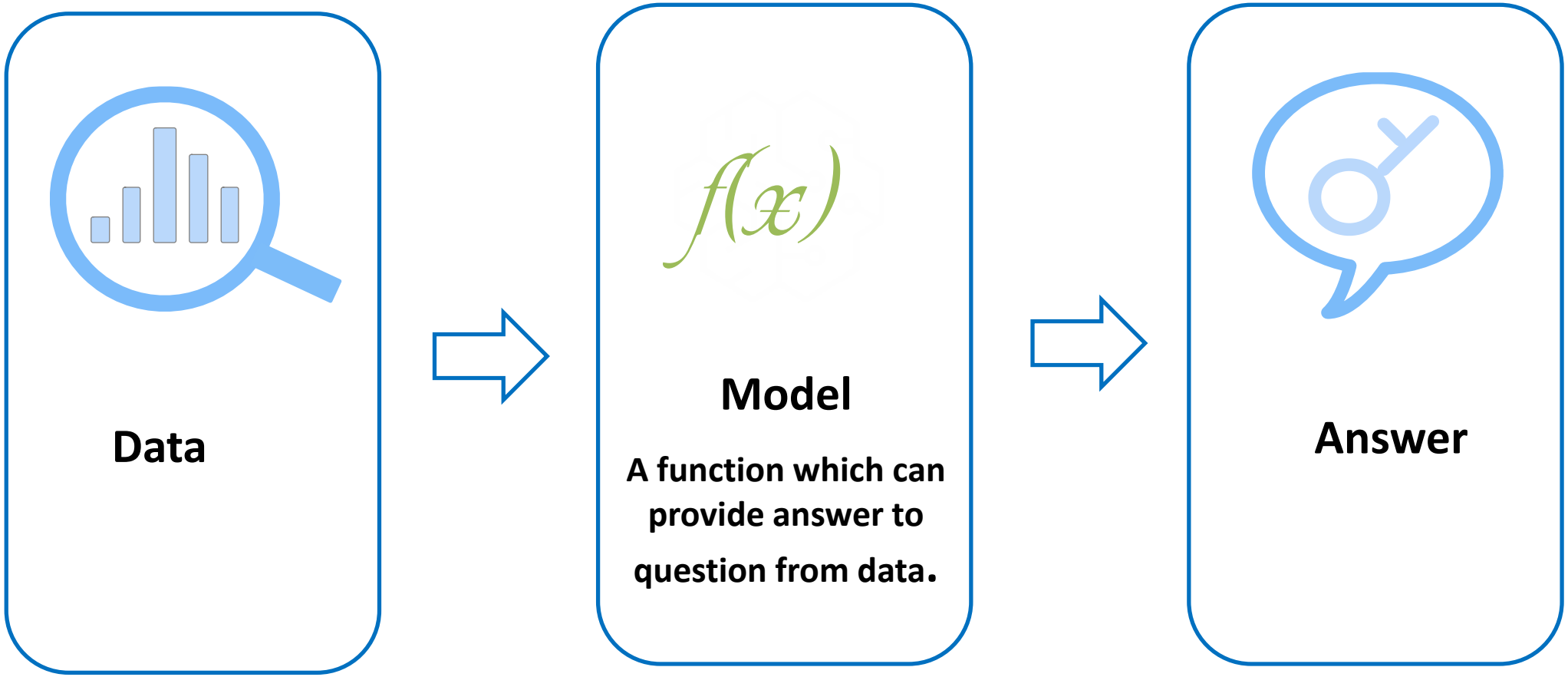
hrglst0

What is Machine Learning (ML)?



What is a Model?

↳ A Function



What is Machine Learning (ML)?

> "The field of study that gives computers the ability to learn without being explicitly programmed"

Arthur Samuel
Pioneer of AI research

ML Terminologies

ML optimizes on **predictive** performance, while statistics places importance on **interpretability** and parsimony/simplicity.

ML / Statistics Jargon	Definition
Label/Target/Output Variable/"y"	The results to predict
Feature/Input Variable/"x"	Input data to help make predictions
Feature Engineering / Transformation	Reshaping raw input data to give more insights
Dimensionality / [1st d, 2nd d, … , nth d]	Number of features
Model Weights / Parameters	A set of numbers embedded in a model to make predictions
Model Training	Applying optimization techniques to find the "best" set of model weights

→ to find the optimal value of parameters

→ extract the features from the data such as images, (يحول) transformate the raw input (data) to features.

Machine Learning Applications

(techniques)

ML Applications

like google , amazon

كأنك لما search على حاجة بيجيبلي الأول الحاجة الأقرب حاجة من الكلمة الي كاتبها الأول فالأبعد

Problem type	Description	Example
Ranking	Helping users find the most relevant *items*	Ranking algorithm within Amazon Search 
Recommendation	Giving users the *items* they may be most interested in	
Classification	Figuring out what category does an *item* belongs to	
Regression	Predicting a numerical value of an *item*	
Clustering	Putting similar *items* together	
Anomaly Detection	Finding uncommon *items*	

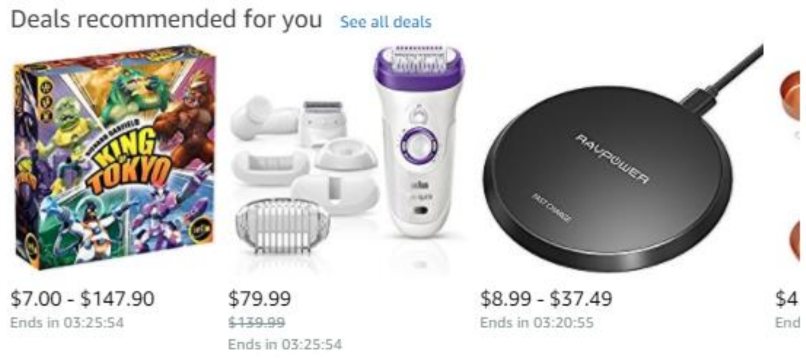
زي مثلاً Facebook لما مثلاً x صديق لـ y وأنا صاحب x هيد recommend y ليا
مثلاً أول لما افتح اي online market أو YouTube بيـrecommend الحاجات المشابهة الى بحثت عنها كتير قبل كدا .

ML Applications

Problem type	Description
Ranking	Helping users find the most relevant *items*
Recommendation	Giving users the *items* they may be most interested in
Classification	Figuring out what category does an *item* belongs to
Regression	Predicting a numerical value of an *item*
Clustering	Putting similar *items* together
Anomaly Detection	Finding uncommon *items*

Example

Recommendations across the website



Amazon's Choice

مثلاً أخبار رياضة والسياسة والفن ٣ مختلفين

Problem type	Description
Ranking	Helping users find the most relevant *items*
Recommendation	Giving users the *items* they may be most interested in
Classification	Figuring out what category does an *item* belongs to
Regression	Predicting a numerical value of an *item*
Clustering	Putting similar *items* together
Anomaly Detection	Finding uncommon *items*

محددة وليشا عدد محدود

Example

Product classification for our catalog

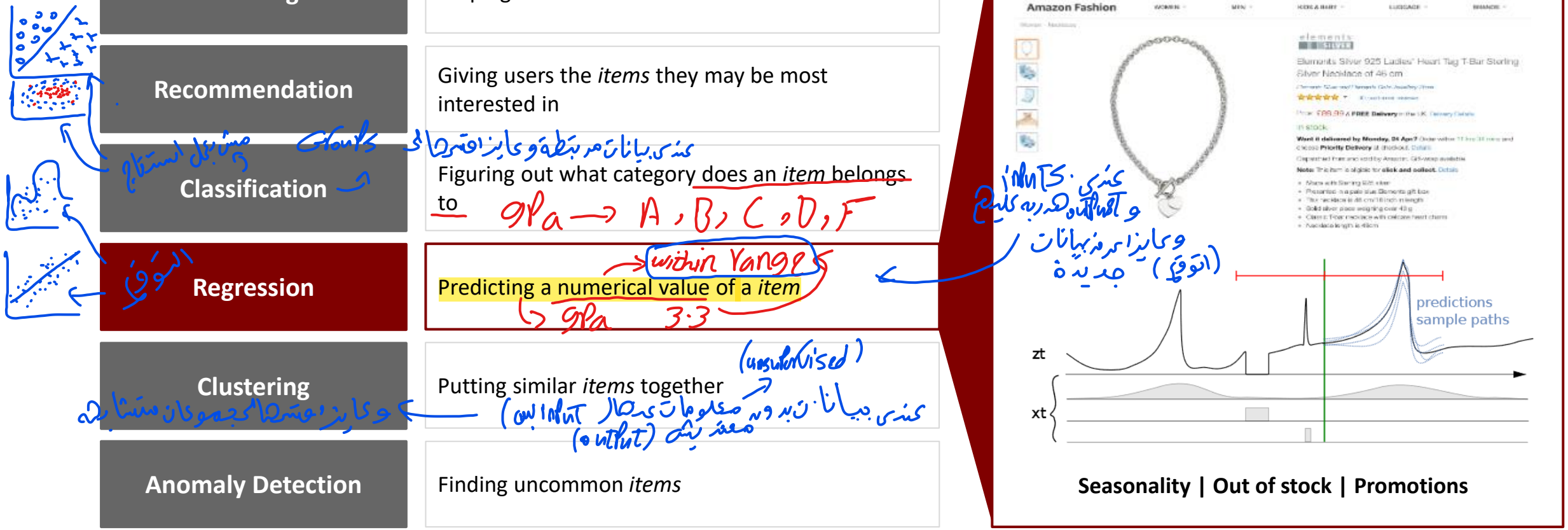
High-Low Dress

Straight Dress

Striped Skirt

Graphic Shirt

Problem type	Description	Example
Ranking	Helping users find the most relevant *items*	Predicting sales for specific ASINs
Recommendation	Giving users the *items* they may be most interested in	
Classification	Figuring out what category does an *item* belongs to	
Regression	Predicting a numerical value of a *item*	
Clustering	Putting similar *items* together	
Anomaly Detection	Finding uncommon *items*	



Seasonality | Out of stock | Promotions

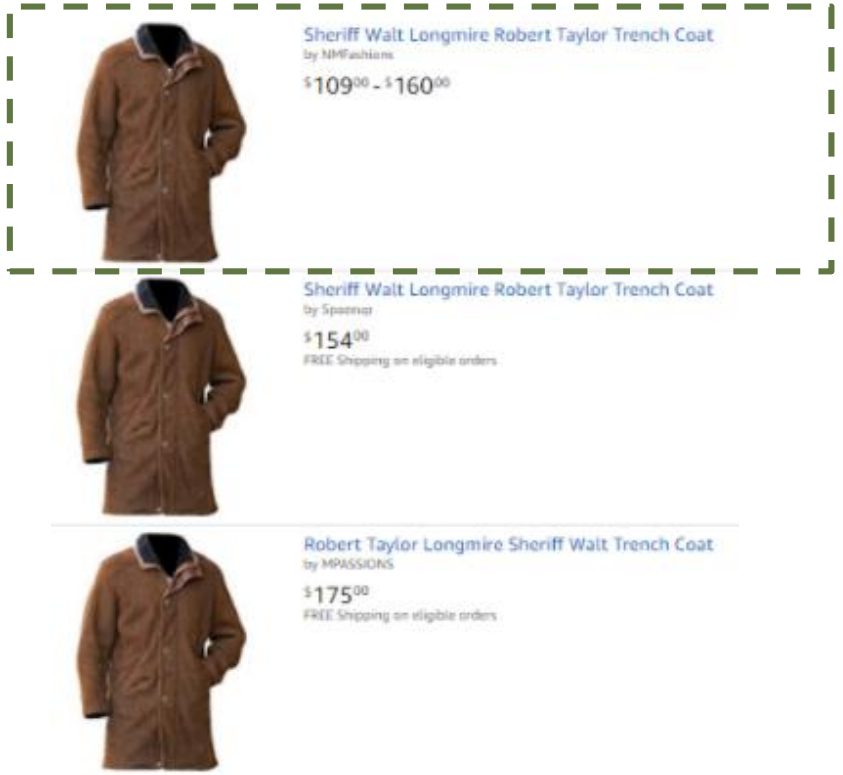
ML Applications

Grouping similar items together
category wise

Problem type	Description
Ranking	Helping users find the most relevant *items*
Recommendation	Giving users the *items* they may be most interested in
Classification	Figuring out what category does an *item* belongs to
Regression	Predicting a numerical value of an *item*
Clustering	Putting similar *items* together
Anomaly Detection	Finding uncommon *items*

