Unit – IV Introduction to ML

Teaching Hours: 07 **Marks: 10**

Topics & Subtopics

4.1 History and Evaluation of ML, AI vs ML
4.2 Machine Learning Life Cycle
4.3 Different forms of Data
4.4 Dataset for ML: Training Dataset, Testing Datasets.
4.5 Data Cleaning: Missing Data, Outliers

4.1 History and Evaluation of ML, AI vs ML

History of Machine Learning

1. 1950 - 1960

   The first machine learning algorithms were developed. And were based on ideas from psychology and statistics.

2. In the 1970s,

   Machine learning began to be used for practical applications, such as image recognition and natural language processing.

3. In the 1980s,

   Support vector machines (SVMs) were introduced, which allowed for more accurate and efficient classification of data.

4.1 History and Evaluation of ML, AI vs ML(Cont…)

4. In the 1990s,

   The rise of statistical learning algorithms, such as Bayesian networks and Hidden Markov Models (HMMs), led to breakthroughs in areas such as speech recognition and machine translation.

5. In 1997,

   The IBM computer Deep Blue, which was a chess playing computer, beat the world chess champion.

6. In the 2000s,

   The advent of deep learning algorithms, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), allowed for more complex and sophisticated machine learning applications.

4.1 History and Evaluation of ML, AI vs ML(Cont…)

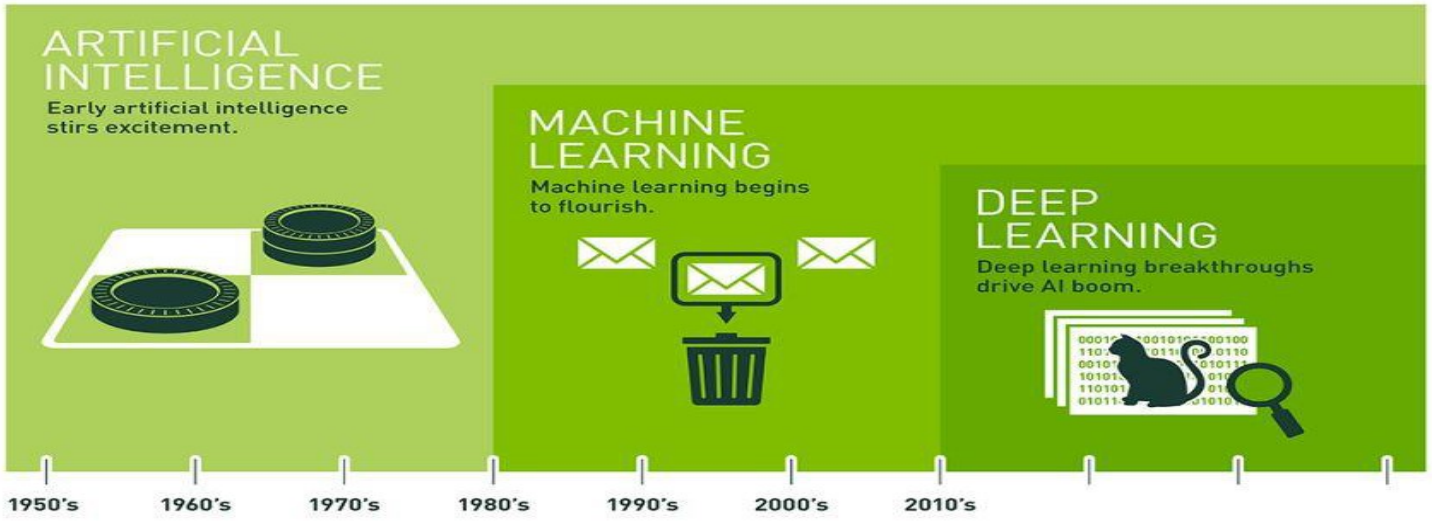
7. The 21st Century

Today, machine learning is used in a wide range of fields :

a) AlphaGo
b) Image recognition and object detection: self-driving cars, security cameras.
c) Speech recognition: such as Siri, Google Assistant, and Alexa.
d) Recommender systems: such as Netflix and Amazon.
e) Medical diagnosis and prediction
f) Fraud detection
g) Natural language processing : sentiment analysis, and chatbots.

WHAT IS MACHINE LEARNING

Arthur Samuel described it as: “The field of study that gives computers the ability to learn from data without being explicitly programmed.”



MACHINE LEARNING

Machine learning is a scientific discipline that is concerned with the design and development of algorithms that allow computers to learn based on data, such as from sensor data or databases.

A **major focus** of machine learning research is to **automatically learn to recognize complex patterns and make intelligent decisions** based on data .

