CS395: Selected CS1 (Introduction to Machine Learning)

Associate Prof. Wessam El-Behaidy

Fall 2021

References:

https://www.coursera.org/learn/machine-learning (Andrew Ng)

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

Agenda

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

Course Objectives

- To introduce students to the basic concepts and techniques of Machine Learning.
- To develop skills of using recent machine learning software for solving practical problems.

References

- **Lectures (Course slides) are based on :**
  - https://www.coursera.org/learn/machine-learning (Andrew Ng)
- **Practical Labs based on:**
  - Machine learning A to Z: **Kirill Eremenko ©superdatascience**

Topics to Cover

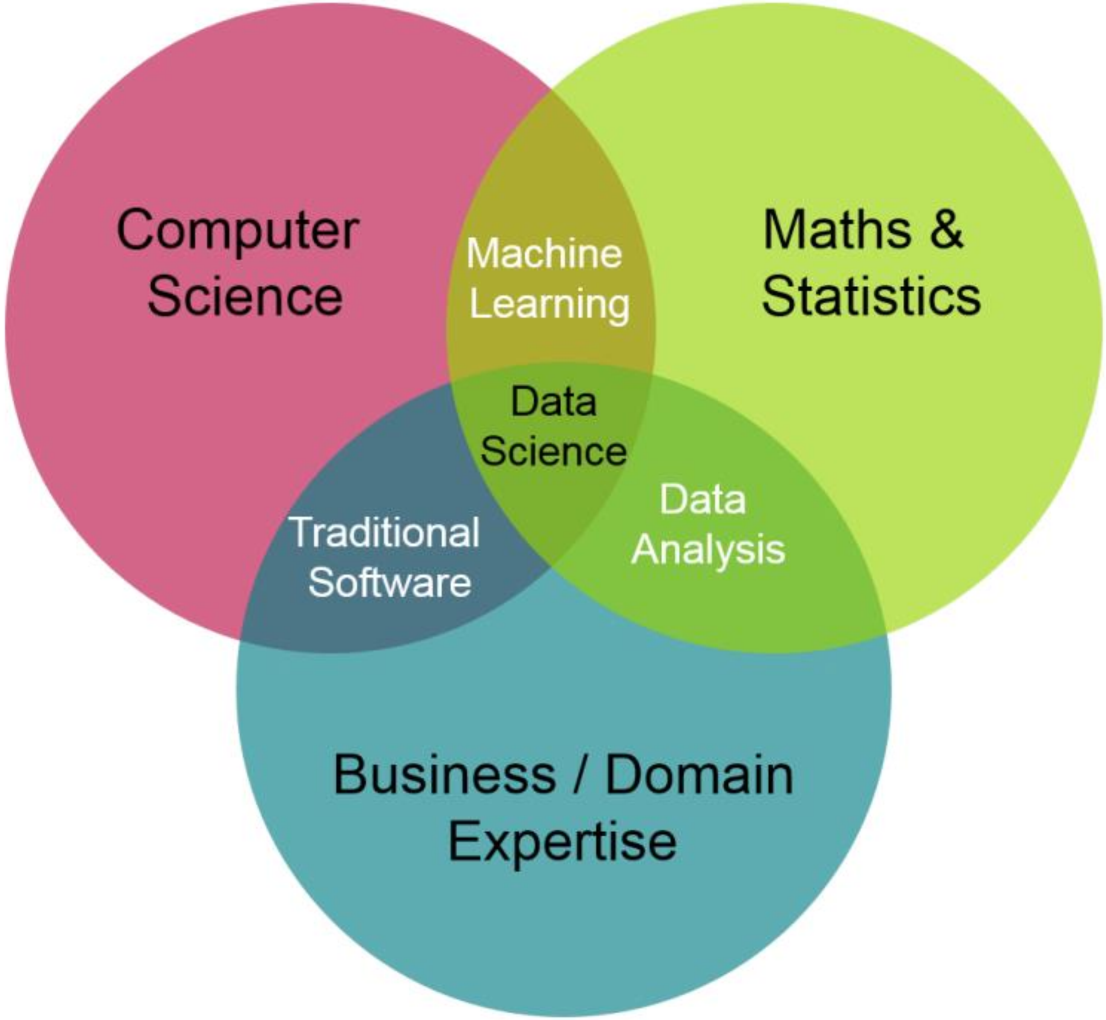
- **Introduction to machine learning**
- **Linear Regression with one variable**
- **Gradient Descent**
- **Linear Algebra – review**
- **Linear Regression with Multiple Variables**
- **Logistic Regression**
- **Regularization**
- **Neural Networks – Representation**
- **Neural Networks – Learning**
- **Support Vector Machines**
- **Evaluation Metrics**

Grading Distribution

- Project: (20%)
  - Code “github”, pitching video, presentation.
- Midterm exam (20%)
- Final exam (60%)

Where are we?

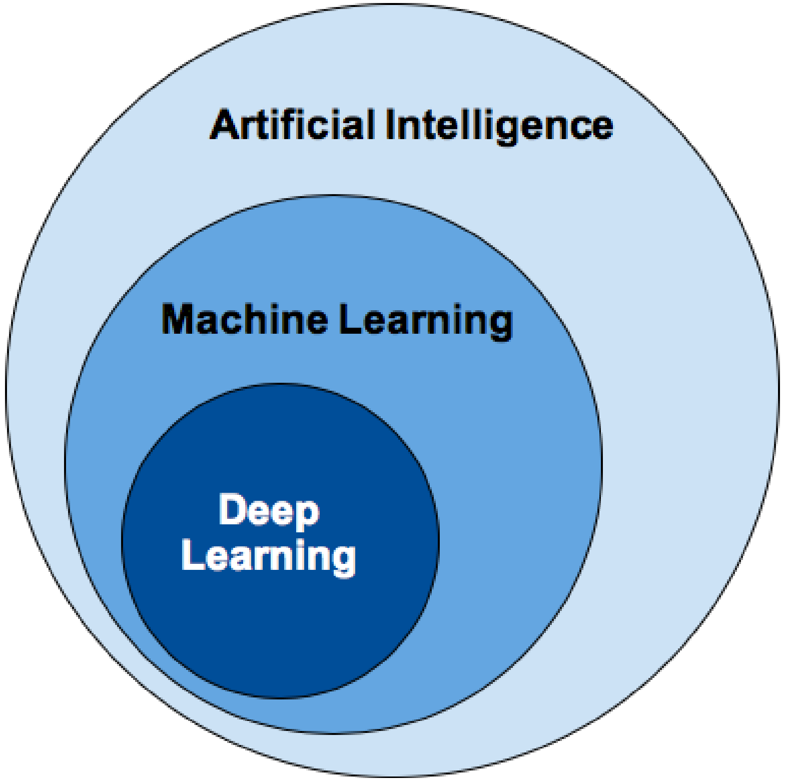
- **Data science** is a broad, interdisciplinary field that aims to:
    - use a scientific approach to extract meaning from data, and
    - transform data into important business information
- Data Science is a combination of many disciplines, including statistics, information science, and mathematics. It is difficult to master all fields and be fairly competent in all of them.
- One of the most exciting technologies in modern data science is **machine learning.**



https://www.jigsawacademy.com/what-is-the-difference-between-data-science-and-machine-learning/

What is Machine Learning?

- The term machine learning was first introduced by **Arthur Samuel** in **1959**.
- Machine learning (ML):
  - Is said as a subset of **artificial intelligence.**
  - Is a growing technology which enables computers to learn automatically from past data.
  - Uses various algorithms for **building mathematical models and making predictions using historical data or information**.

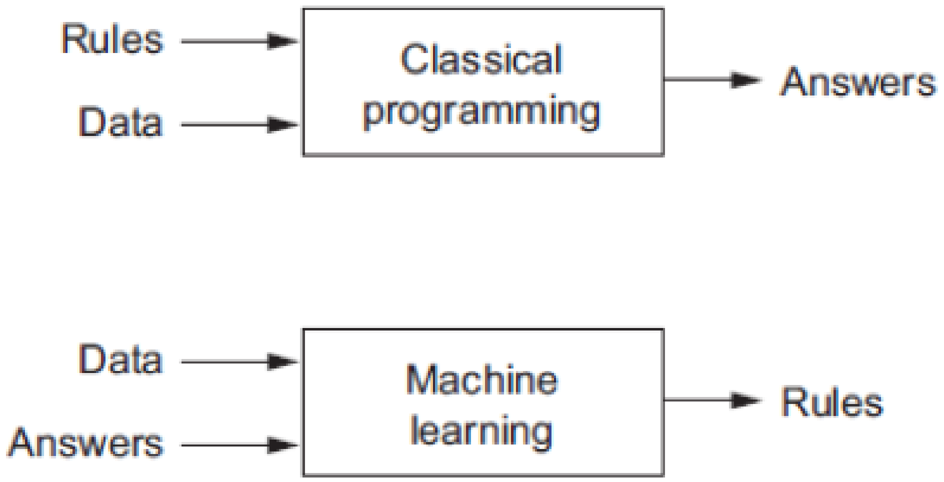


https://www.sumologic.com/wp-content/uploads/compare_AI_ML_DL.png

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

Machine Learning Definition

- **Arthur Samuel (1959)**
  - **"**Field of study that gives computers the ability to learn without being explicitly programmed**"**
    - Samuels wrote a checkers playing program
      - Had the program play 10000 games against itself
      - Work out which board positions were good and bad depending on wins/losses



General model of machine Learning

Machine Learning Definition

- **Tom Michel (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**."
    - The checkers example,
      - E = 10000s games
      - T is playing checkers
      - P if you win or not

Machine Learning Example

- "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**.“
- **Example:** Suppose your **email program** watches which emails you do or do not mark as spam, and based on that learns **how to better filter spam**. What is the task T in this setting?

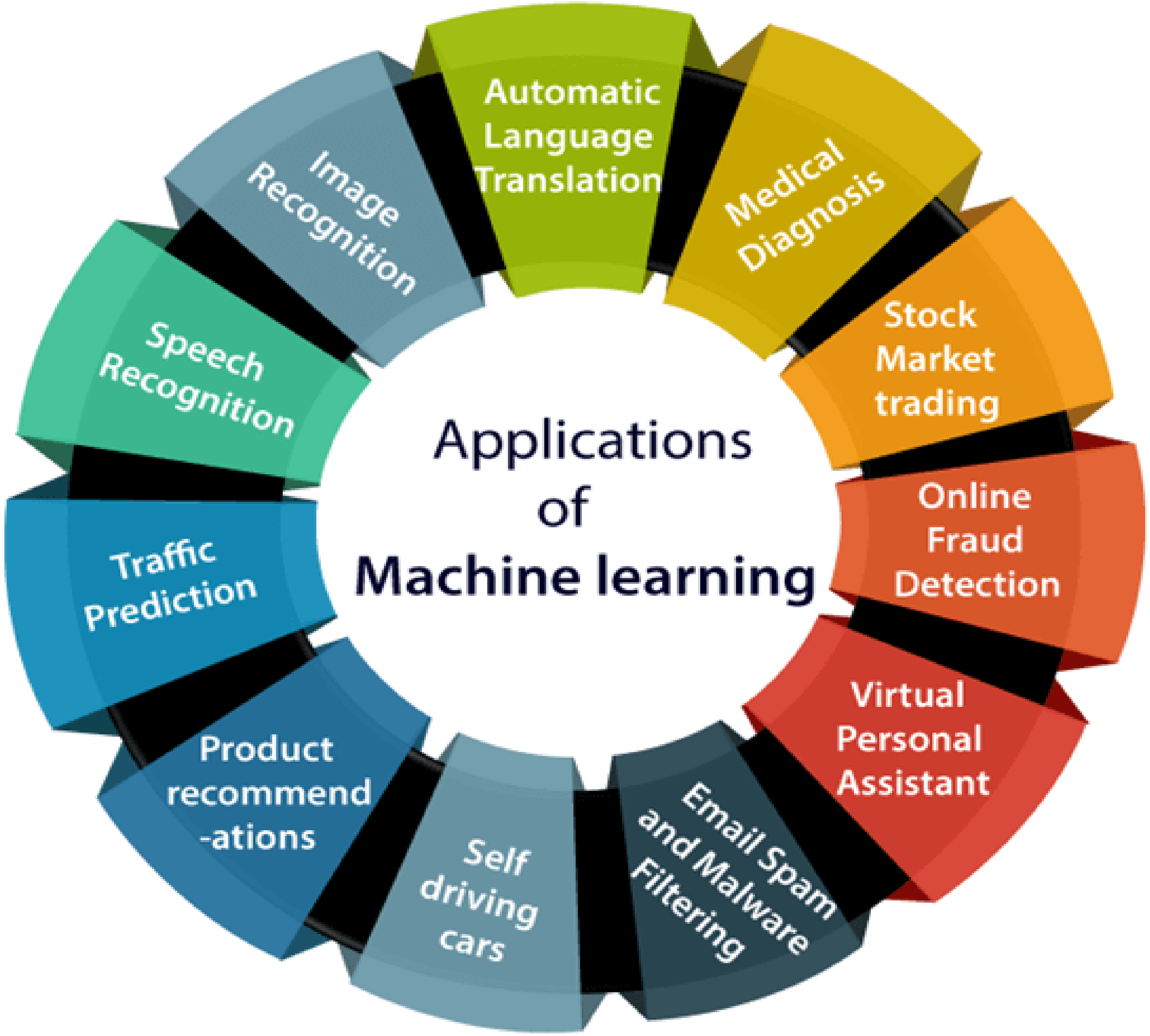
□ Classify emails as spam or not spam. ✓