Example

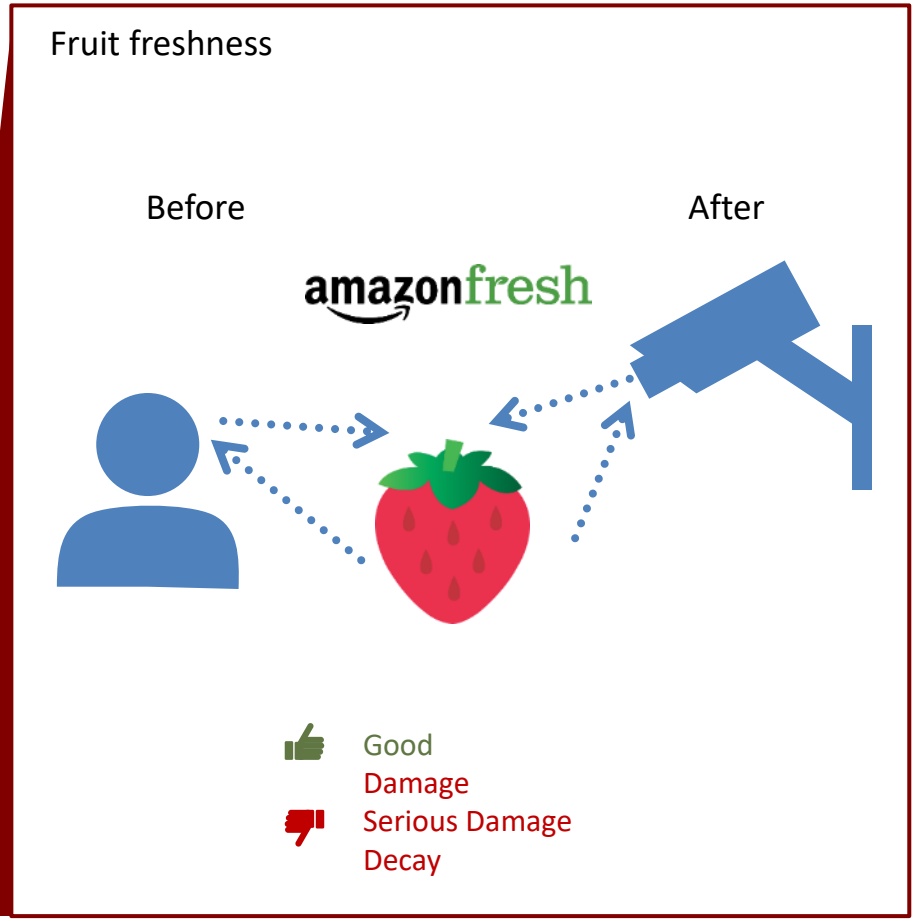
Close-matching for near-duplicates



ML Applications

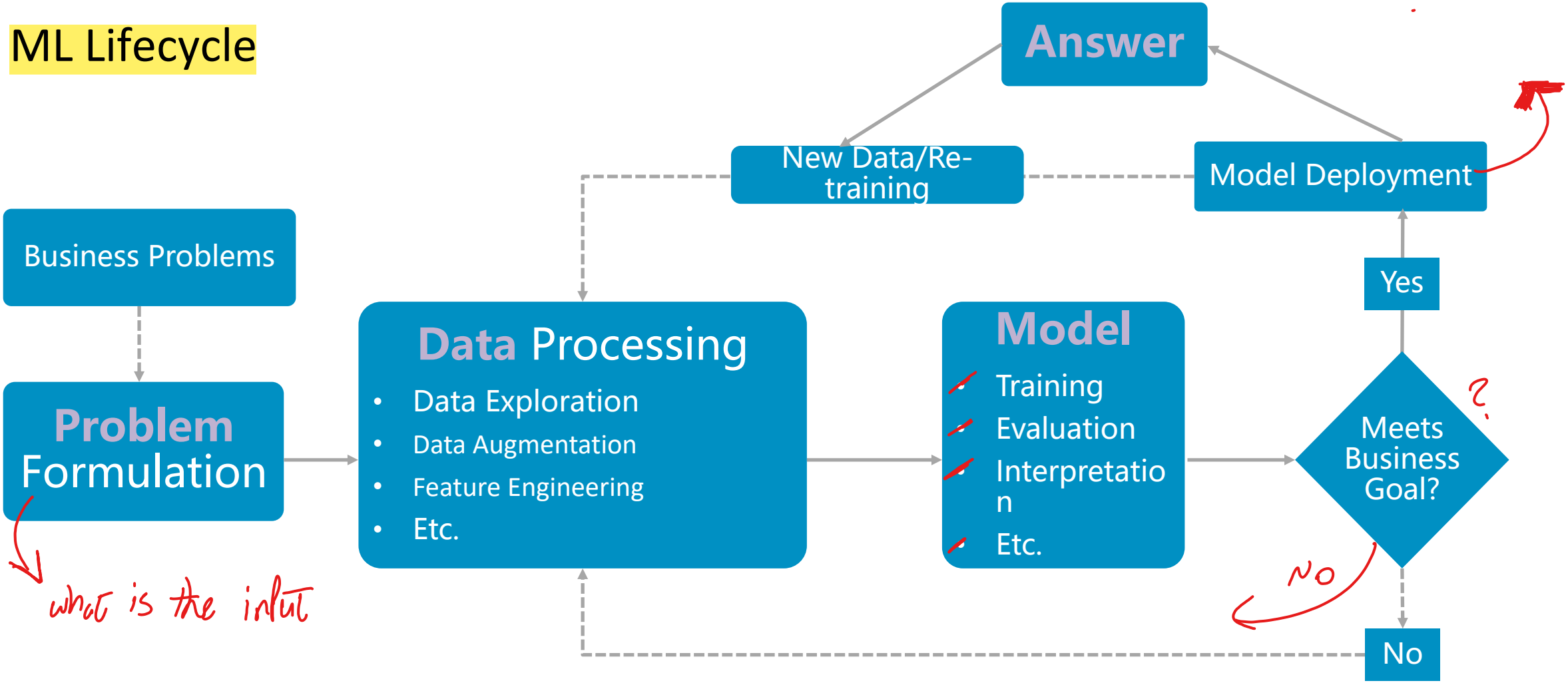
Problem type	Description
Ranking	Helping users find the most relevant *items*
Recommendation	Giving users the *items* they may be most interested in
Classification	Figuring out what category does an *item* belongs to
Regression	Predicting a numerical value of an *item*
Clustering	Putting similar *items* together
Anomaly Detection	Finding uncommon *items*

Example

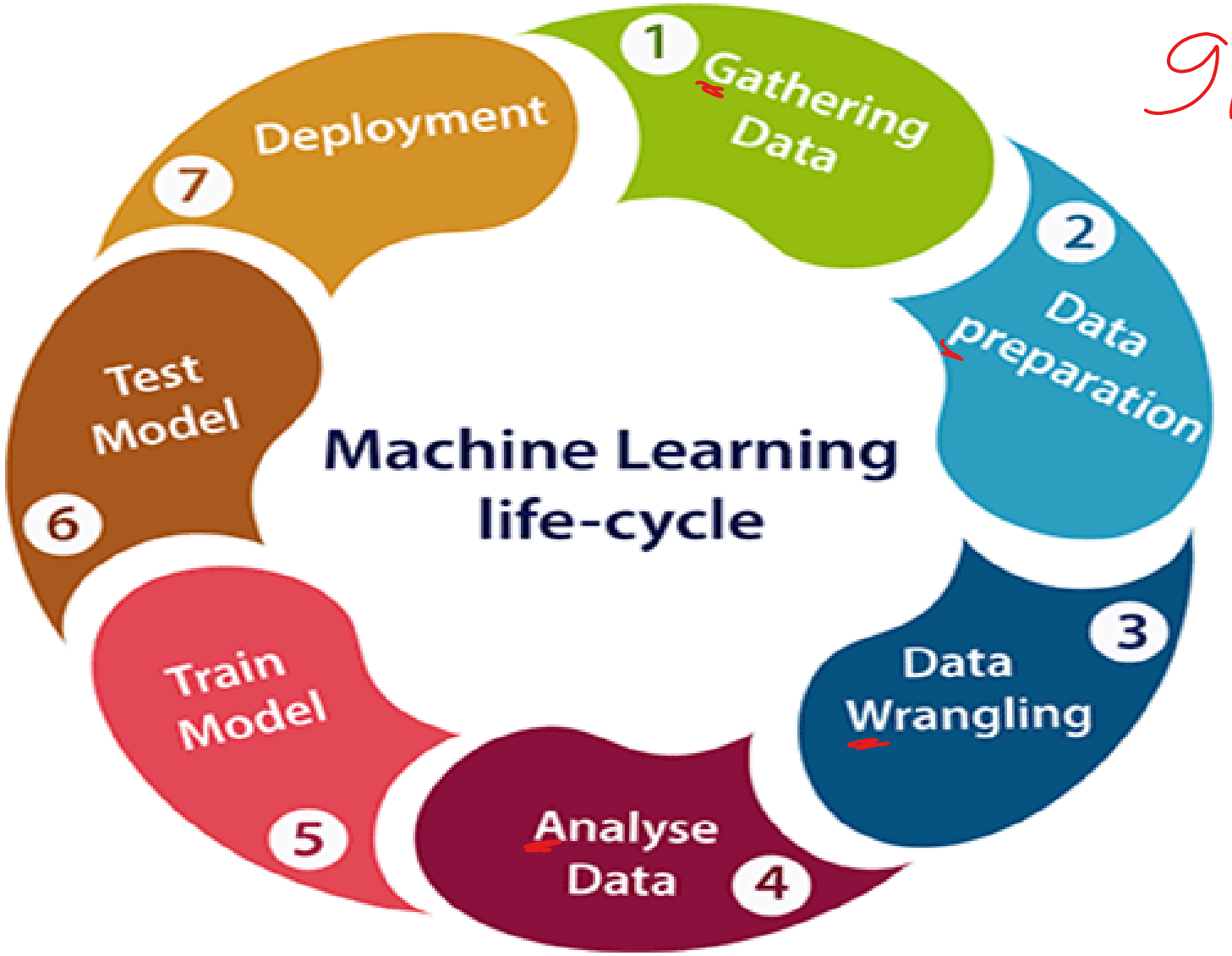


Machine Learning Life Cycle

ML Lifecycle



Problem → Data → Model → Answer



https://www.javatpoint.com/machine-learning-life-cycle

1- Gathering data (Acquiring Dataset)

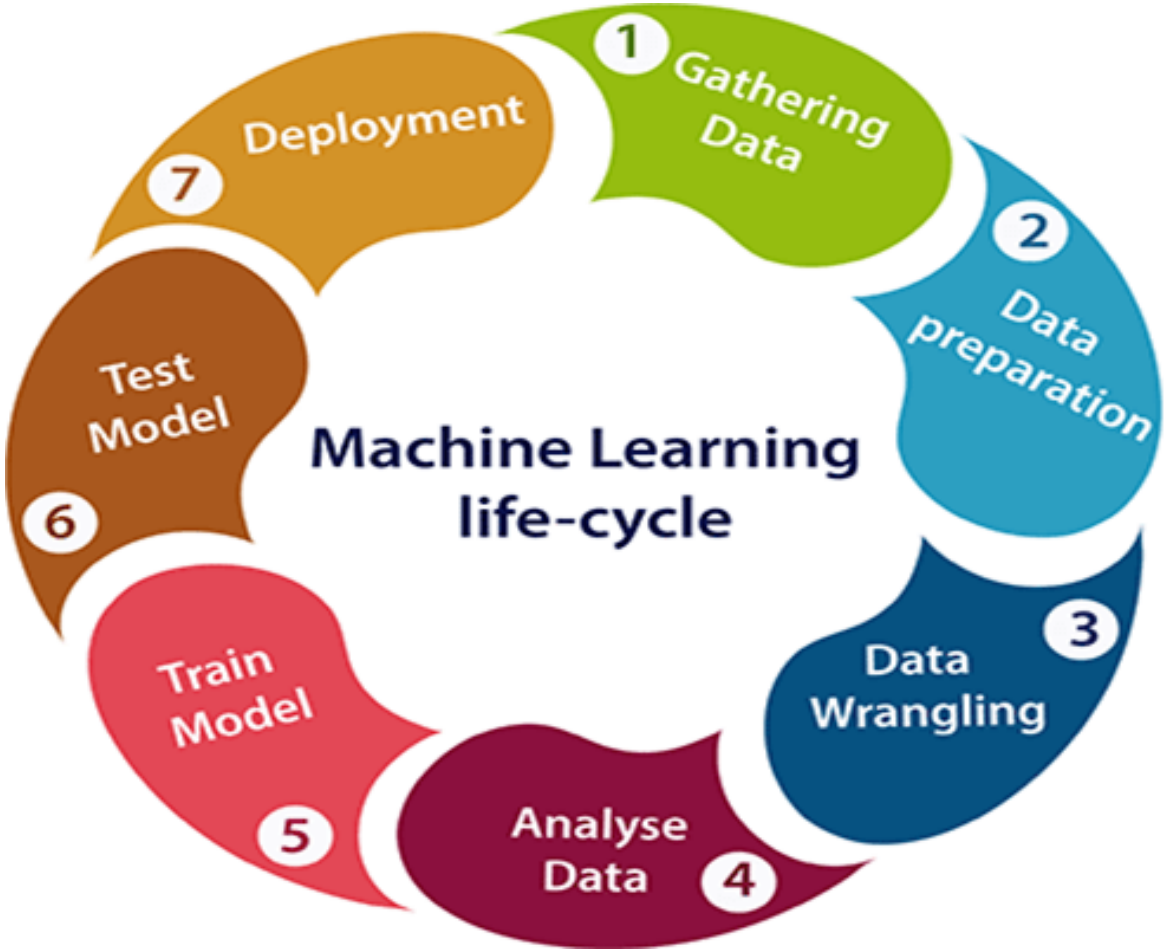
Data collection

The goal of this step is to identify and obtain **all data-related problems**.
we need to identify the different data sources, as data can be collected from **various sources** such as files, database, internet, or mobile devices.