□ Watching you label emails as spam or not spam. ✗

□ The number (or fraction) of emails correctly classified as spam/not spam. ✗

□ None of the above, this is not a machine learning algorithm. ✗

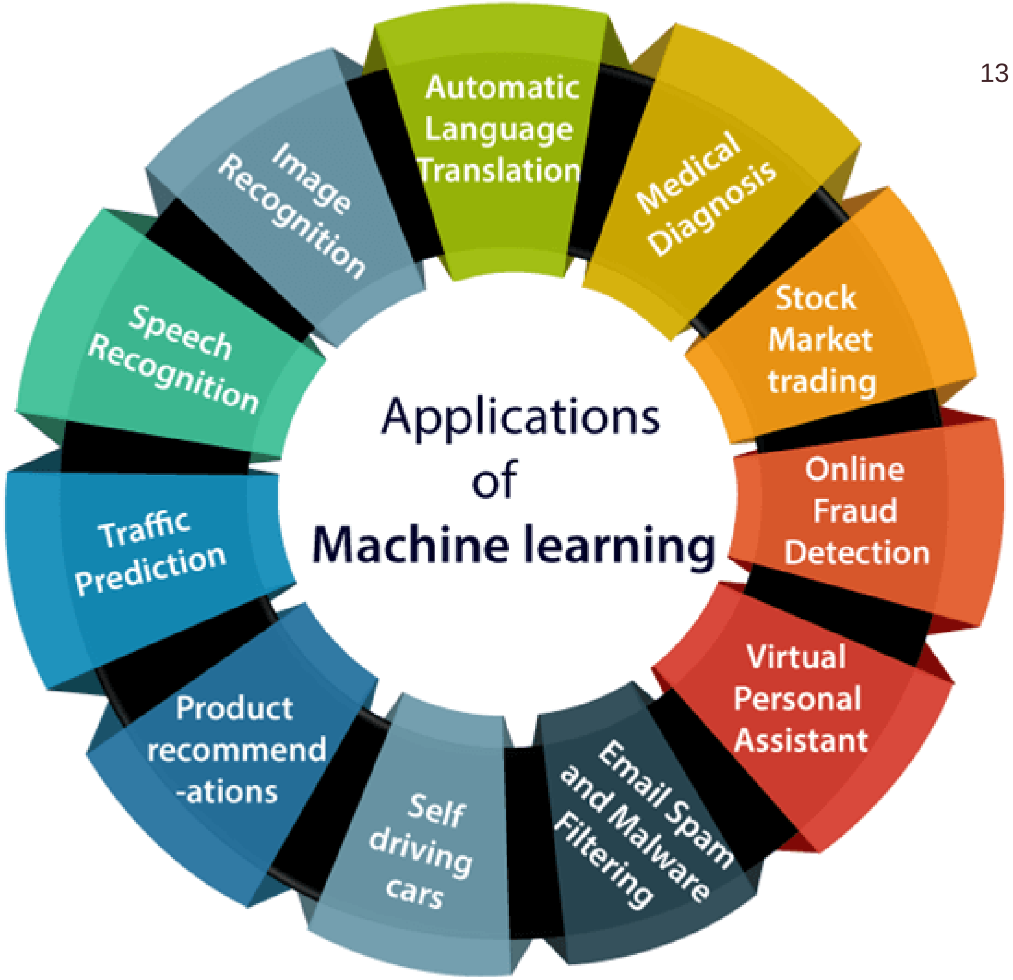
ML Applications


Applications of Machine learning
Automatic Language Translation
Medical Diagnosis
Stock Market trading
Online Fraud Detection
Virtual Personal Assistant
Email Spam and Malware Filtering
Self driving cars
Product recommend -ations
Traffic Prediction
Speech Recognition
Image Recognition

https://www.javatpoint.com/applications-of-machine-learning

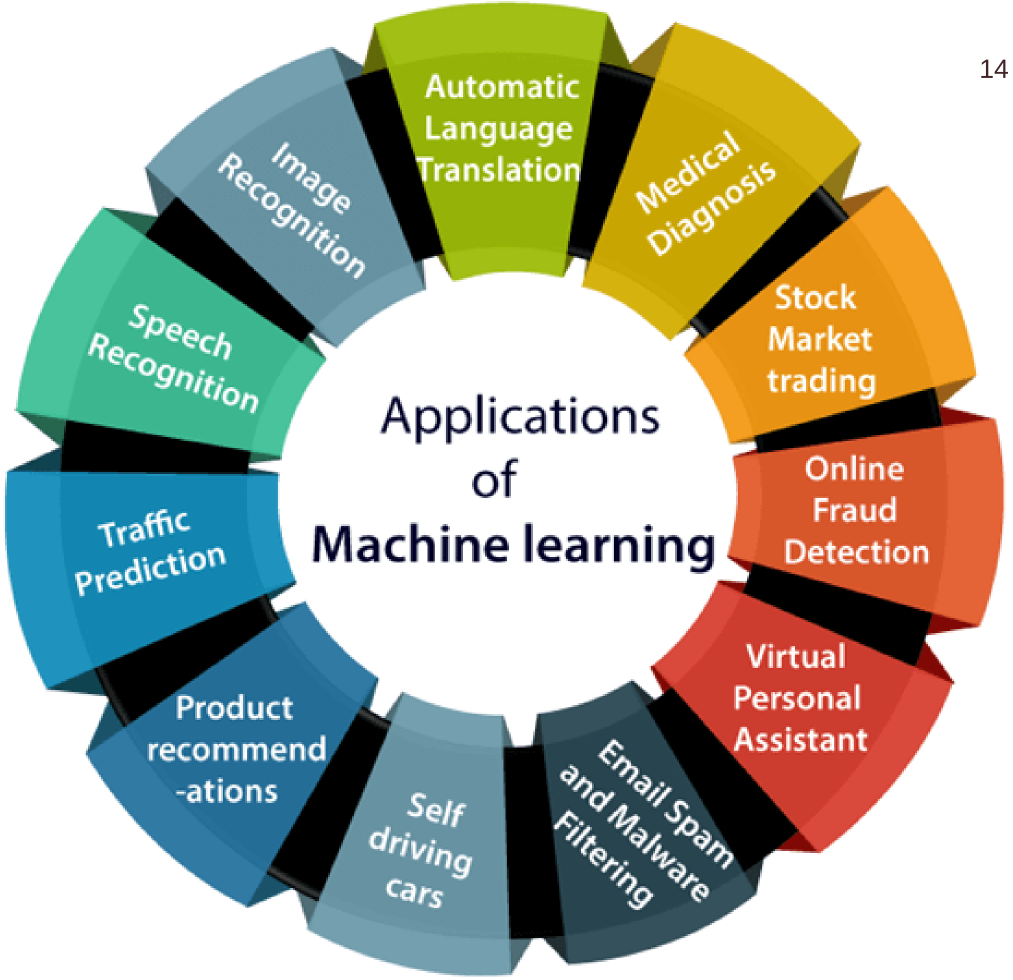
Image Recognition

- It is used to identify objects, persons, places, digital images, etc.
- The popular use case of image recognition and face detection is, **Automatic friend tagging suggestion**:
- .


Applications of Machine learning
Automatic Language Translation
Medical Diagnosis
Stock Market trading
Online Fraud Detection
Virtual Personal Assistant
Email Spam and Malware Filtering
Self driving cars
Product recommend -ations
Traffic Prediction
Speech Recognition
Image Recognition

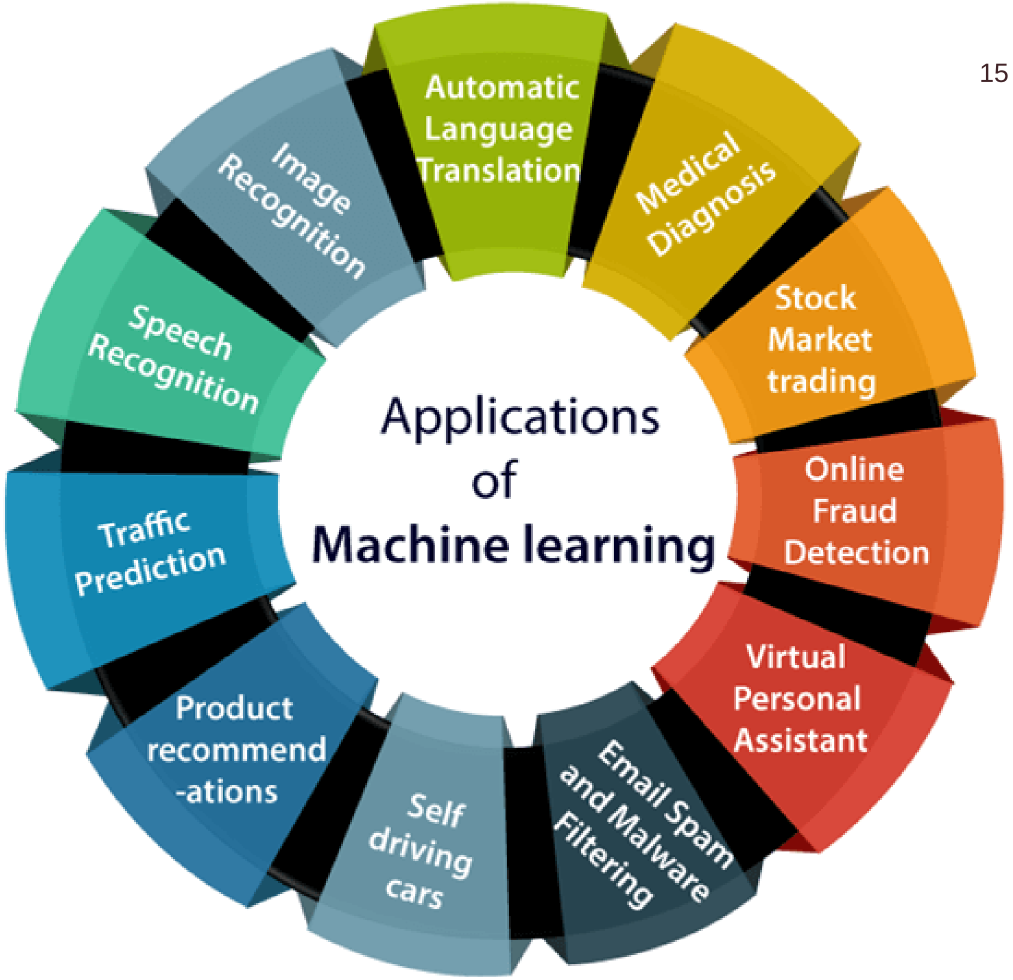
Speech Recognition

- Speech recognition is a process of converting voice instructions into text, and it is also known as "**Speech to text**“.
- Various applications of speech recognition: **Google assistant**, **Siri**, **Cortana**, and **Alexa** are using speech recognition technology to follow the voice instructions.