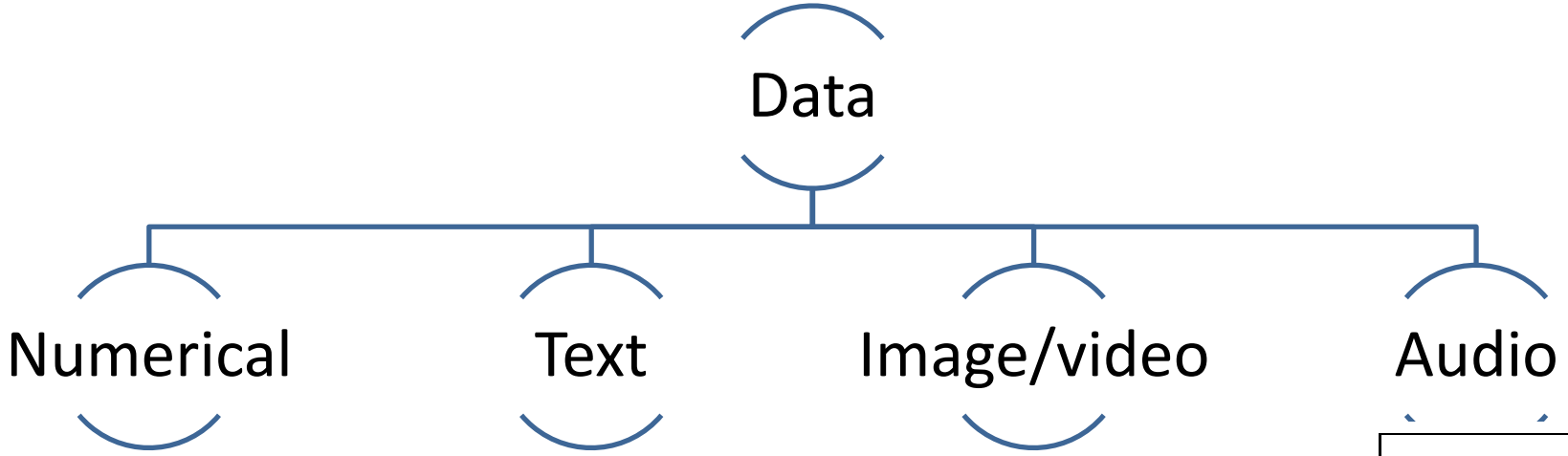
It is one of **the most important steps** of the life cycle.

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.

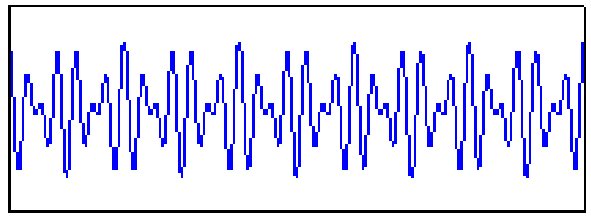
تنبؤ



Types of data in datasets



Country	Age	Salary	Purchased
India	38	48000	No
France	43	45000	Yes
Germany	30	54000	No
France	48	65000	No
Germany	40		Yes
India	35	58000	Yes



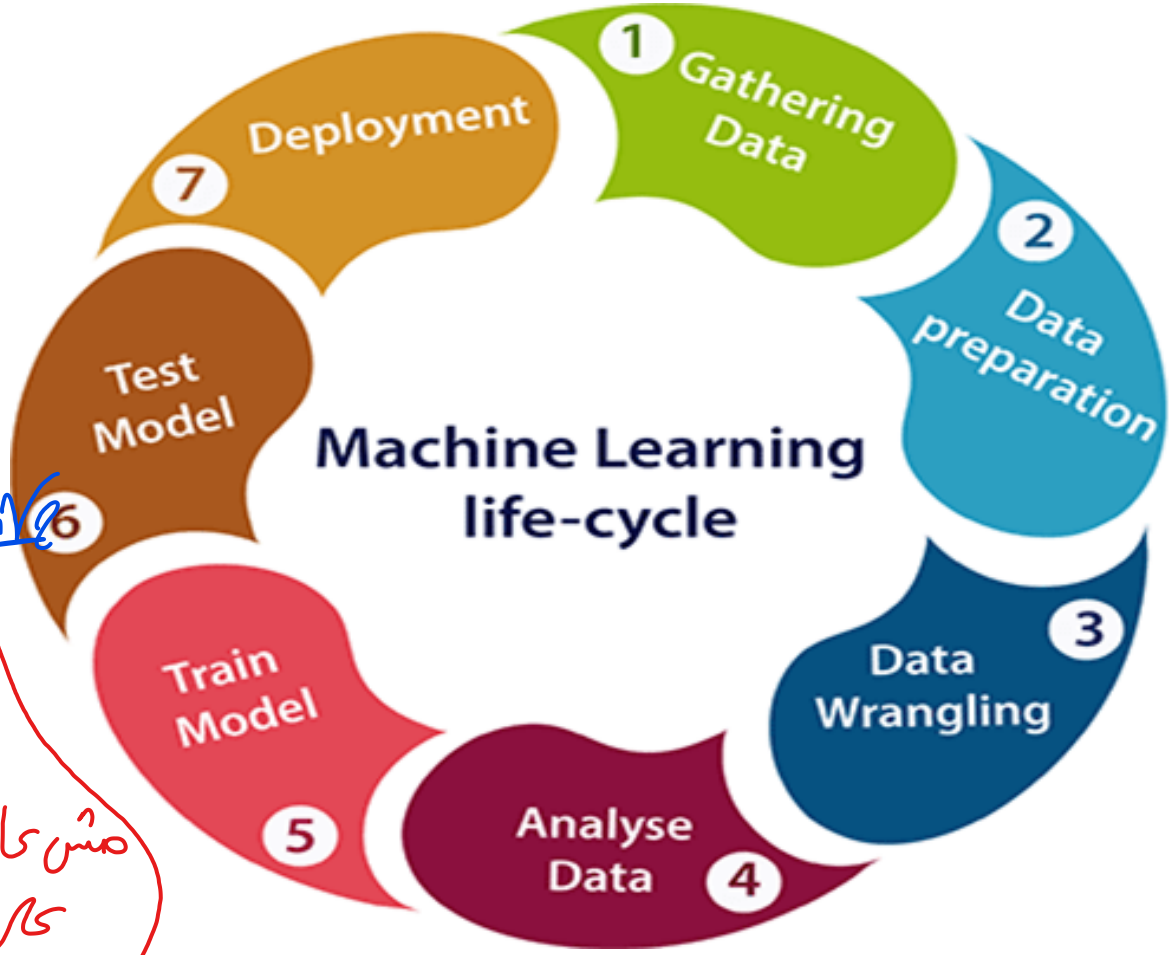
https://www.javatpoint.com/how-to-get-datasets-for-machine-learning

Popular sources for ML datasets

Kaggle datasets: https://www.kaggle.com/datasets.
UCI machine learning repository: https://archive.ics.uci.edu/ml/index.php.
AWS resources: https://registry.opendata.aws/.
Google dataset search engine: https://toolbox.google.com/datasetsearch.

2- Data preparation

It 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**.
Then, **understand the nature of data** that we have to work with. We need to understand the characteristics, format, and quality of data.

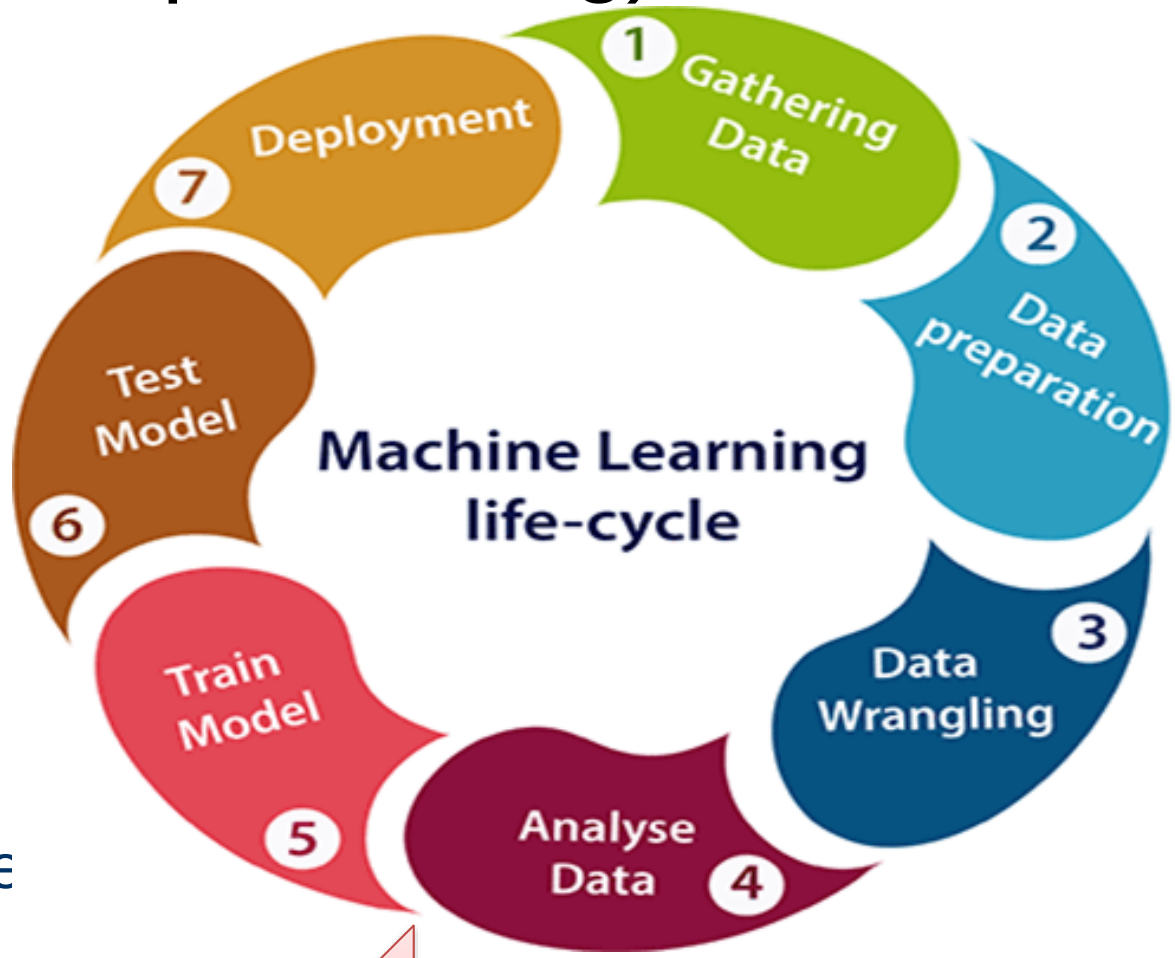


3- Data Wrangling (Cleansing / Preprocessing)

is the process of cleaning and converting raw data into a useable format.
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.