Applications of Machine learning
Automatic Language Translation
Medical Diagnosis
Stock Market trading
Online Fraud Detection
Virtual Personal Assistant
Email Spam and Malware Filtering
Self driving cars
Product recommend -ations
Traffic Prediction
Speech Recognition
Image Recognition

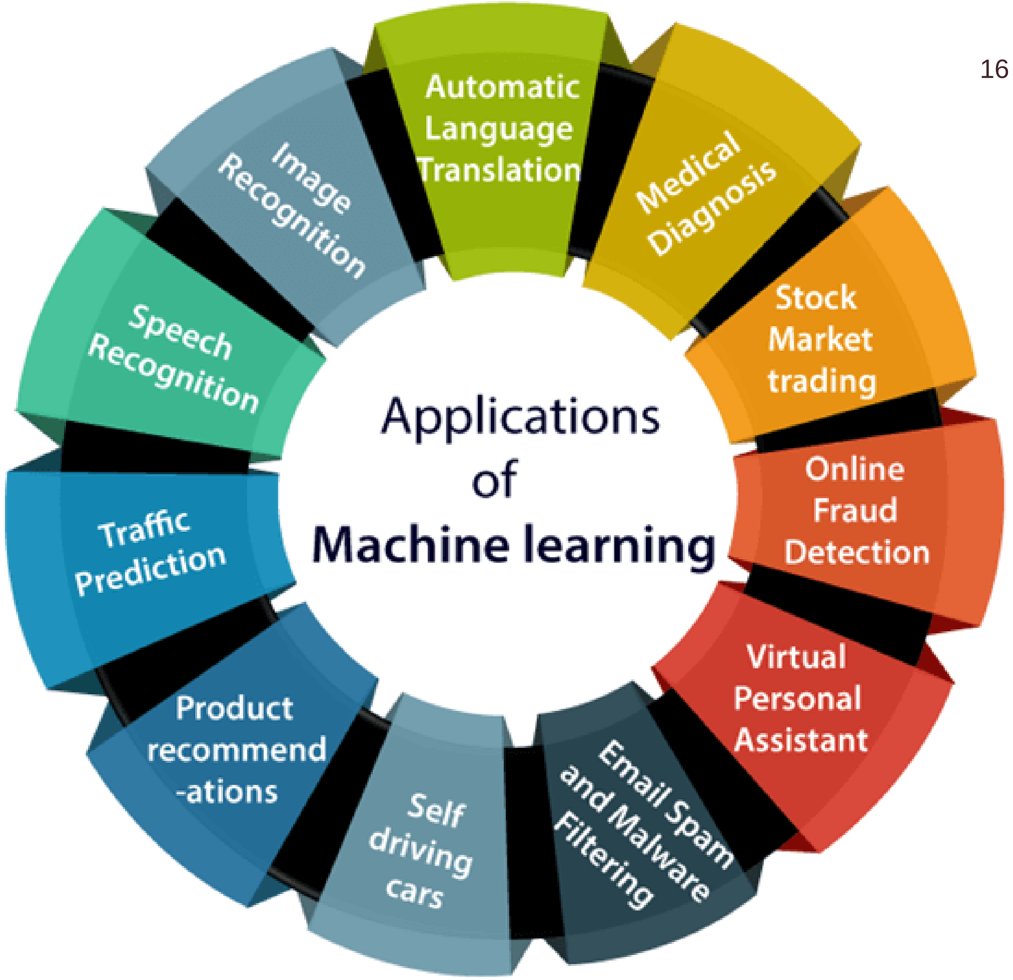
Product recommendations

- Machine learning is widely used by various e-commerce and entertainment companies such as **Amazon**, **Netflix**, etc., for product recommendation to the user.


Applications of Machine learning
Automatic Language Translation
Medical Diagnosis
Stock Market trading
Online Fraud Detection
Virtual Personal Assistant
Email Spam and Malware Filtering
Self driving cars
Product recommend -ations
Traffic Prediction
Speech Recognition
Image Recognition

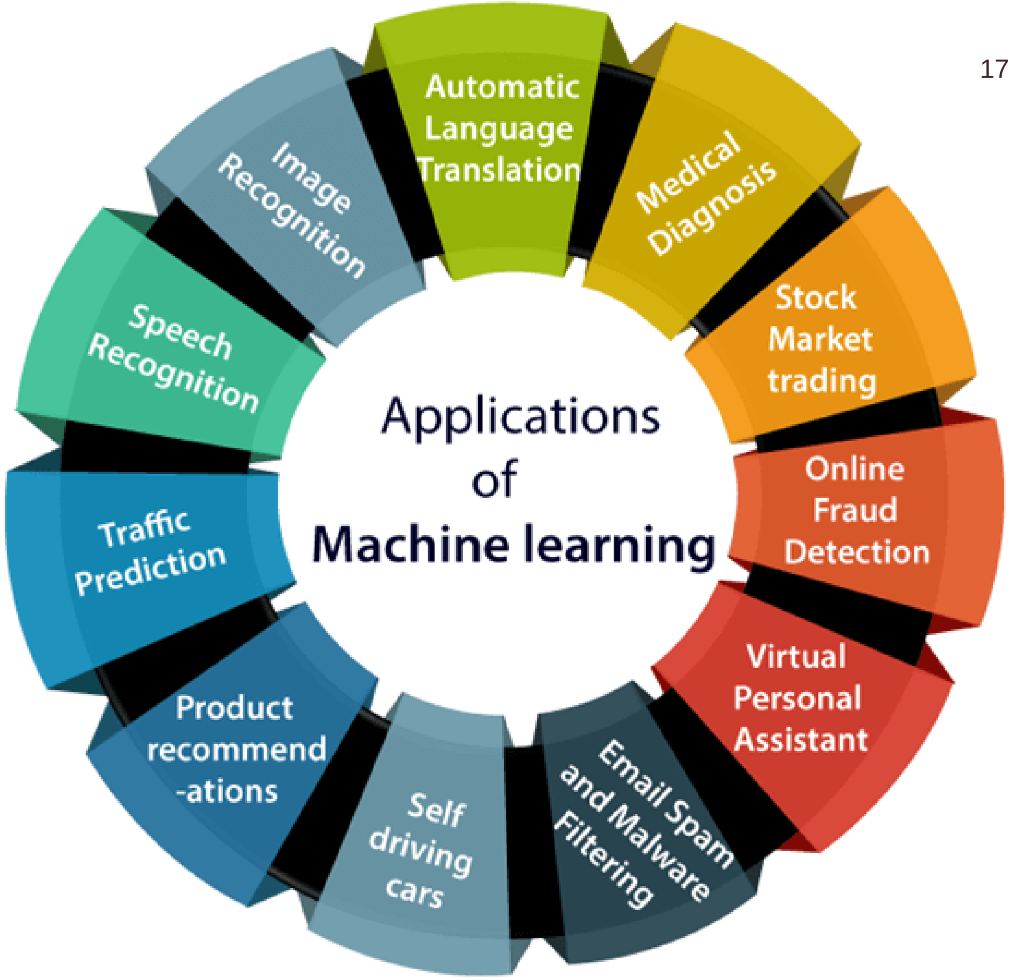
Stock Market trading

- Machine learning is widely used in stock market trading. In the stock market, there is always a risk of up and downs in shares, so for this machine learning's **long short term memory neural network** is used for the prediction of stock market trends.


Applications of Machine learning
Automatic Language Translation
Medical Diagnosis
Stock Market trading
Online Fraud Detection
Virtual Personal Assistant
Email Spam and Malware Filtering
Self driving cars
Product recommend -ations
Traffic Prediction
Speech Recognition
Image Recognition

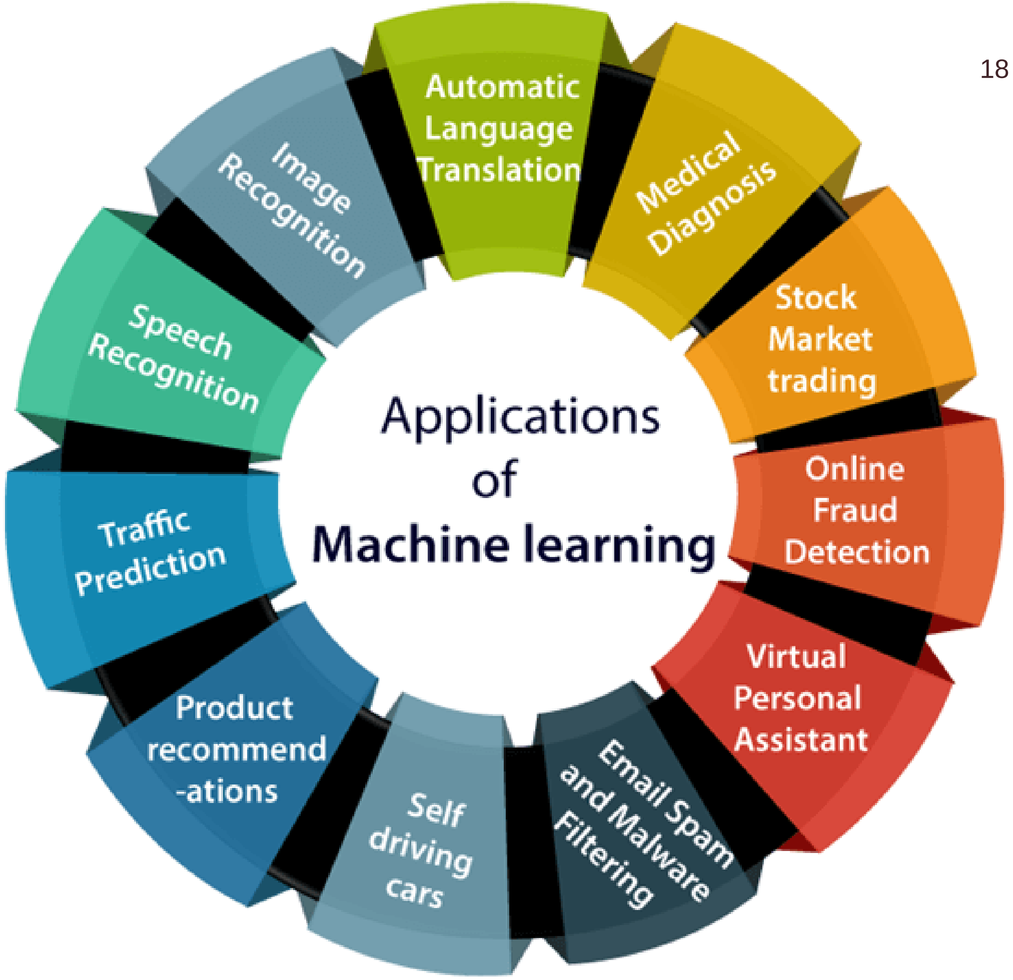
Medical Diagnosis

- In medical science, machine learning is used for diseases diagnoses.