Our Course starting

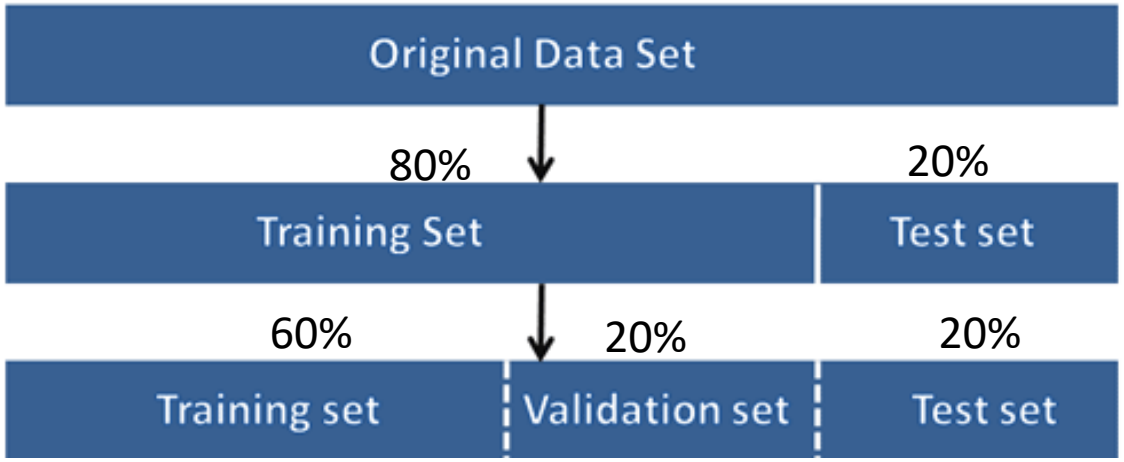
Split Dataset

During the development of the ML project, the developers completely rely on the datasets.

In building ML applications, datasets are divided into two parts:

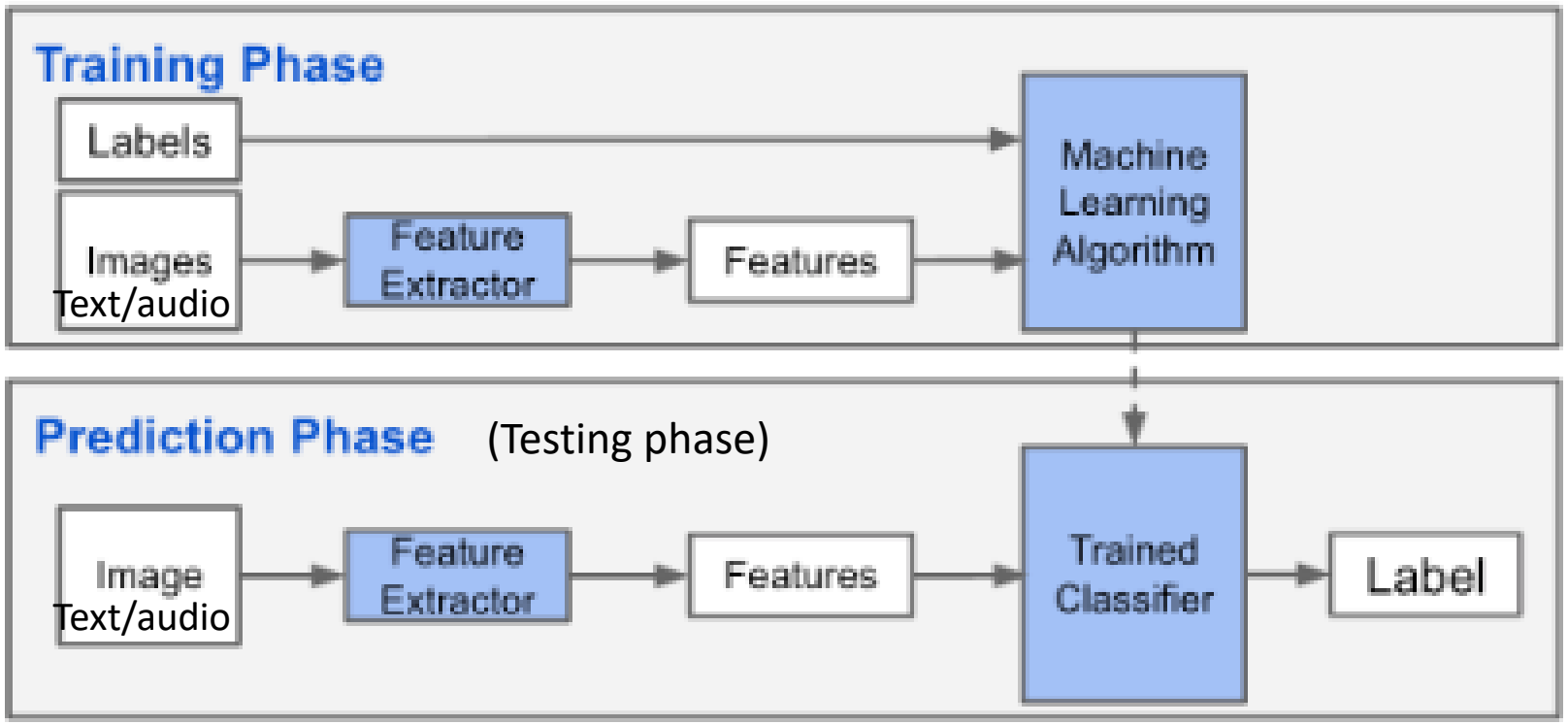
Training dataset:

Test Dataset



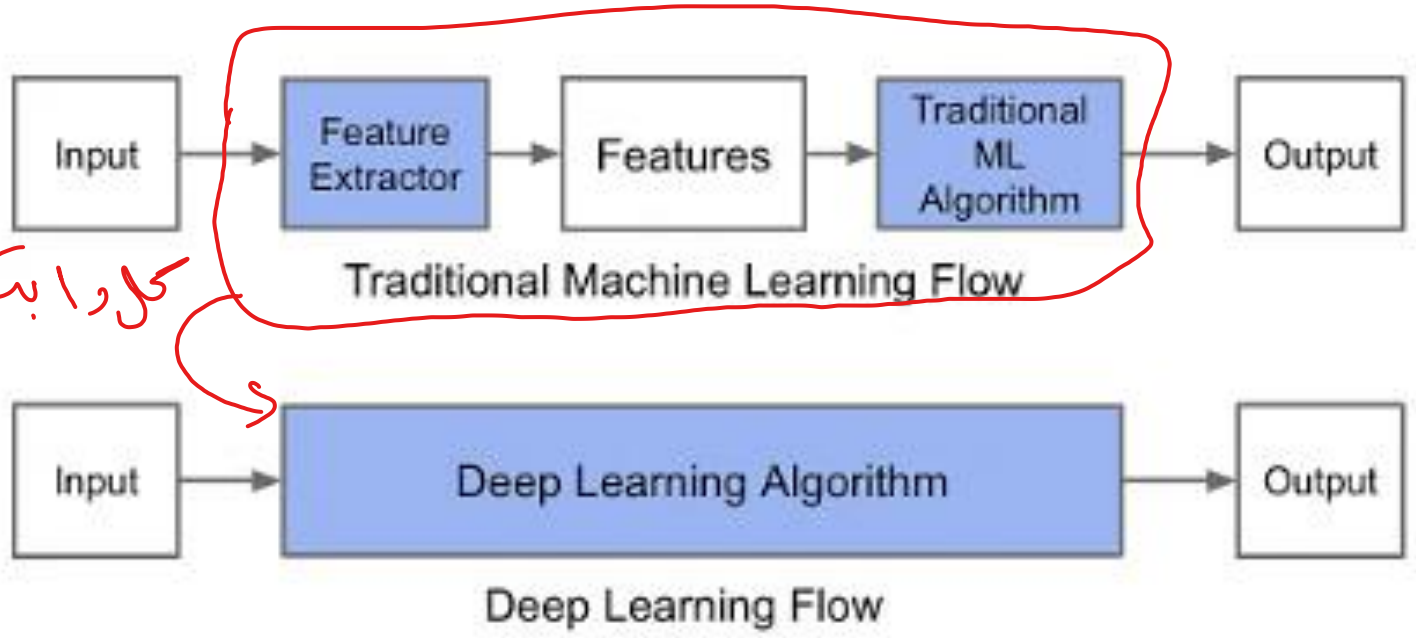
Feature Extraction (Selection)

In **traditional machine learning** algorithms, hand-craft the features, i.e. select the features, is needed.



Feature Extraction (Selection) Cont.

In **traditional machine learning** algorithms, hand-craft the features is needed.
By contrast, in **deep learning algorithms** feature engineering is done automatically by the algorithm.



Deep learning versus Traditional ML algorithms

4- Data Analysis

Select techniques , Build, and evaluate models

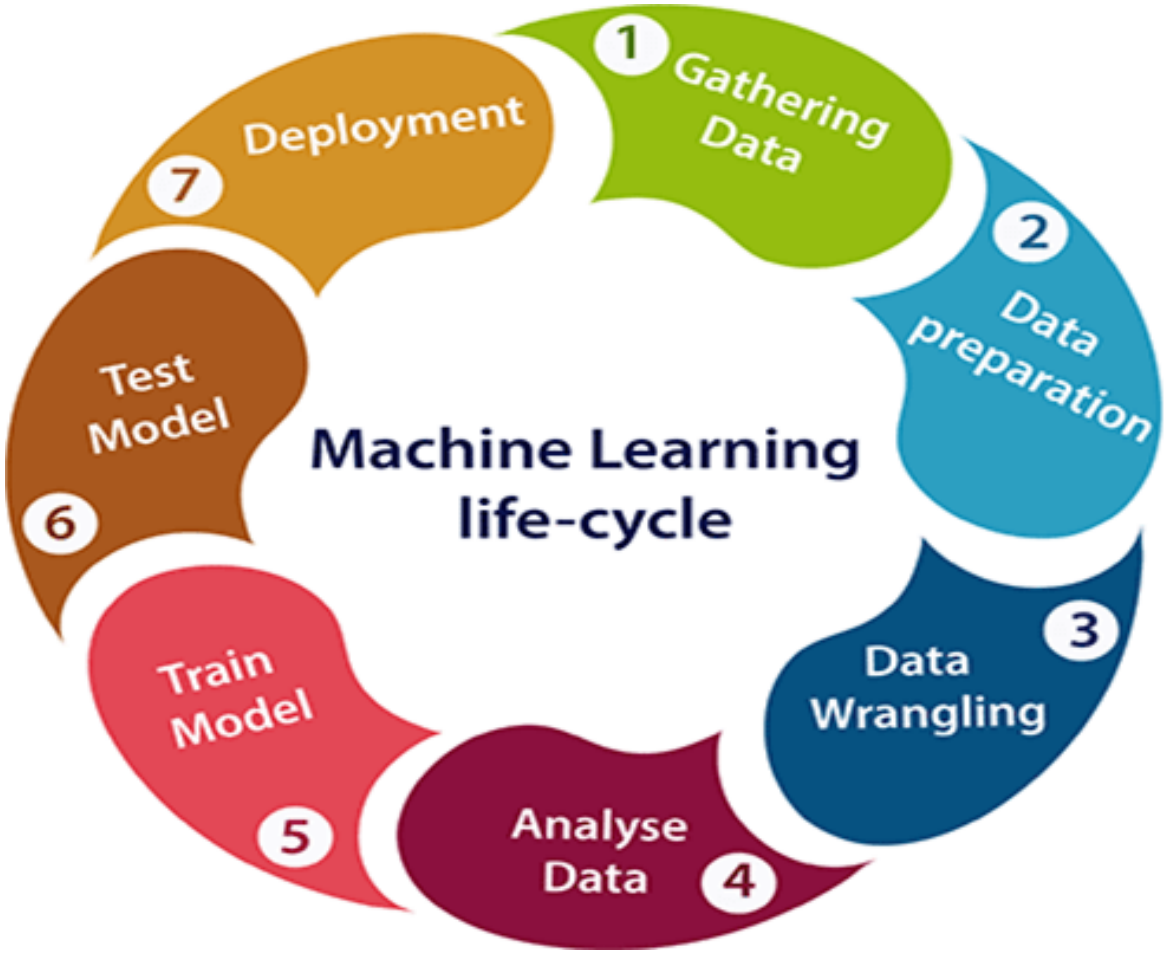
This step involves:

a) Selection of analytical techniques

Select ML techniques such as Classification, Regression, Cluster analysis, Association, etc.