Applications of Machine learning
Automatic Language Translation
Medical Diagnosis
Stock Market trading
Online Fraud Detection
Virtual Personal Assistant
Email Spam and Malware Filtering
Self driving cars
Product recommend -ations
Traffic Prediction
Speech Recognition
Image Recognition

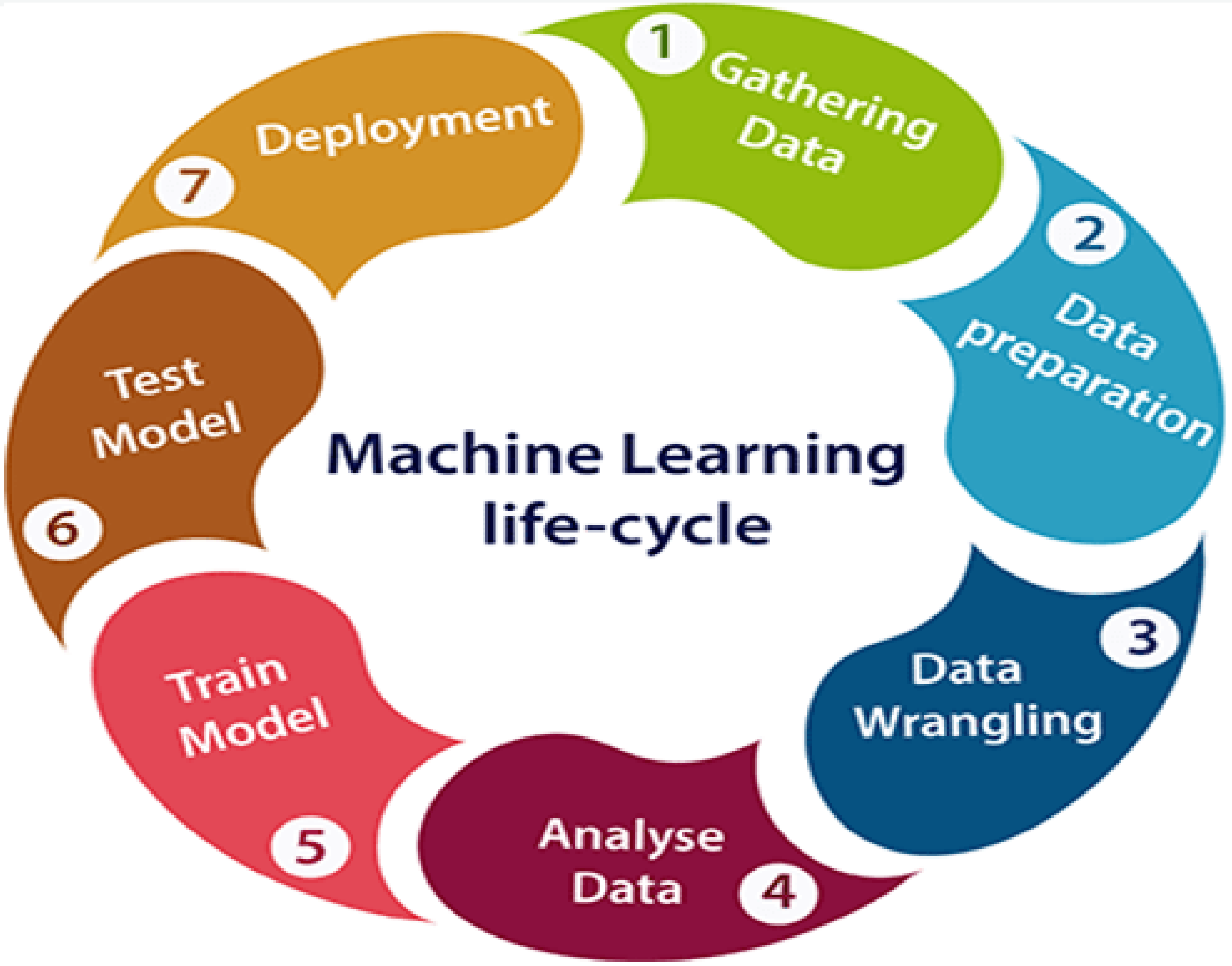
Automatic Language Translation

- Google's GNMT (Google Neural Machine Translation) provide this feature, which is a Neural Machine Learning that translates the text into our familiar language, and it called as automatic translation.


Applications of Machine learning
Automatic Language Translation
Medical Diagnosis
Stock Market trading
Online Fraud Detection
Virtual Personal Assistant
Email Spam and Malware Filtering
Self driving cars
Product recommend -ations
Traffic Prediction
Speech Recognition
Image Recognition

ML Life cycle

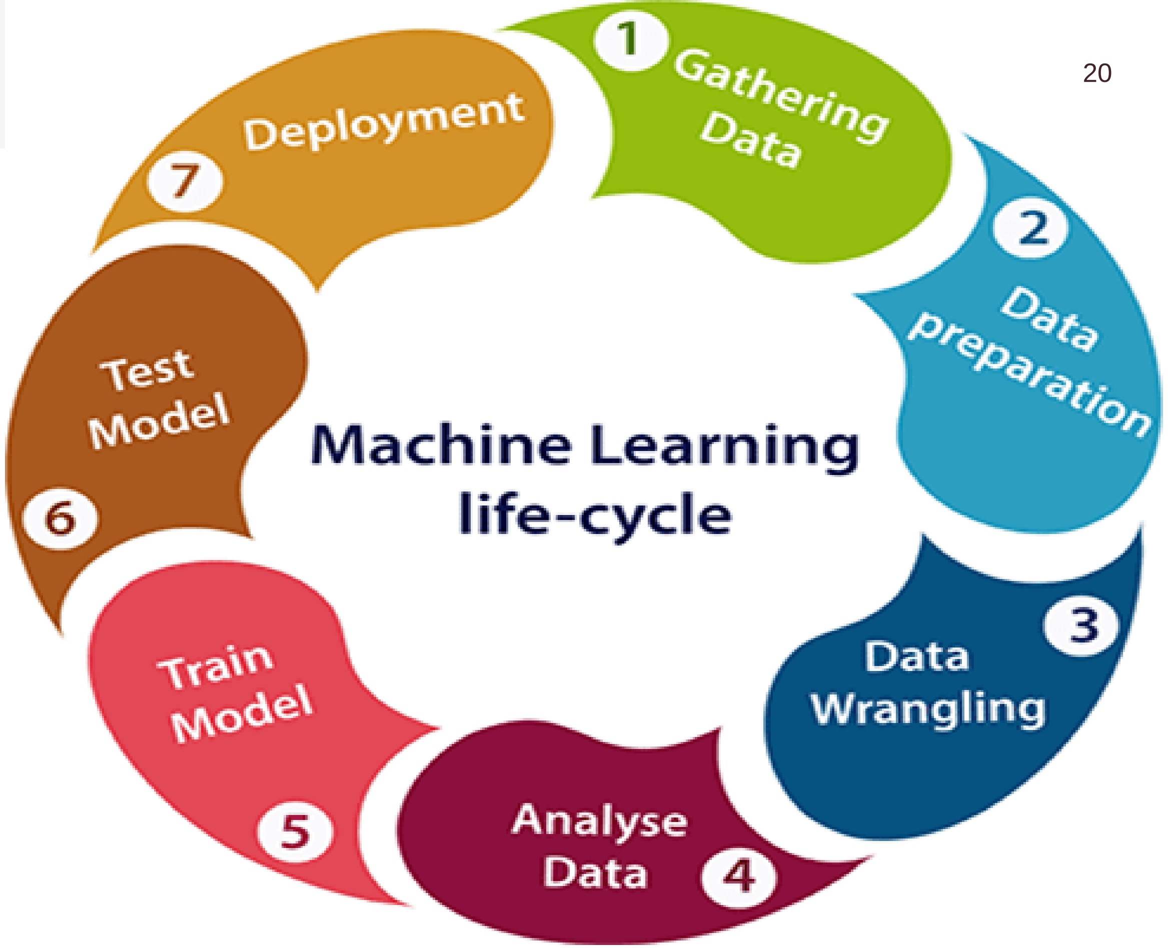
- Machine learning life cycle is a cyclic process to build an efficient machine learning problem.



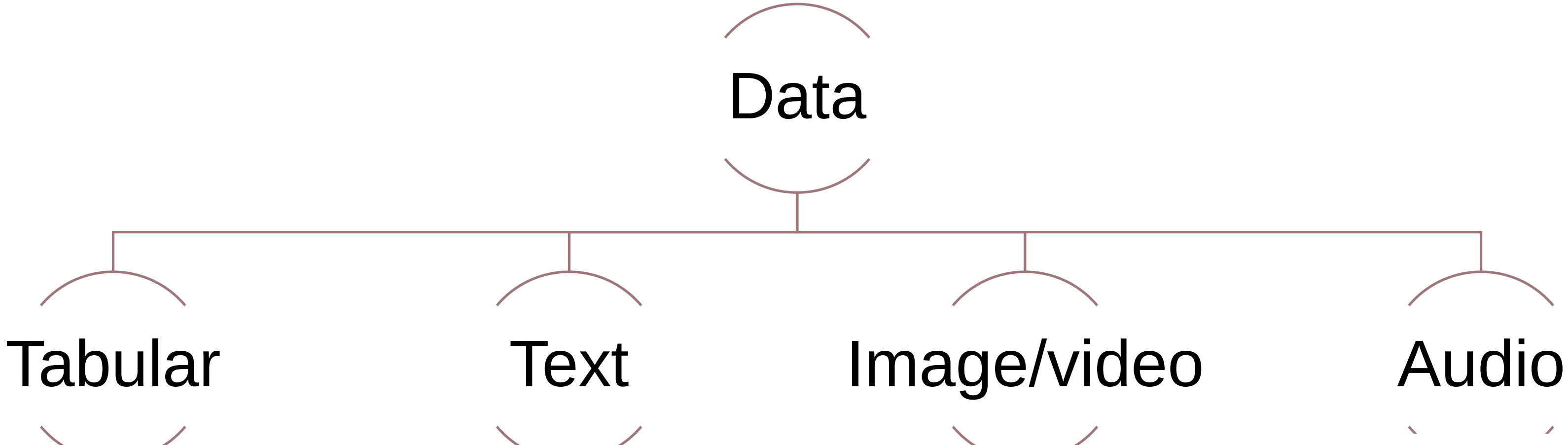
https://www.javatpoint.com/machine-learning-life-cycle

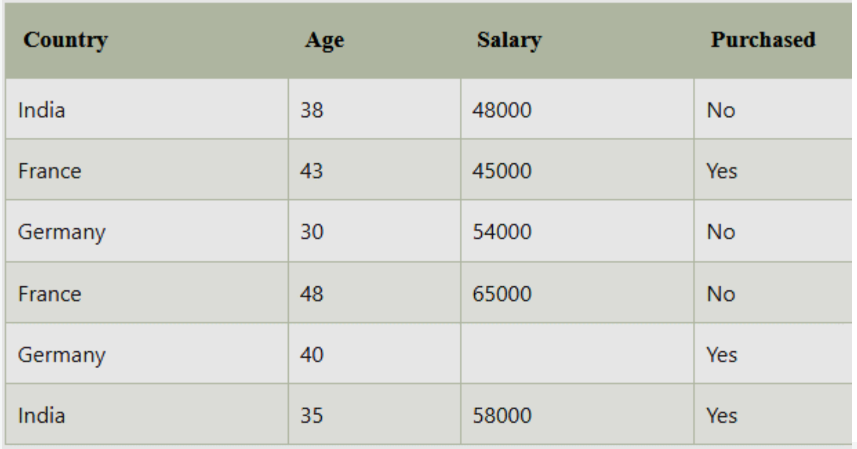
1- Gathering data (Dataset)

- 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.

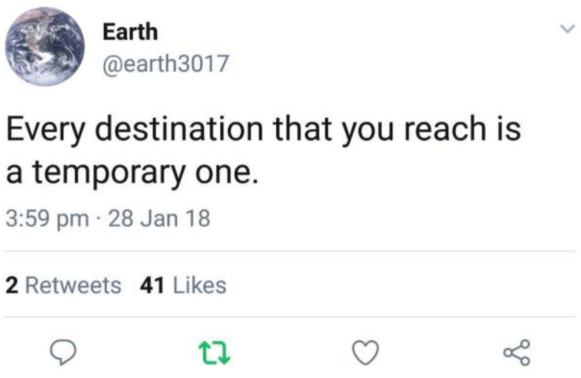


Types of data in datasets





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

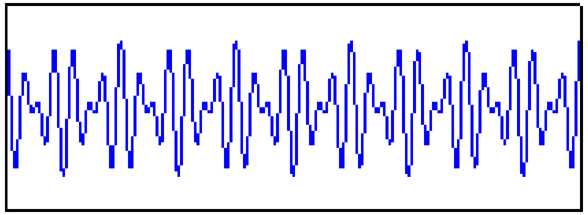


Earth
@earth3017

Every destination that you reach is a temporary one.