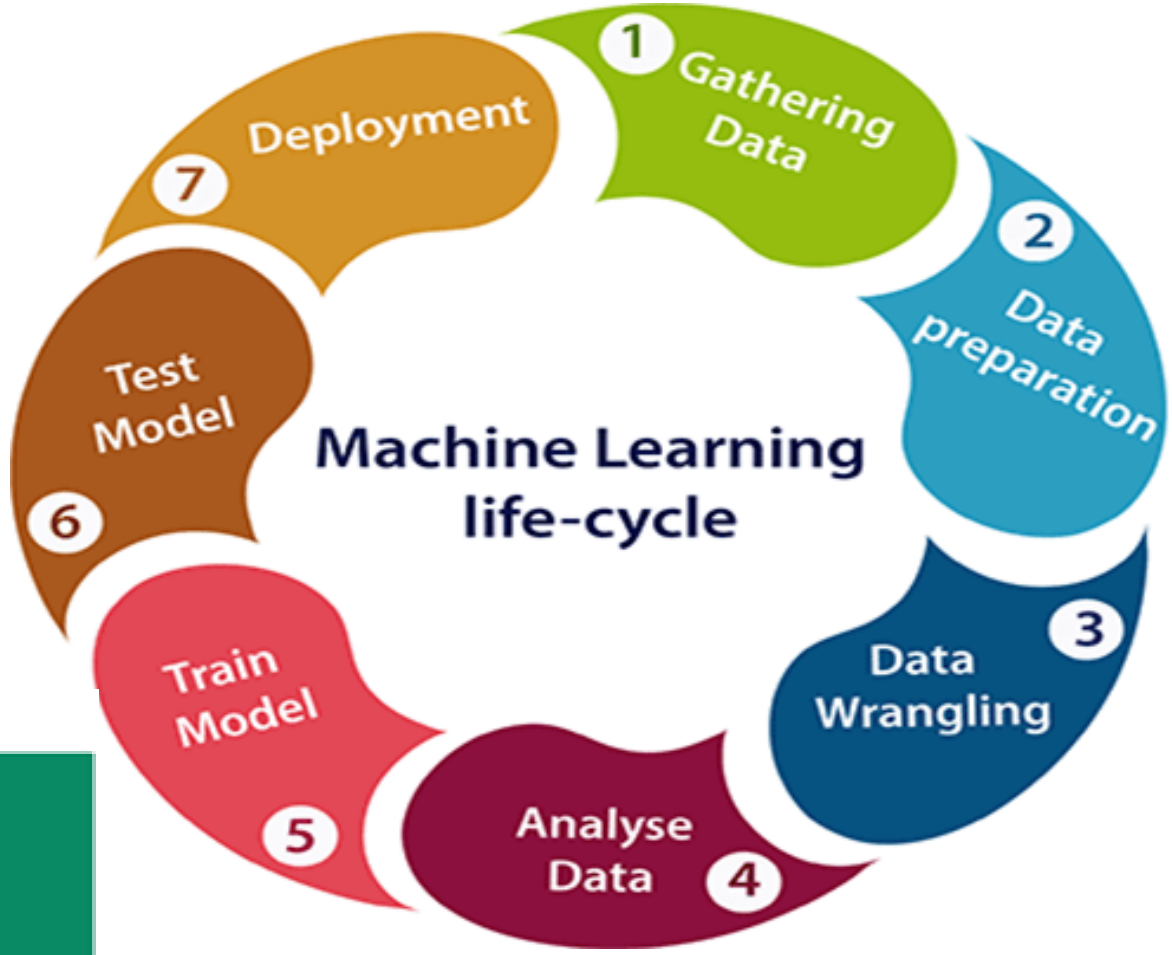
b) Building & evaluate models

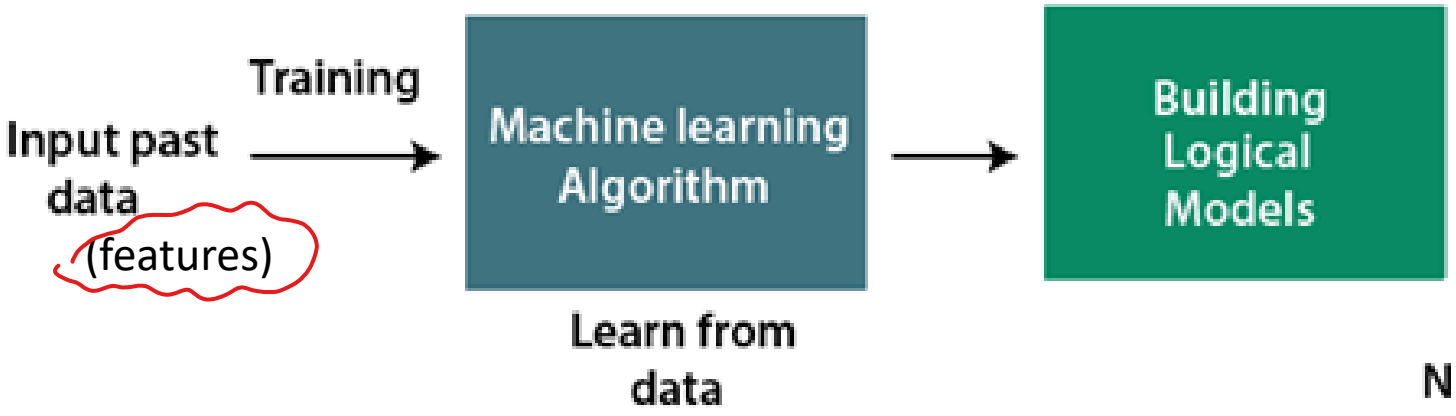
build the selected model using prepared data, and evaluate the model.



5- Train Model

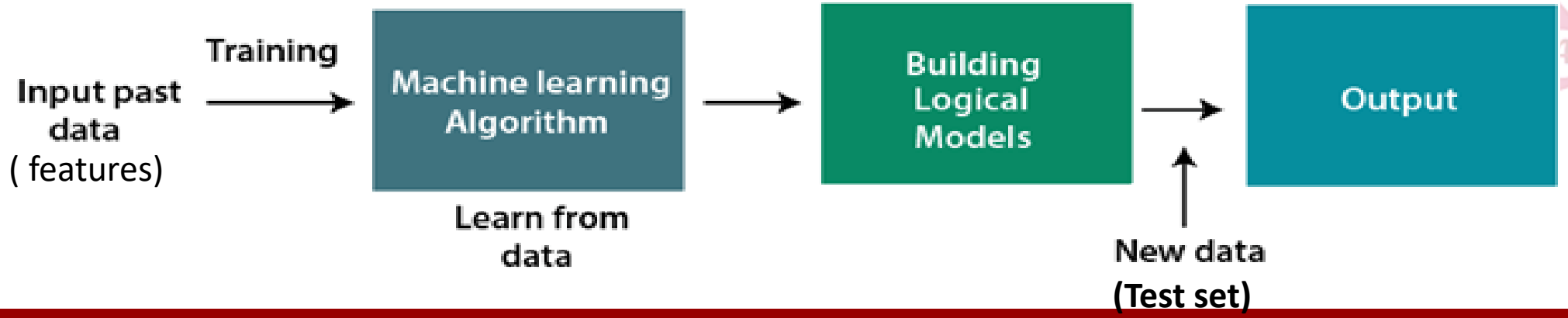
We use datasets to train the model using various machine learning algorithms. Training a model is required so that it can understand the various **patterns, and rules.**





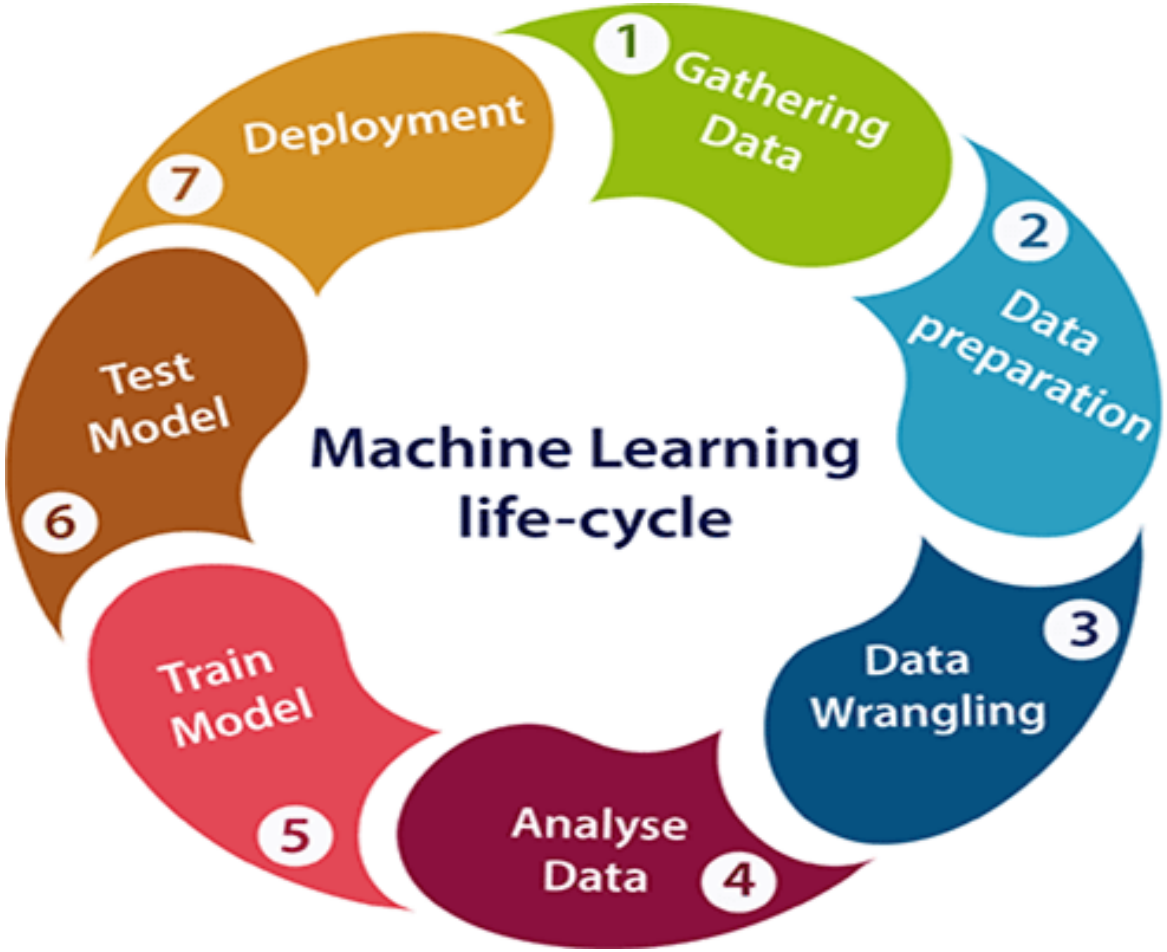
6- Test Model

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

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**.