3:59 pm · 28 Jan 18

2 Retweets 41 Likes



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

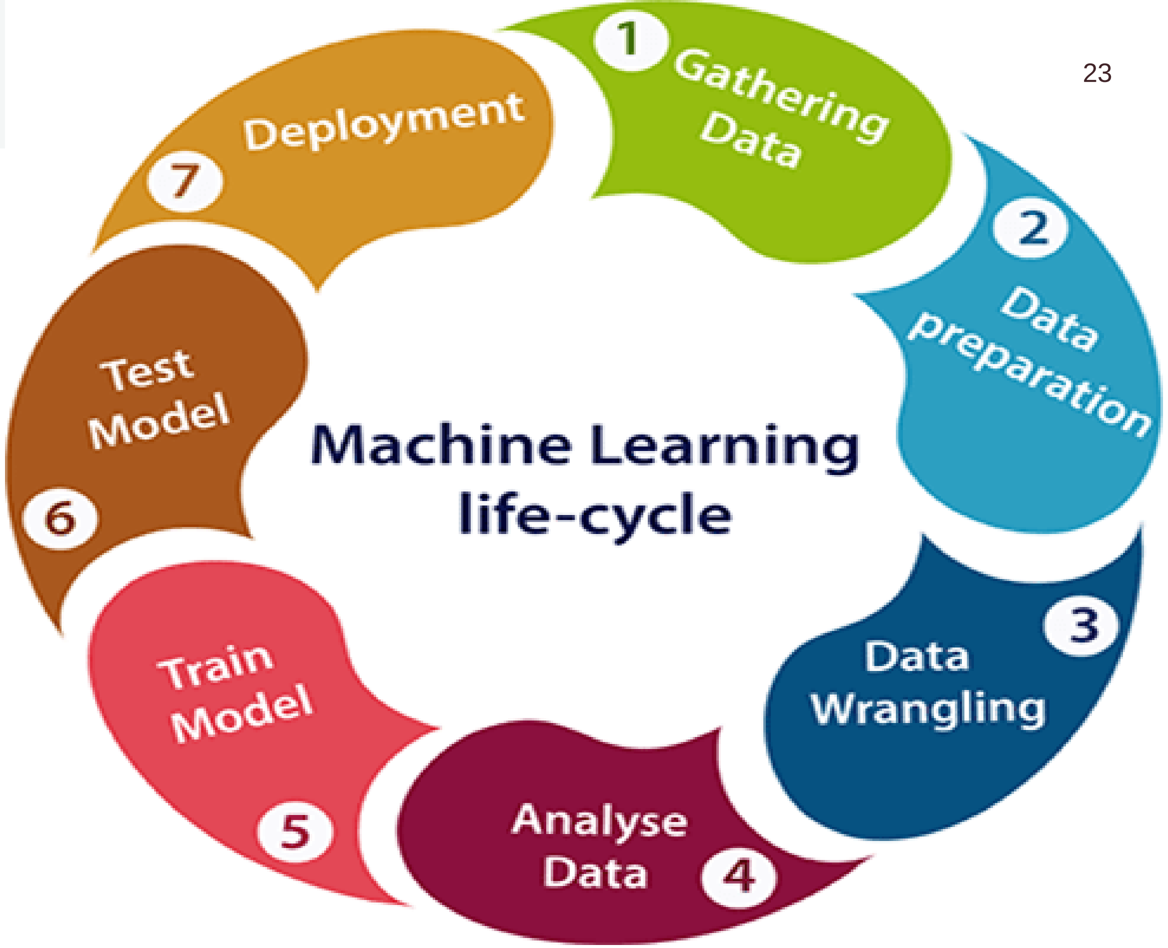
Popular sources for ML datasets

- **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.

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

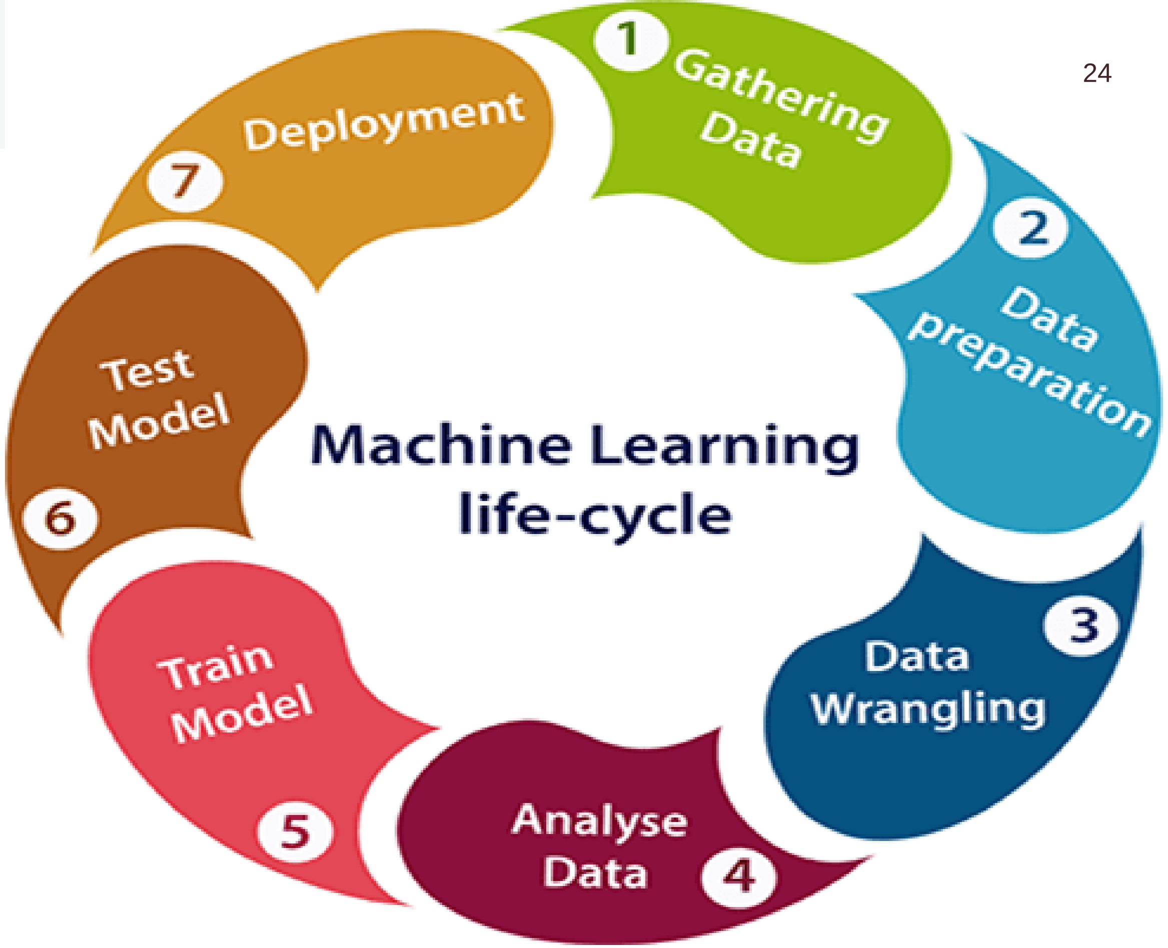
2- Data preparation

- 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.



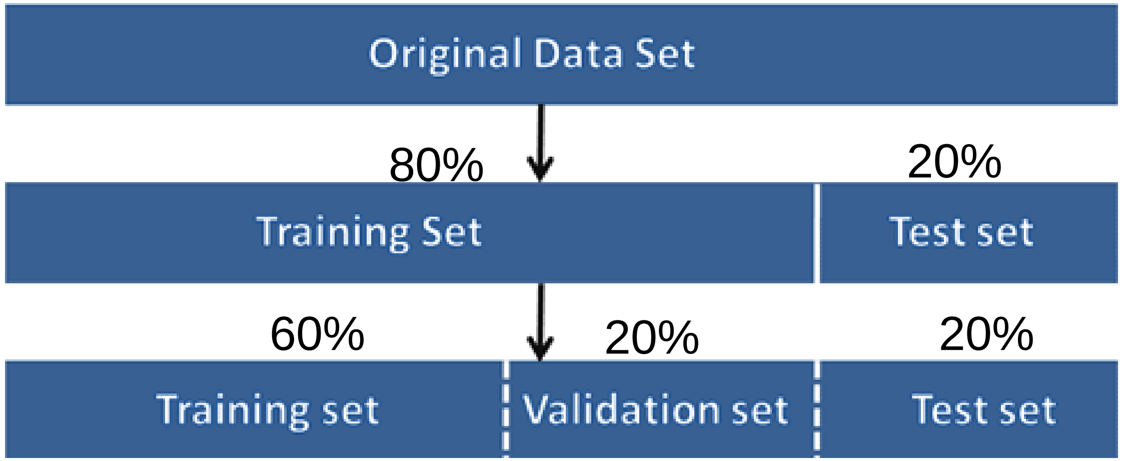
3- Data Wrangling (Pre-processing)

- 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.



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**



Original Data Set

80%

20%

Training Set

Test set

60%

20%

20%

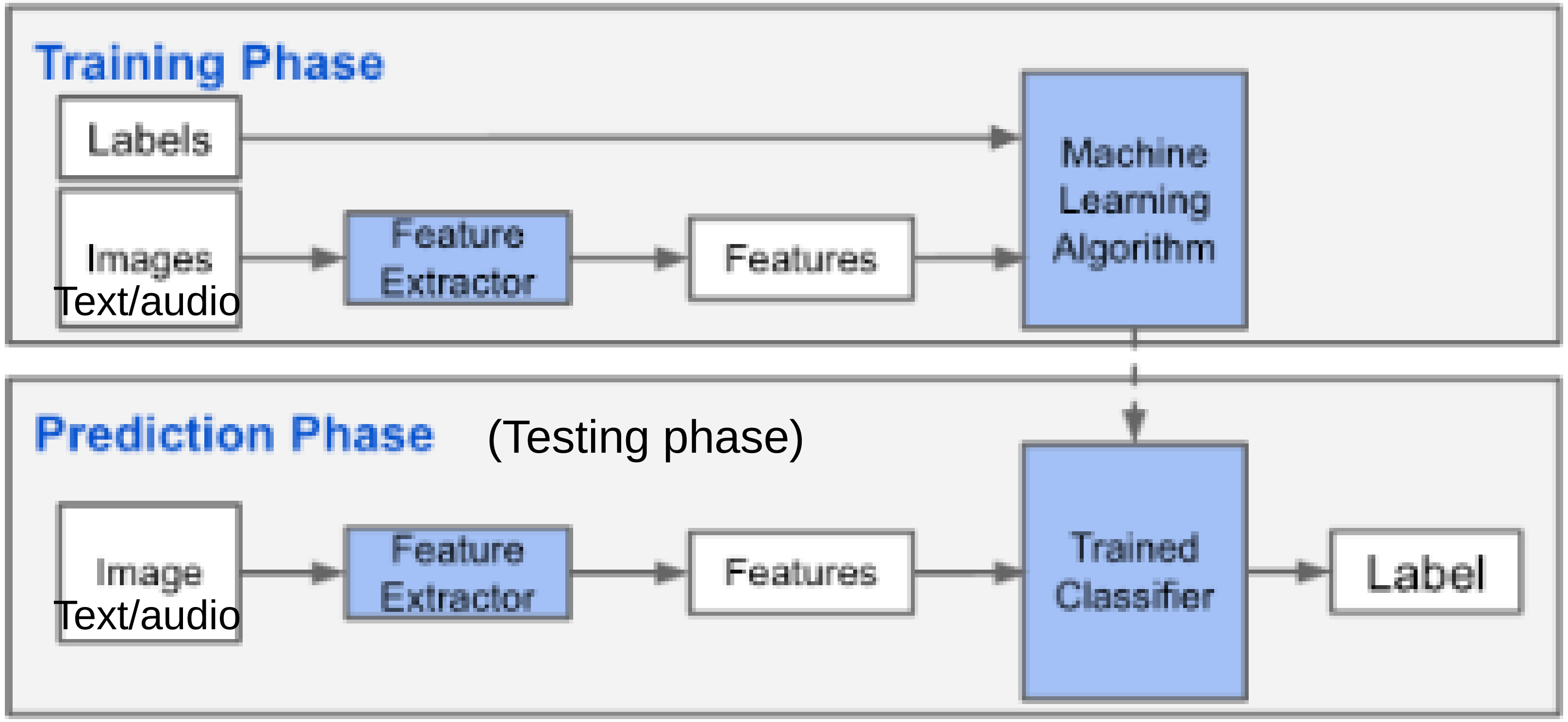
Training set

Validation set

Test set

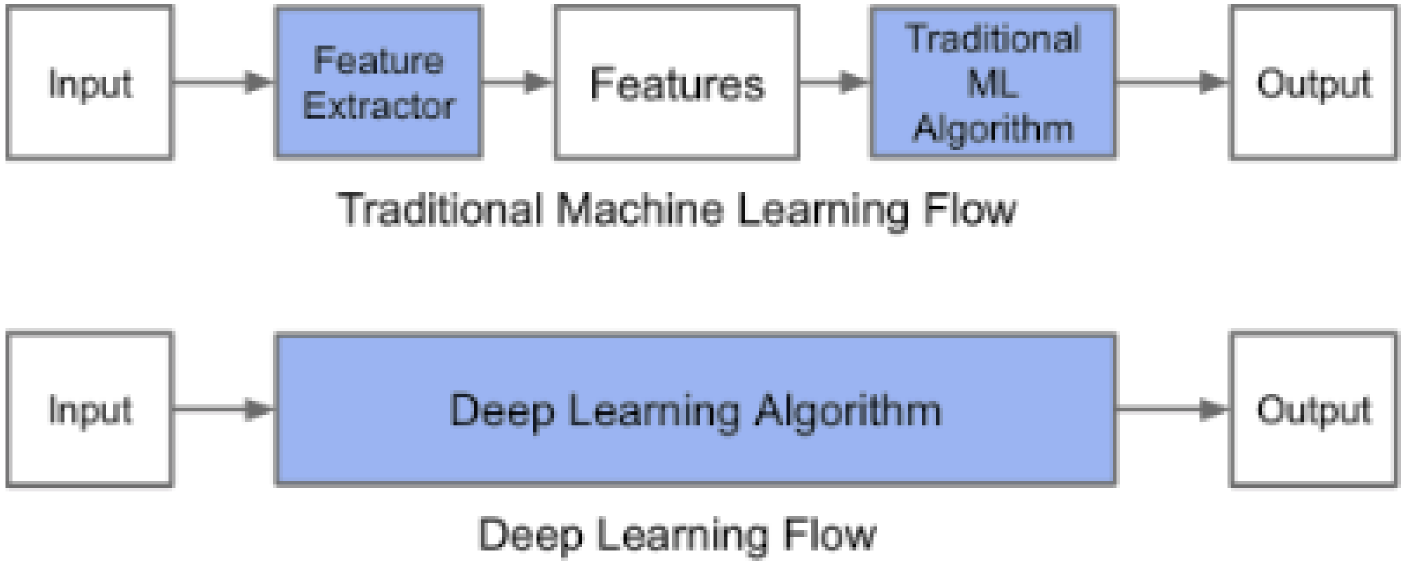
Feature Extraction (Selection)

- 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

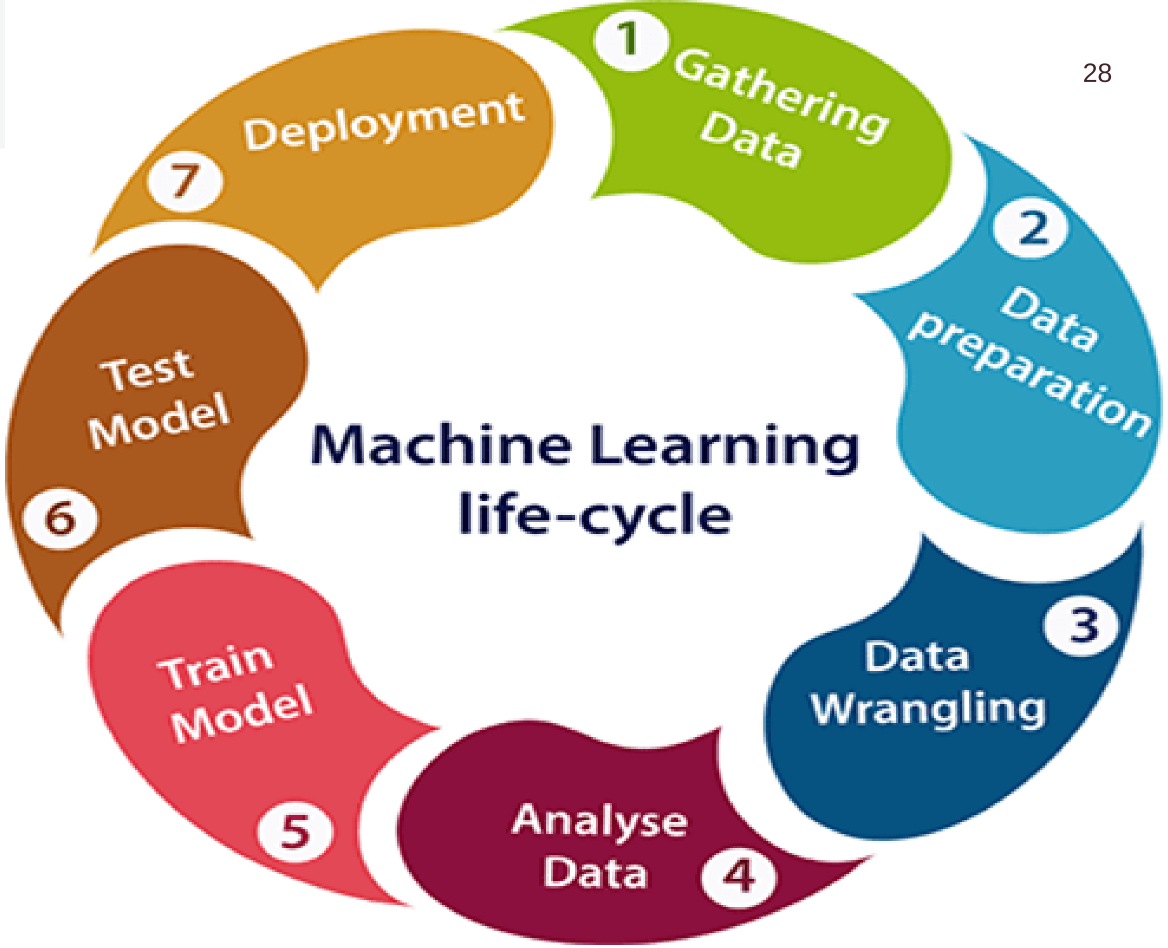
- This step involves:

a)Selection of analytical techniques

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

b) Building & evaluate models

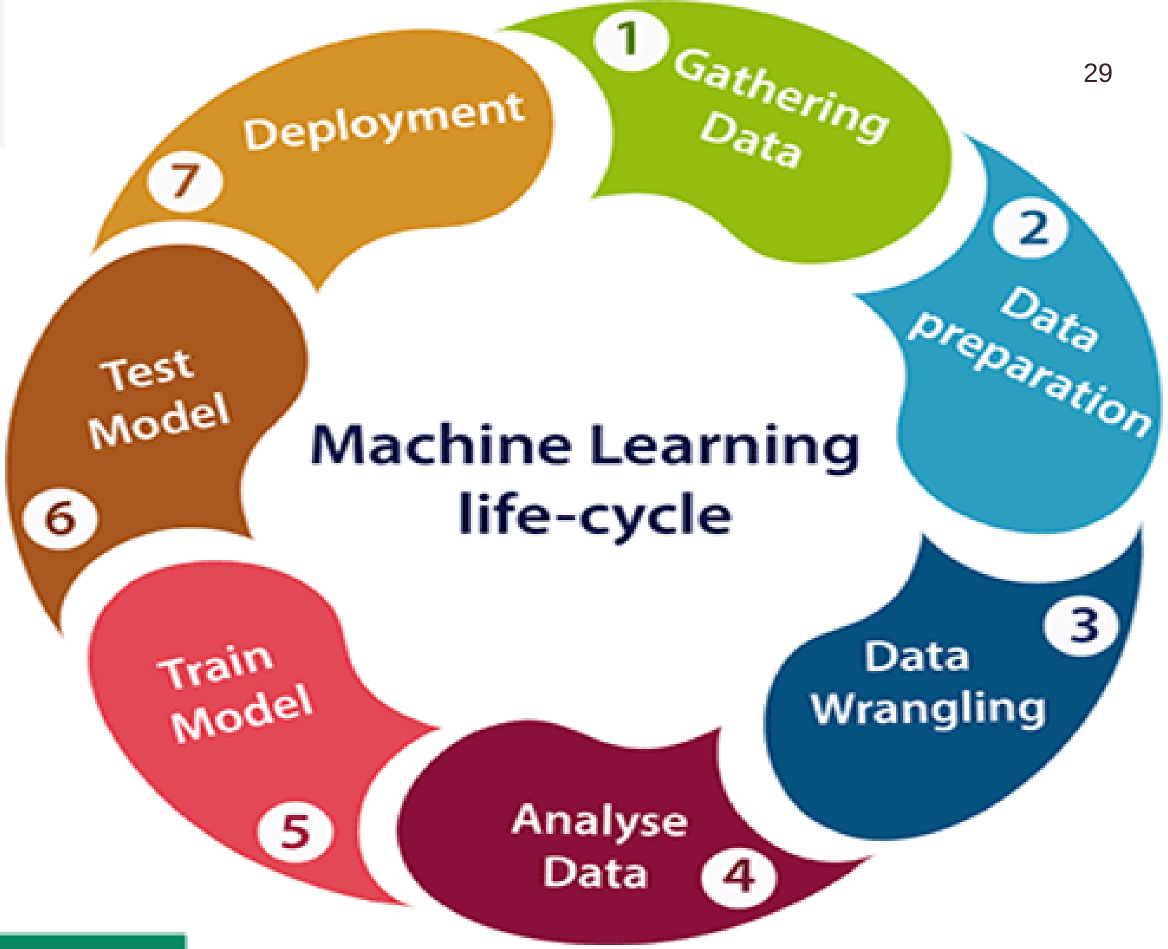
- build the selected model using prepared data, and evaluate the model.