Types of Machine Learning

Types of Machine Learning

labeled data

Supervised

- Classification
- Regression
- Ranking

Training data includes desired outputs

make Prediction

this Technique is used in the Pre-Processing data stage

unlabeled data

don't make Prediction

Unsupervised

- Clustering
- Dim. Reduction
- Association

dimensional Reduction → reduce the number of variables in data

the algorithms looks for → relationships between variables in the data

Training data does not include desired outputs

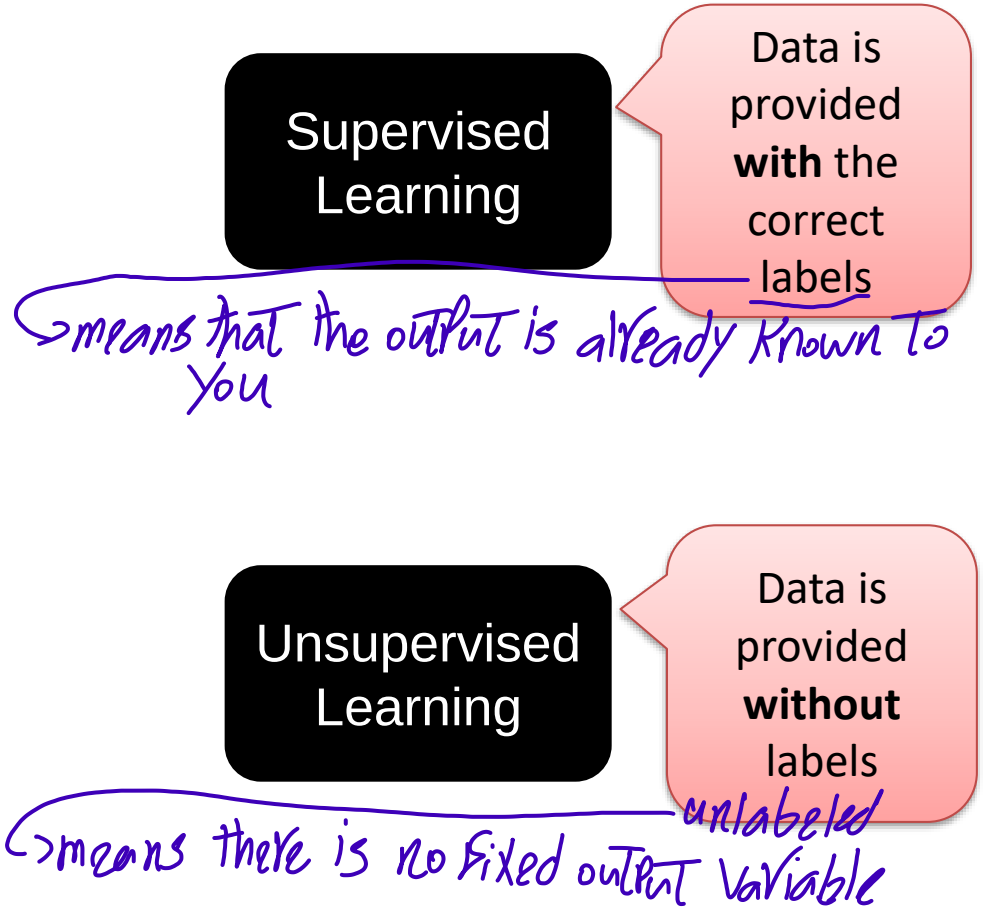
the algorithm knows what the correct is but unsupervised doesn't

Reinforcement

Learn from mistakes & rewards

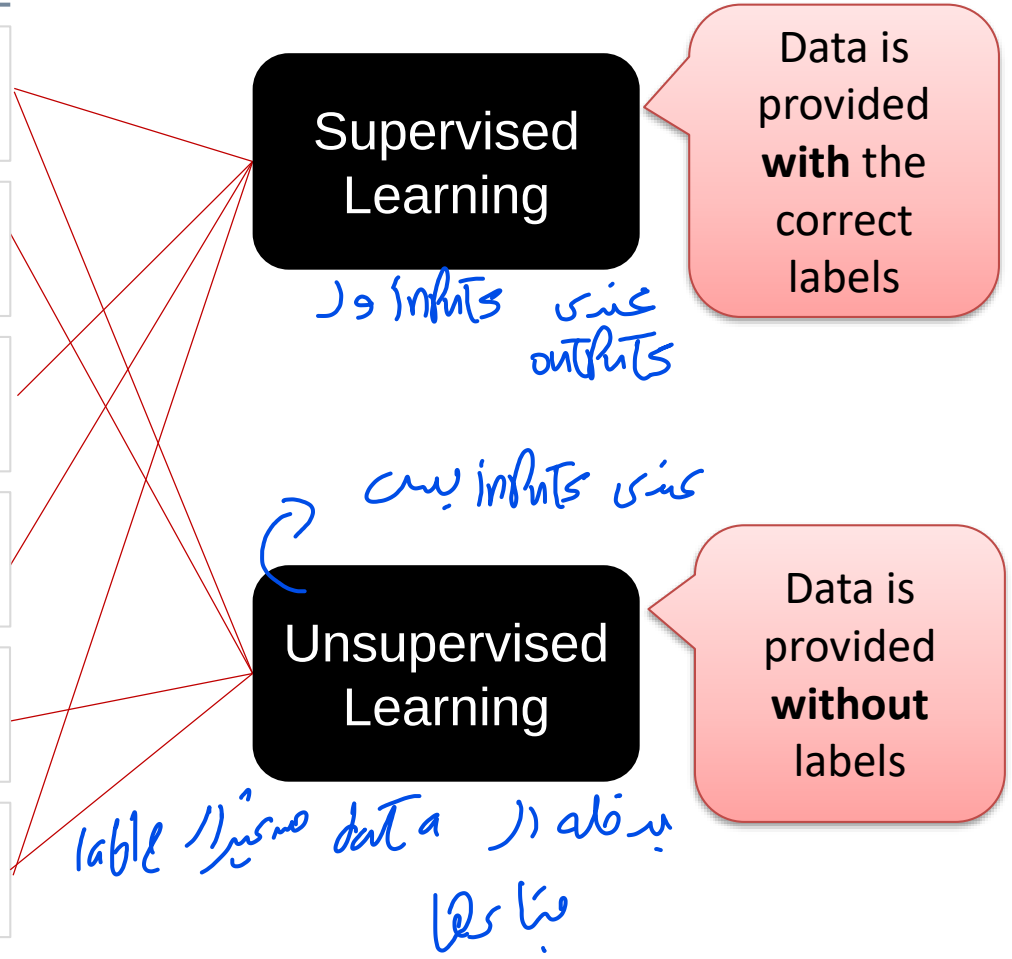
Supervised vs. Unsupervised Learning

Problem type	Description
Ranking	Helping users find the most relevant *items*
Recommendation	Giving users the *items* they may be most interested in
Classification	Figuring out what category does an *item* belongs to
Regression	Predicting a numerical value of an *item*
Clustering	Putting similar *items* together
Anomaly Detection	Finding uncommon *items*

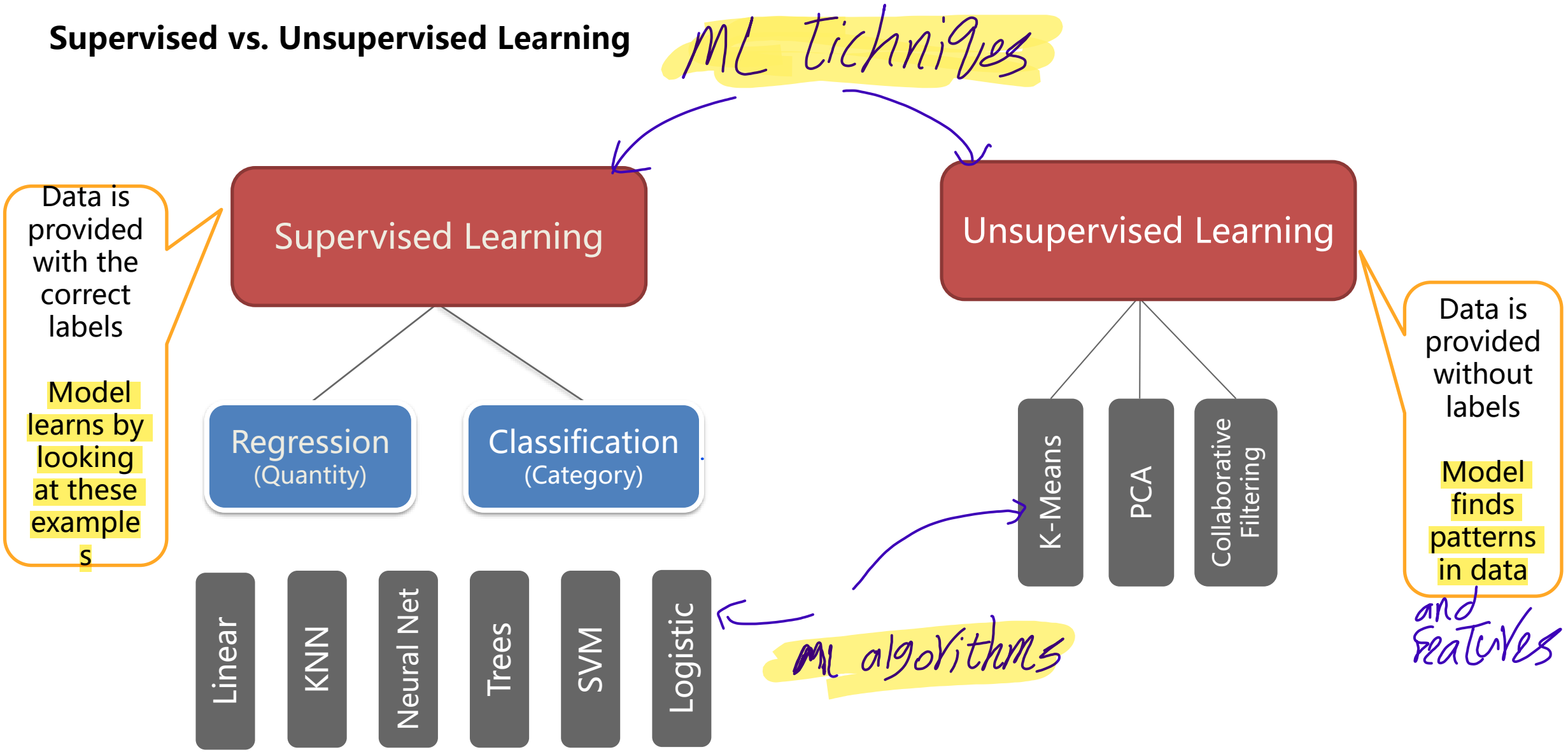


Supervised vs. Unsupervised Learning

Problem type	Description
Ranking	Helping users find the most relevant *items*
Recommendation	Giving users the *items* they may be most interested in
Classification	Figuring out what category does an *item* belongs to
Regression	Predicting a numerical value of an *item*
Clustering	Putting similar *items* together
Anomaly Detection	Finding uncommon *items*