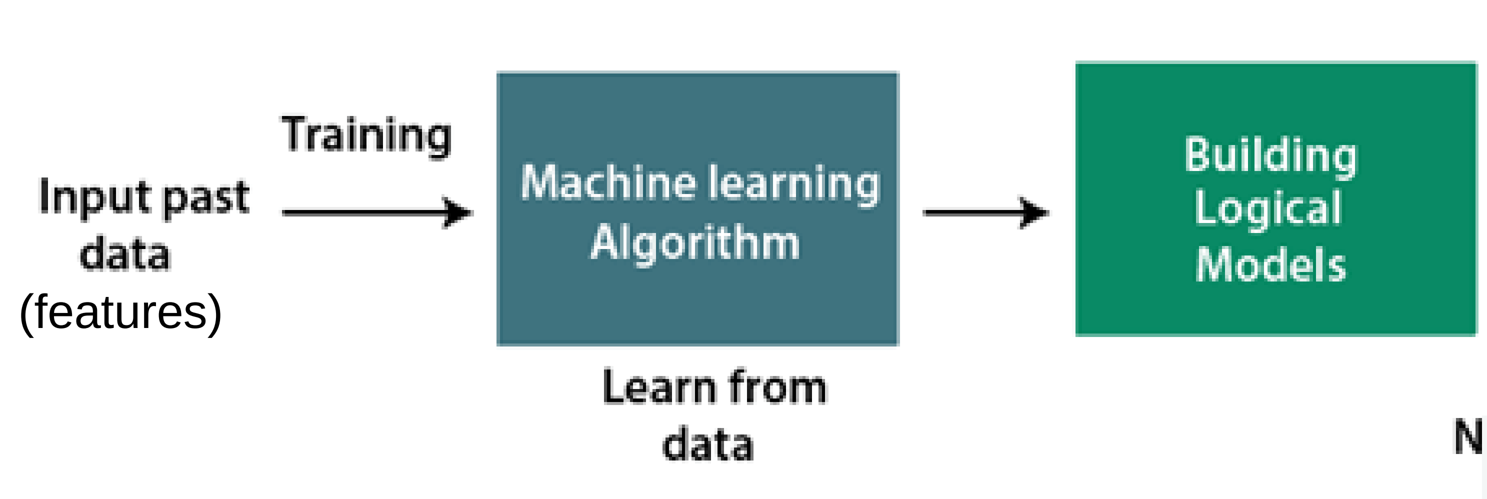


Our course start

5- Train Model

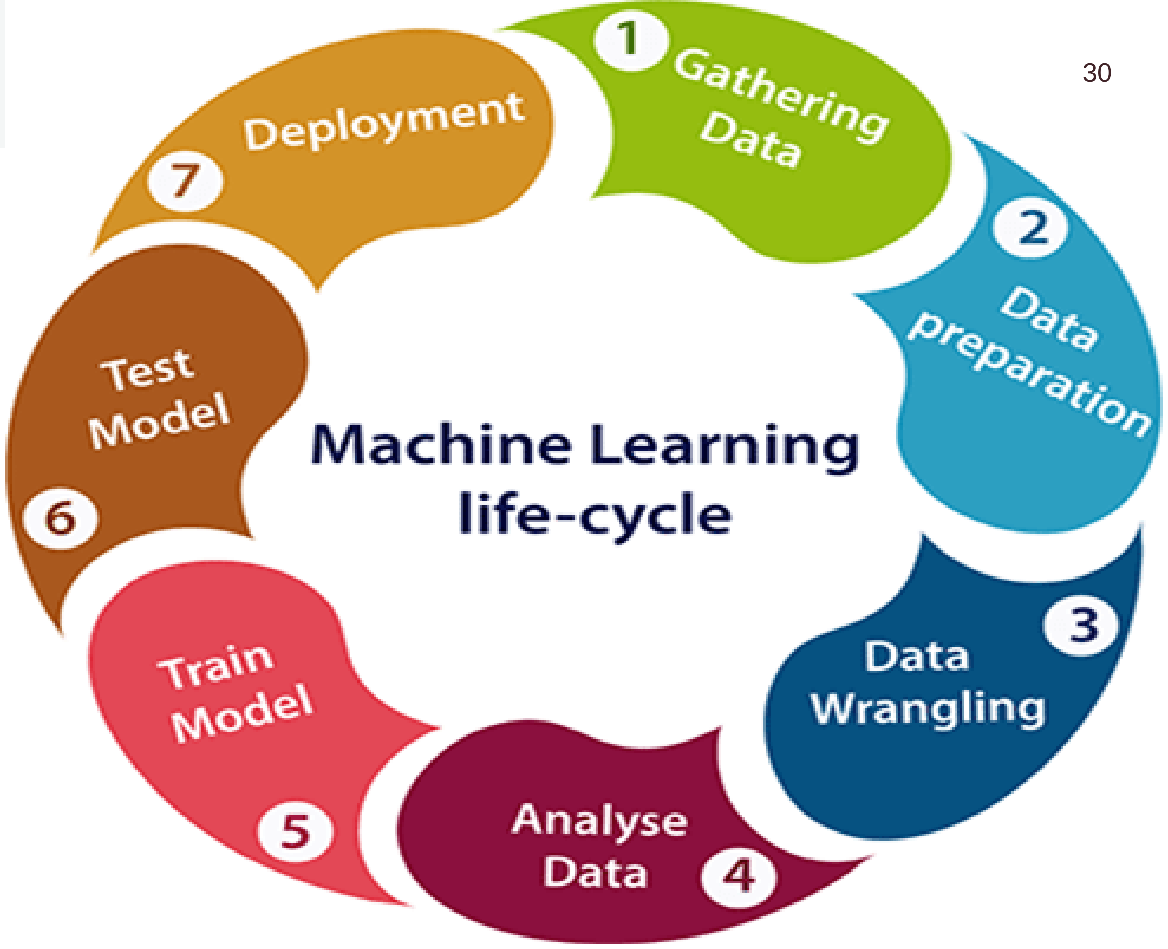
- We use datasets to train the model using various machine learning algorithms. Training a model is required so that it can understand the various **patterns, and rules.**

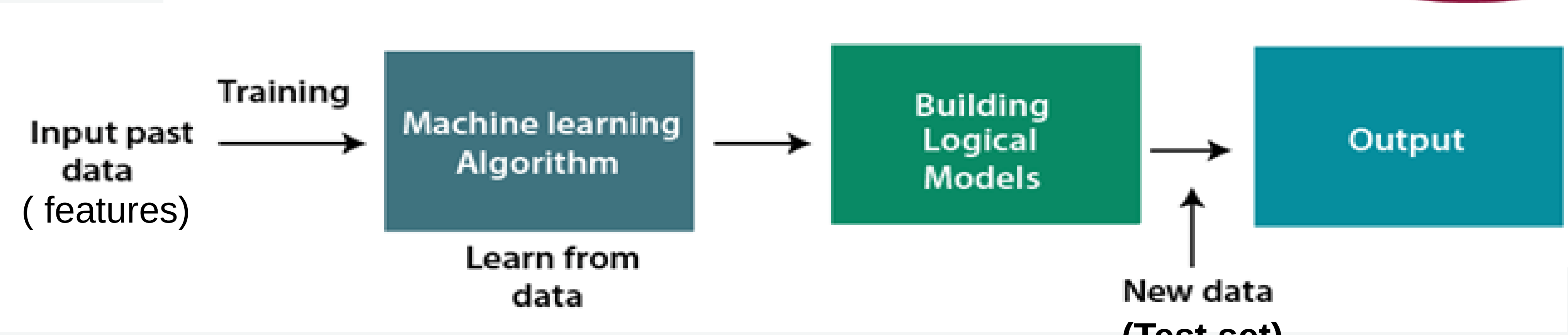




6- Test Model

- 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.



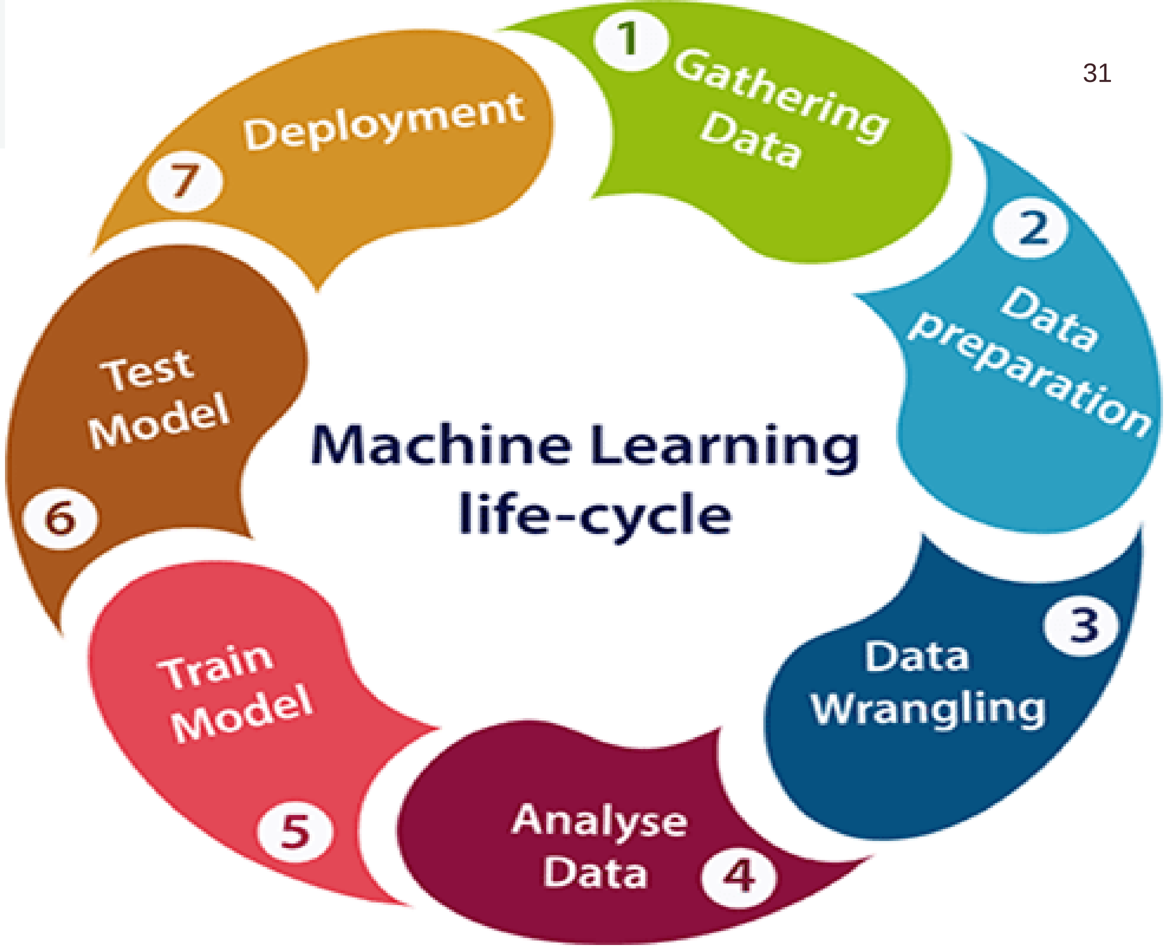


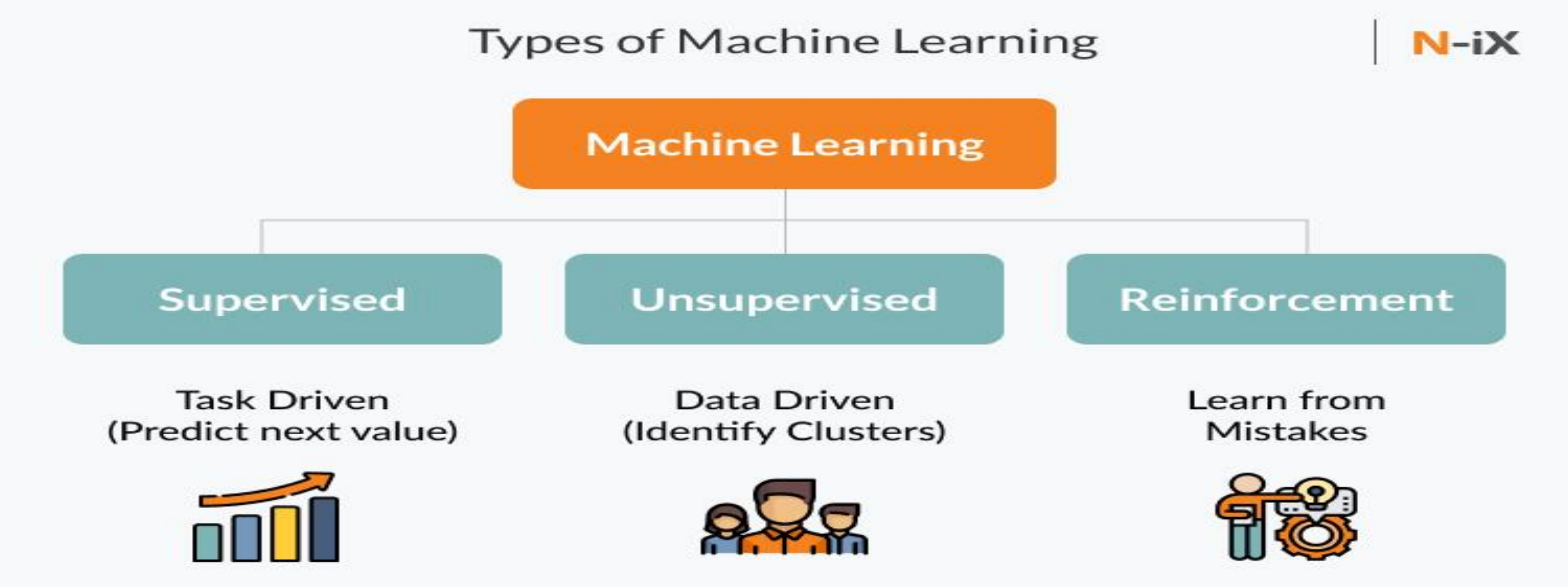
(features)

(Test set)

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

N-iX

Machine Learning

Supervised

Unsupervised

Reinforcement

Task Driven
(Predict next value)

Data Driven
(Identify Clusters)

Learn from
Mistakes

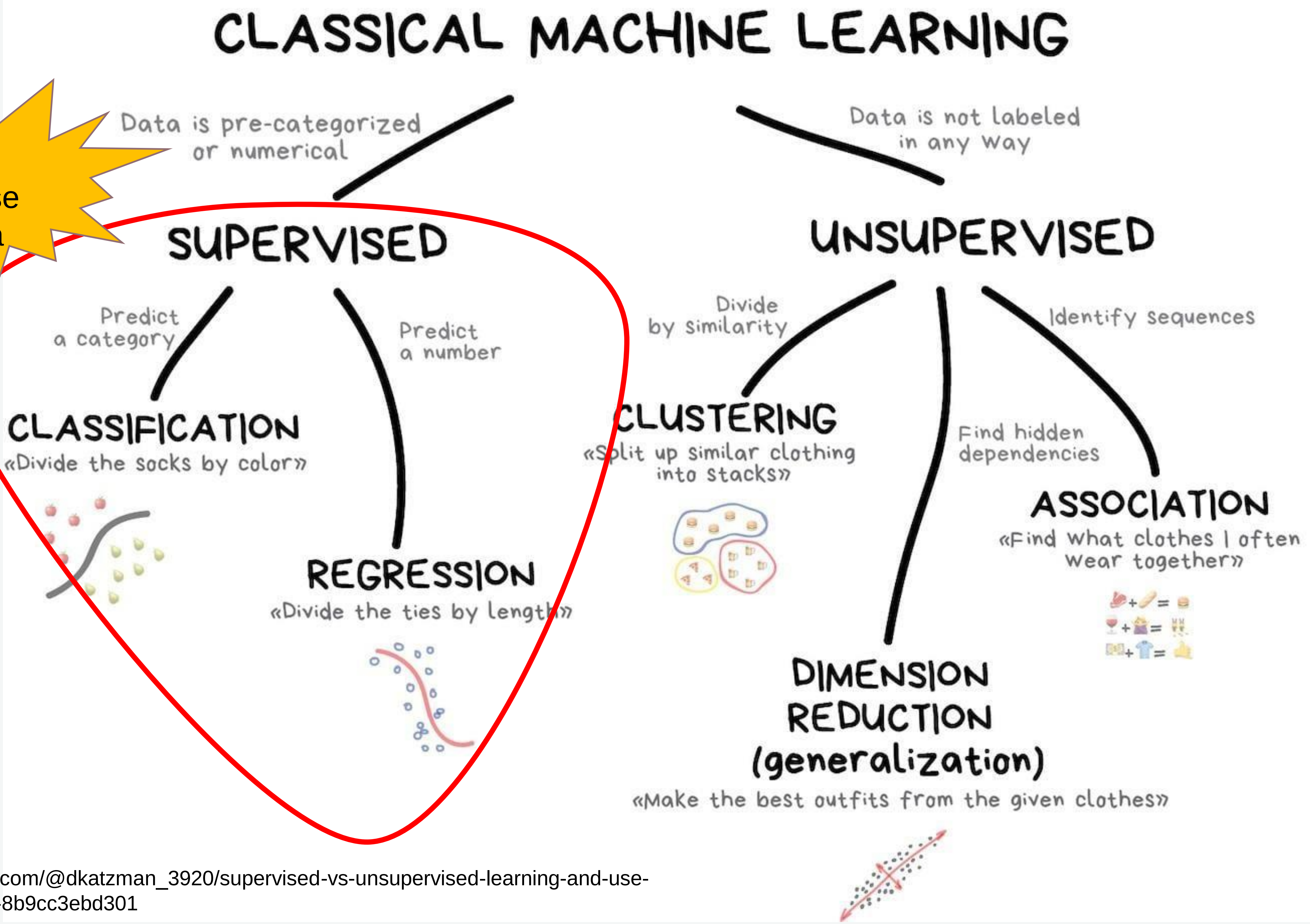
Example:
analyzing customer data
to predict if a person clicks
on an ad or not.

Example:
market segmentation.

Examples:
driverless cars, games.

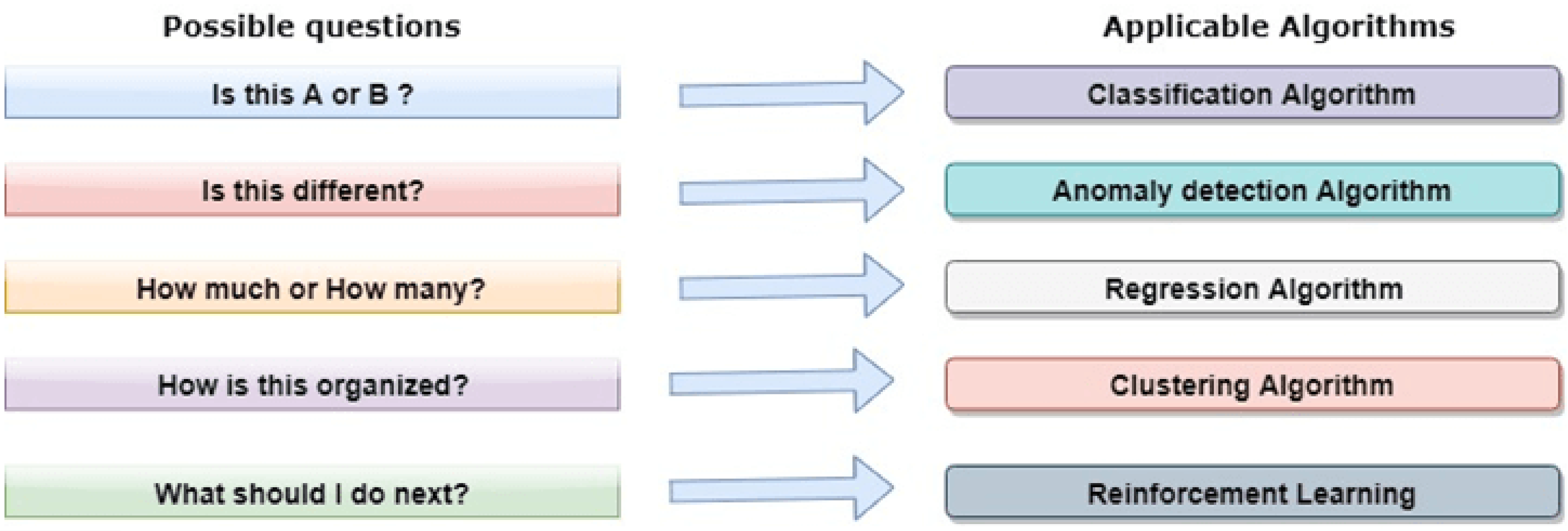
https://www.n-ix.com/deep-learning-vs-machine-learning/

Our course area


CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
Divide by similarity
Identify sequences
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
CLUSTERING
«Split up similar clothing into stacks»
Find hidden dependencies
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

https://medium.com/@dkatzman_3920/supervised-vs-unsupervised-learning-and-use-cases-for-each-8b9cc3ebd301

Applicable algorithms for possible questions

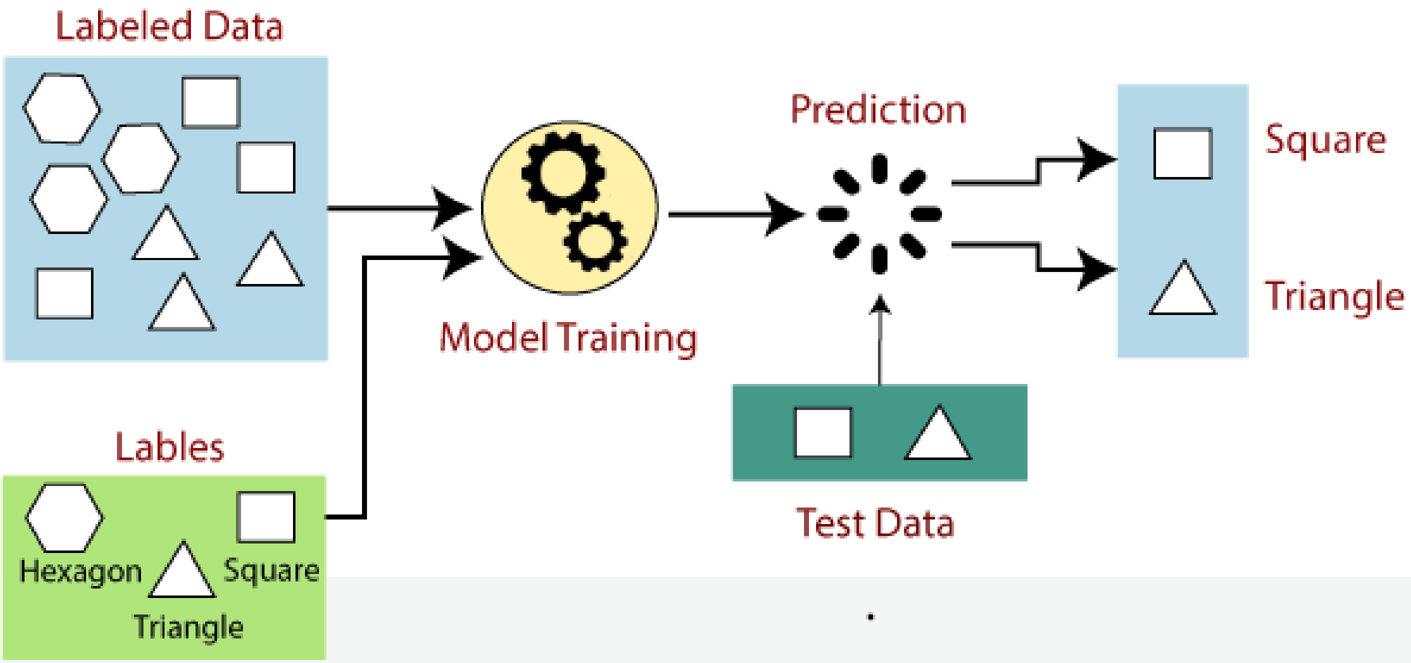


Possible questions	Applicable Algorithms
Is this A or B ?	Classification Algorithm
Is this different?	Anomaly detection Algorithm
How much or How many?	Regression Algorithm
How is this organized?	Clustering Algorithm
What should I do next?	Reinforcement Learning

https://www.javatpoint.com/data-science

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 an **unlabeled input data**, which means it is not categorized and corresponding outputs are also not given.
- **Clustering**: Clustering is a method **of grouping the objects into clusters** such that objects with most similarities remains into a group and has less or no similarities with the objects of another group.
- **Association**: An association rule is an unsupervised learning method which is used for **finding the relationships between variables** in the large database. It determines the set of items that occurs together in the dataset. Association rule makes marketing strategy more effective. **Example:** People who buy X item (suppose a bread) are also tend to purchase Y (Butter/Jam) item.

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.

Join **Selected CS1(Fall2021)** Team Class

- To communicate with us and get course materials
- Team class code: 2molk1s

Thanks