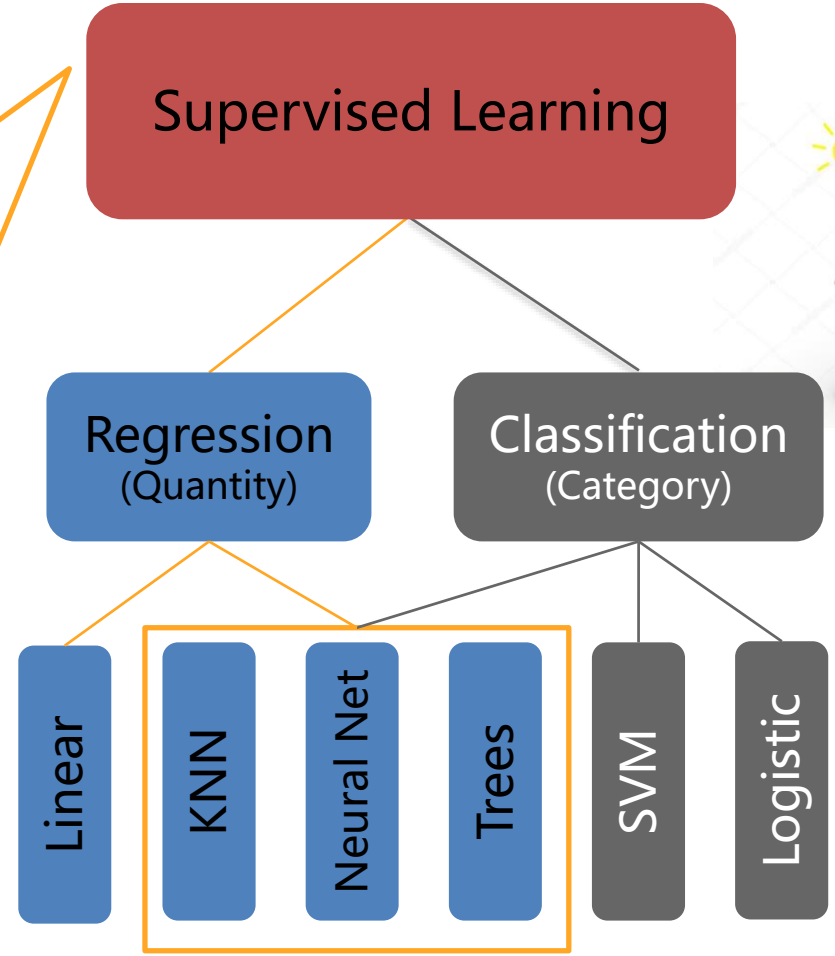
Supervised vs. Unsupervised Learning

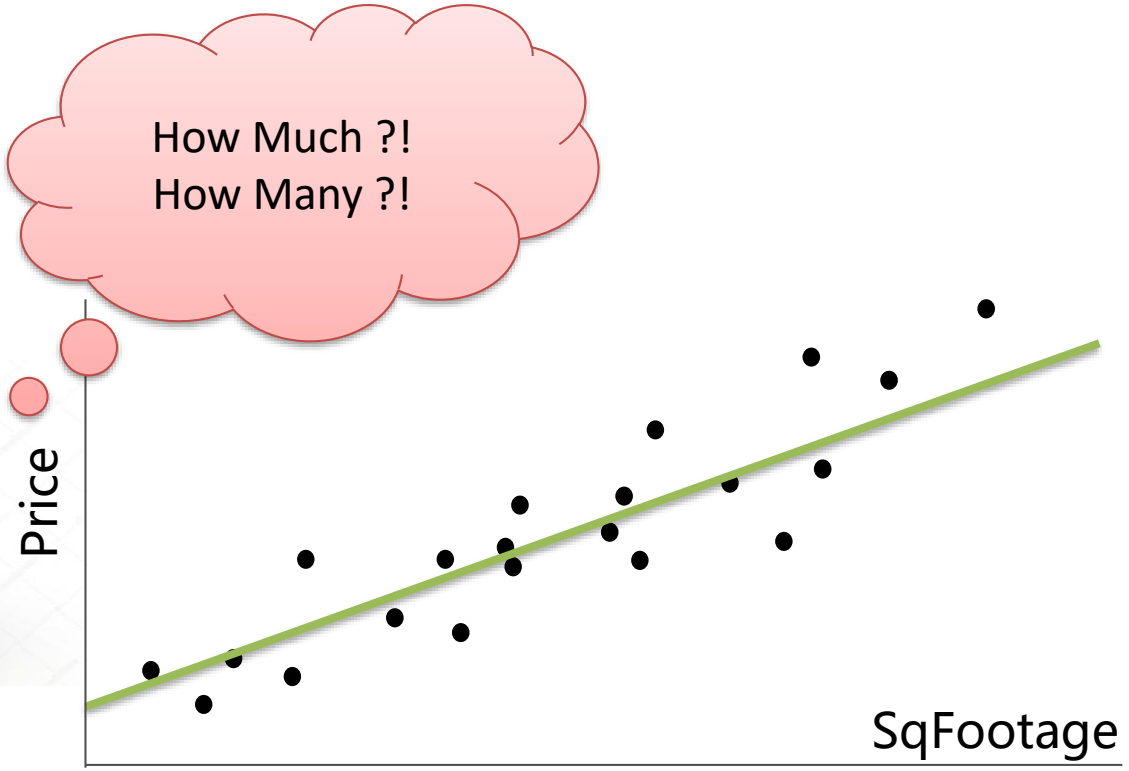


Supervised Learning: Regression

Data is provided with the correct labels

Model learns by looking at these examples





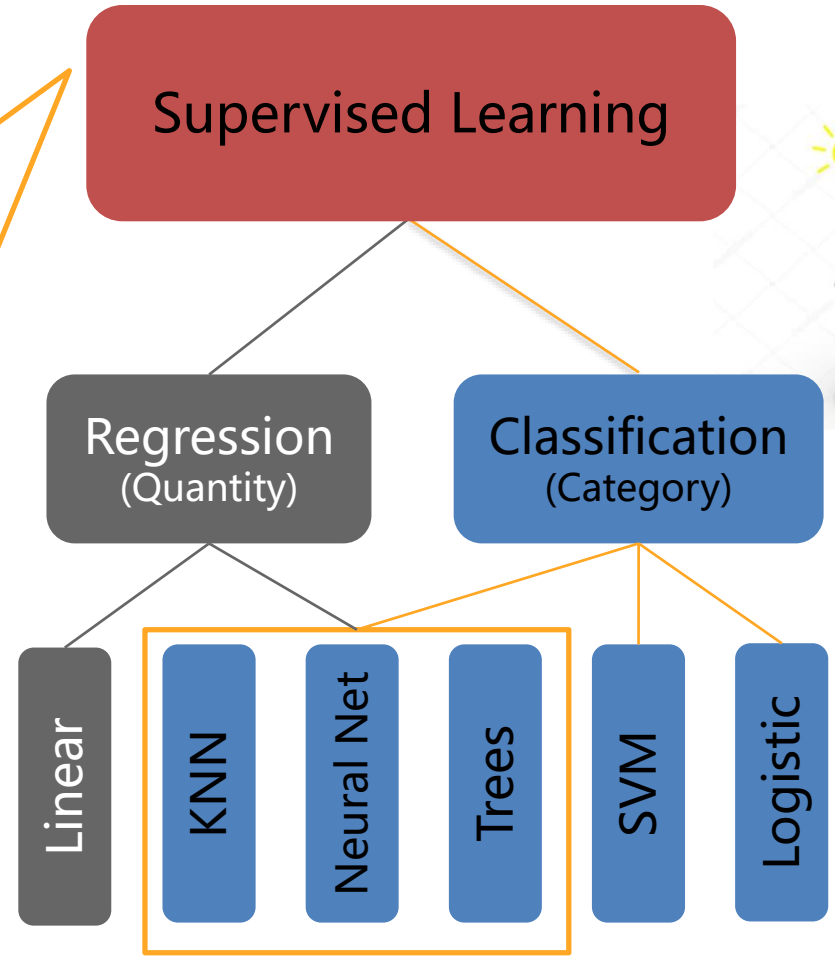
Label | Features

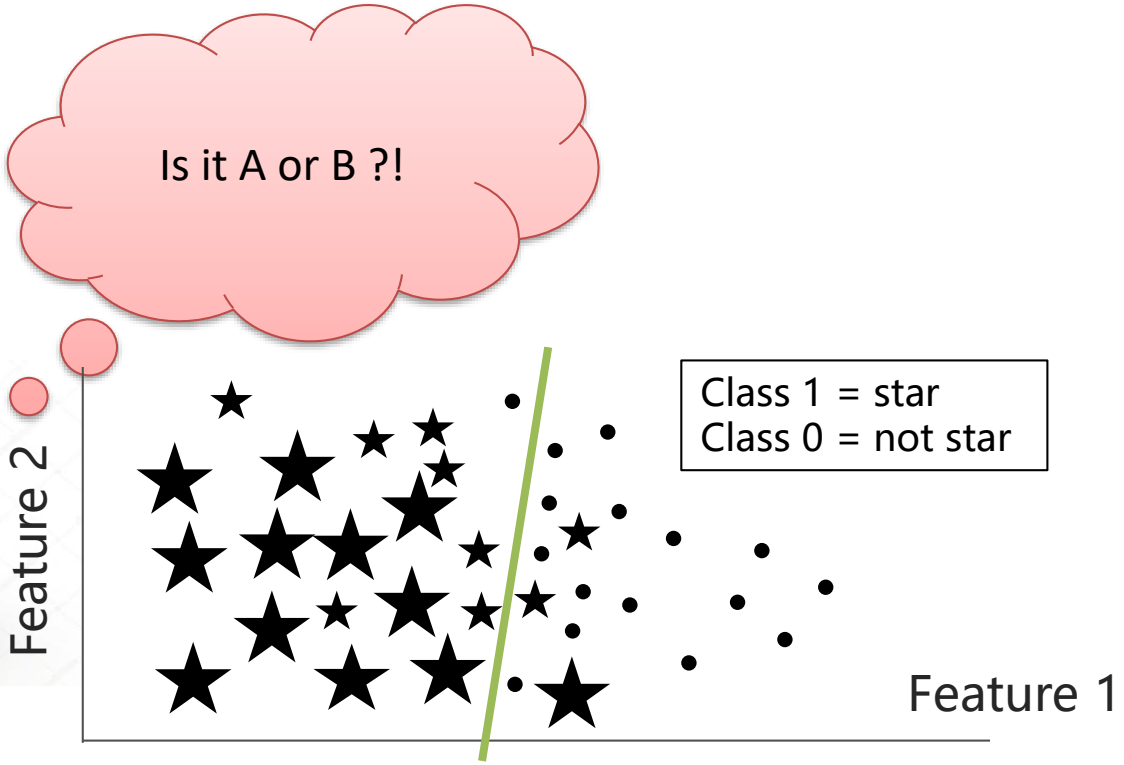
Price	Bedrooms	SqFootage	Age
280.000	3	3292	14
210.030	2	2465	6
…	…	…	…

Supervised Learning: Classification

Data is provided with the correct labels

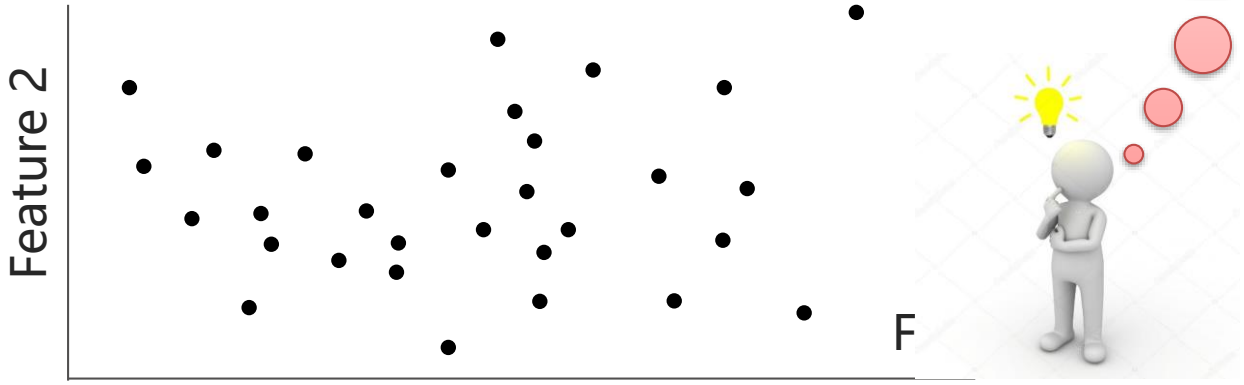
Model learns by looking at these examples





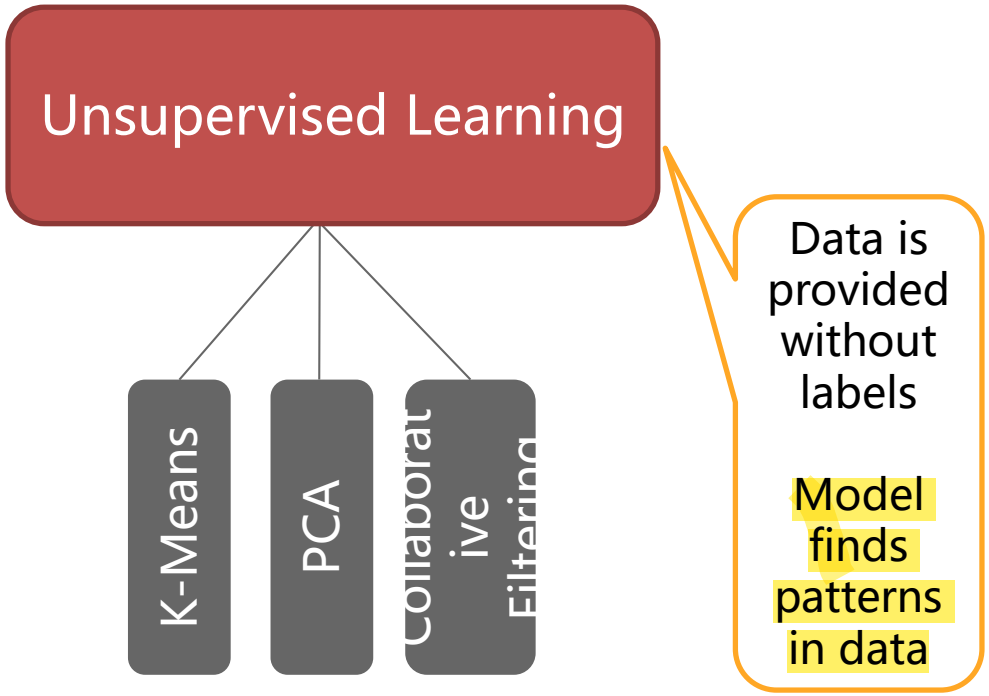
Label	Features		
Star	Points	Edges	Size
1	5	10<	750
0	0	>9	150
…	…	…	…

Unsupervised Learning: Clustering

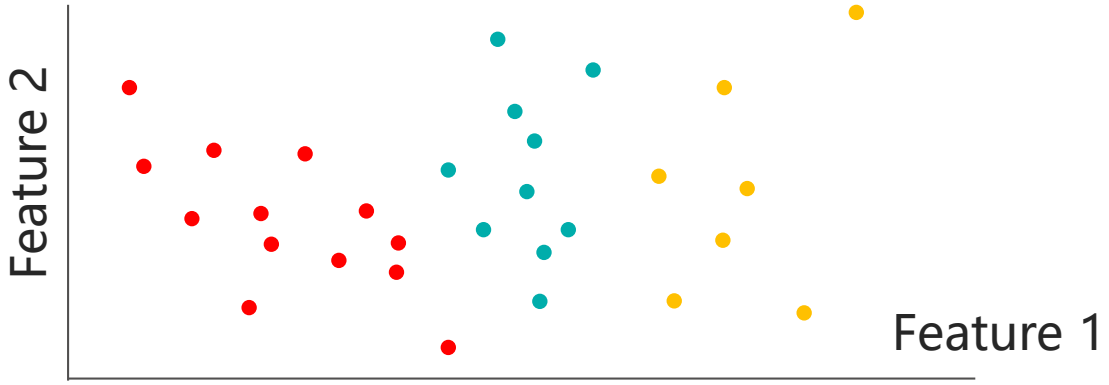


Features

Age	Music	Books
21	Classical	Practical Magic
47	Jazz	The Great Gatsby
…	…	…

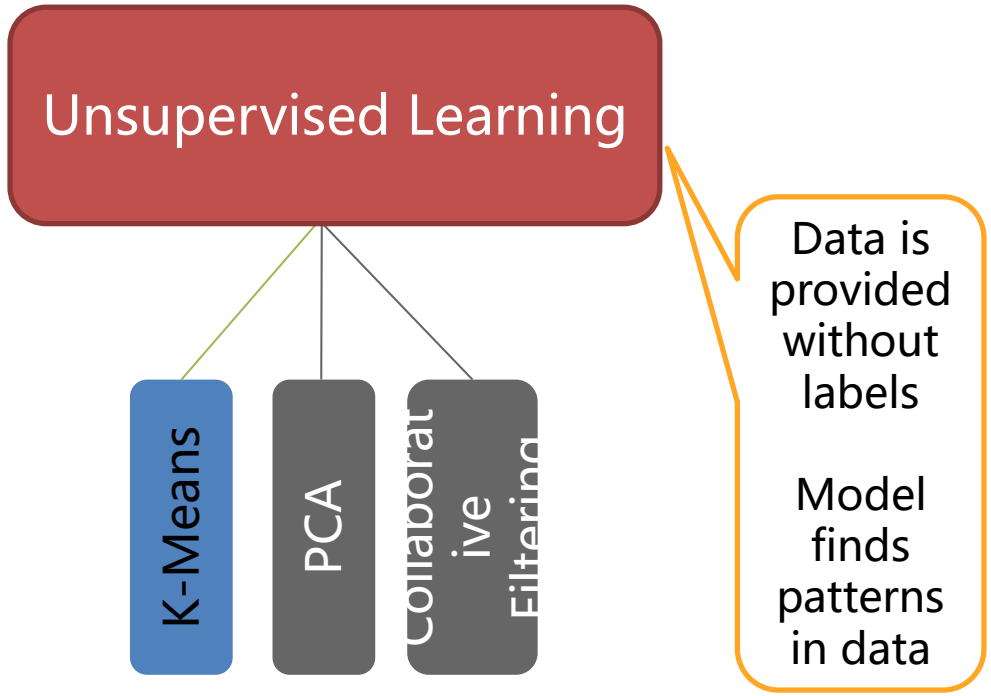


Unsupervised Learning: Clustering



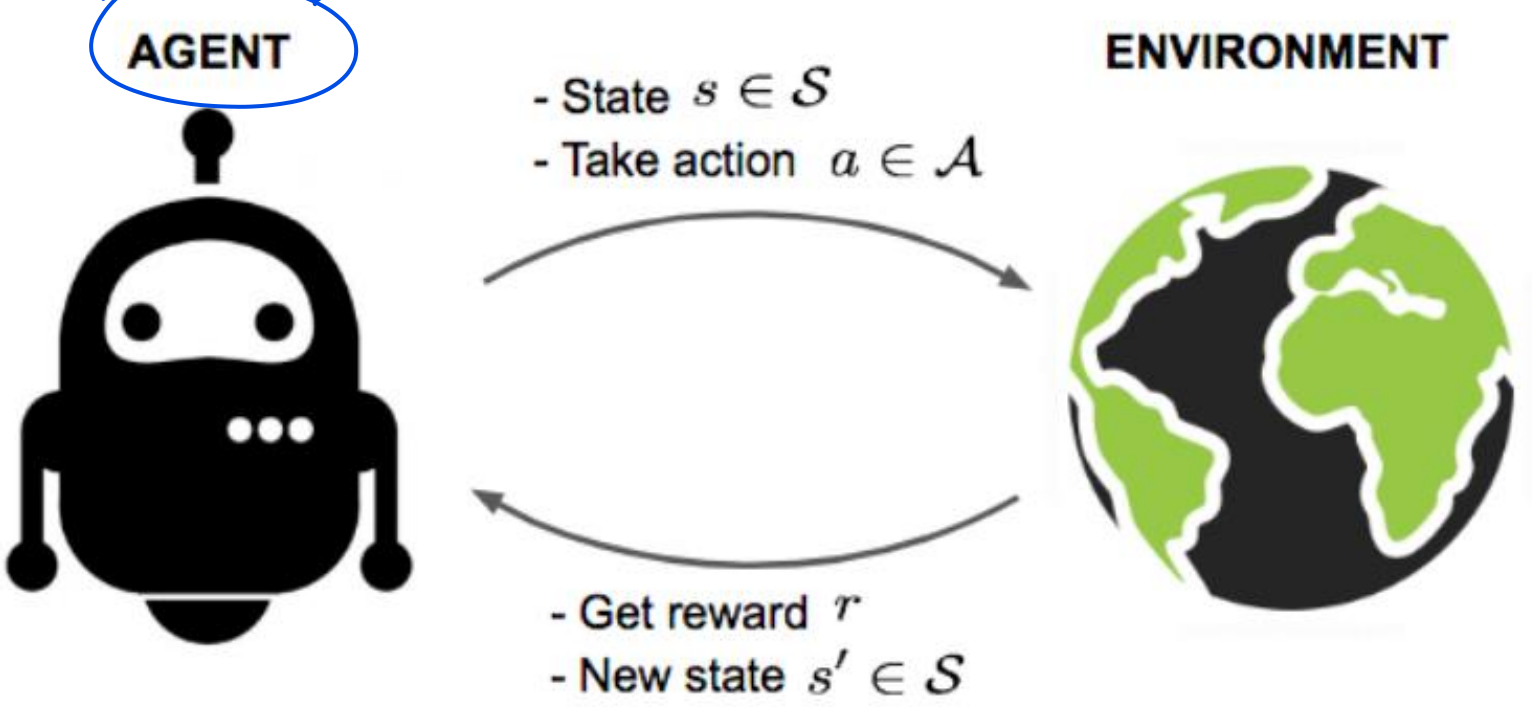
Features

Age	Music	Books
21	Classical	Practical Magic
47	Jazz	The Great Gatsby
…	…	…



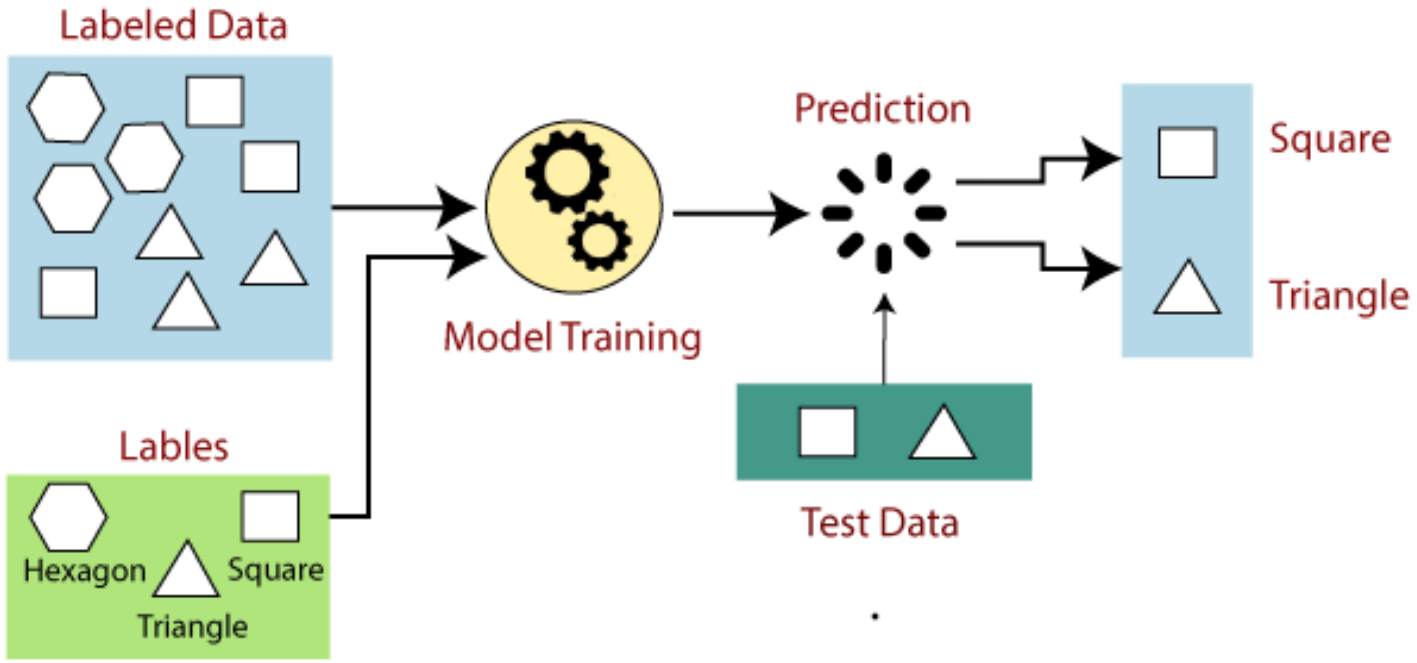
What should I do now ?!

Reinforcement Learning



Supervised Learning

In supervised learning, models are trained using **labeled dataset**, where the model learns about each type of data. Once the training process is completed, the model is tested on the basis of test data (a subset of the training set), and then it predicts the output.



Supervised Learning (Cont.)

Regression

It is used for the prediction of **continuous variables**

Popular regression algorithms:

- **Linear Regression**
- Regression Trees
- Non-Linear Regression
- Bayesian Linear Regression
- Polynomial Regression

Classification

It is used when the output variable is **categorical** (classes).

Popular classification algorithms:

- Random Forest
- Decision Trees
- **Logistic Regression**
- **Support vector Machines (SVMs)**
- **Neural Networks (NNs)**

Advantages and Disadvantages of Supervised learning

Advantages of Supervised Learning

With the help of supervised learning, the model can predict the output on the basis of prior experiences.

Supervised learning model helps us to solve various real-world problems such as **fraud detection, spam filtering**, etc.

Disadvantages of Supervised Learning

Supervised learning cannot predict the correct output if the test data is different from the training dataset.

Training required **lots of computation** times.

In supervised learning, we **need enough knowledge** about the classes of object.

Unsupervised Learning

- Here, we have taken **unlabeled input data**, which means it is not categorized and corresponding outputs are also not given.

Clustering: Clustering is a method **of grouping objects into clusters** such that objects with the most similarities remain in a group and have fewer or no similarities with the objects of another group.

Association: An association rule is an unsupervised learning method that is used for **finding the relationships between variables** in a large database. It determines the set of items that occurs together in the dataset. The Association rule makes marketing strategy more effective.

Example: People who buy X items (bread) also tend to purchase Y (Butter/Jam) items.

Advantages and Disadvantages of Unsupervised Learning

Advantages of Unsupervised Learning

Unsupervised learning is used for **more complex tasks** as compared to supervised learning because, in unsupervised learning, we don't have labeled input data.

Unsupervised learning is preferable as it is **easy to get unlabeled data** in comparison to labeled data.

Disadvantages of Unsupervised Learning

Unsupervised learning is intrinsically **more difficult** than supervised learning as it does not have corresponding output.

The result of the unsupervised learning algorithm might be **less accurate** as input data is not labeled, and algorithms do not know the exact output in advance.

Thanks