KIND OF PROBLEMS WHERE MACHINE LEARNING IS APPLICABLE

1. Problems where a) There is no deterministic algorithm (not even of evil complexity) e.g. Recognizing a 3D object from a given scene, Handwriting recognition, Speech recognition

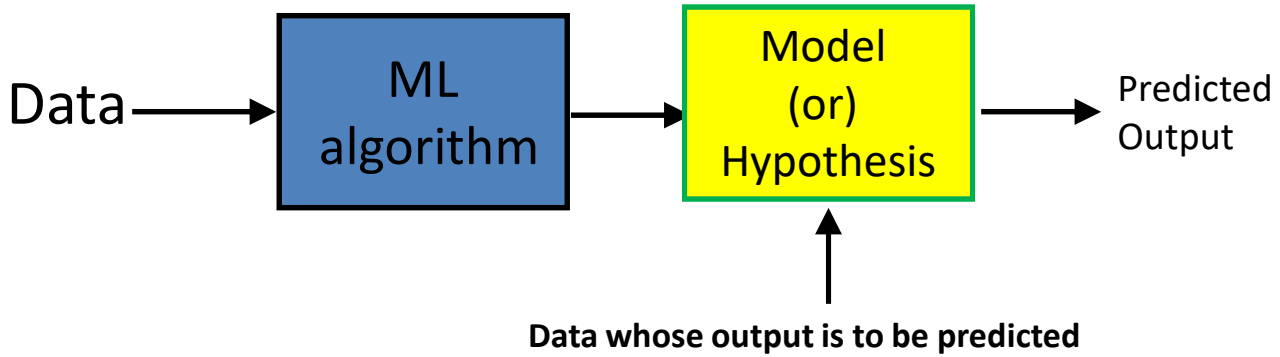
2. Problems which don't have a fix solution and goal posts keep changing. System adapts and learns from experience e.g. SPAM emails, Financial fraud, IT Security Framework

3. Where Solutions are Individual specific or time dependent. e.g. recommendations and targeted advertisements

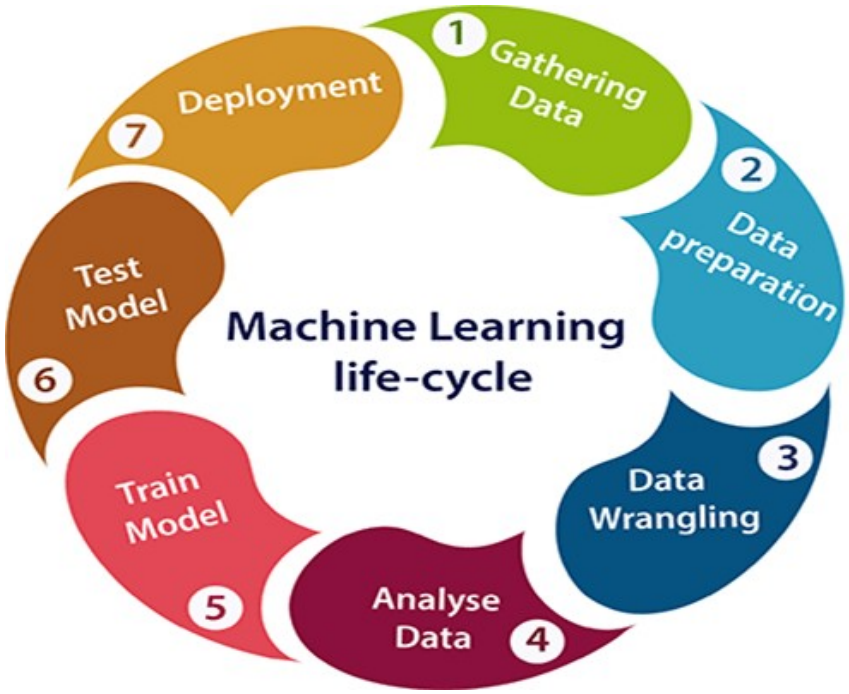
4. For prediction based on past and existing patterns (not defined or defined by huge number of weak rules) e.g. prediction of share prices etc.

Traditional Programming

Machine Learning



4.2 Machine learning Life Cycle



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**

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.
- **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**

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.

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.

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.

4.3 Different forms of Data:

a. Statistics

- Statistics refers to the mathematical and analytical techniques used to collect, interpret, and present data.
- It involves methods of summarizing and drawing inferences from data, such as mean, median, standard deviation, hypothesis testing, and regression analysis.
- This is fundamental component of data analysis, and it is not a form of data itself rather a field of study and a set of tools for working with data.

4.3 Different forms of Data(cont..)

a. Statistics to open a car showroom

Name	Monthly Income ($)
Rob	5000
Rafiq	6000
Nina	4000
Sofia	7500
Mohan	8000
Tao	7000

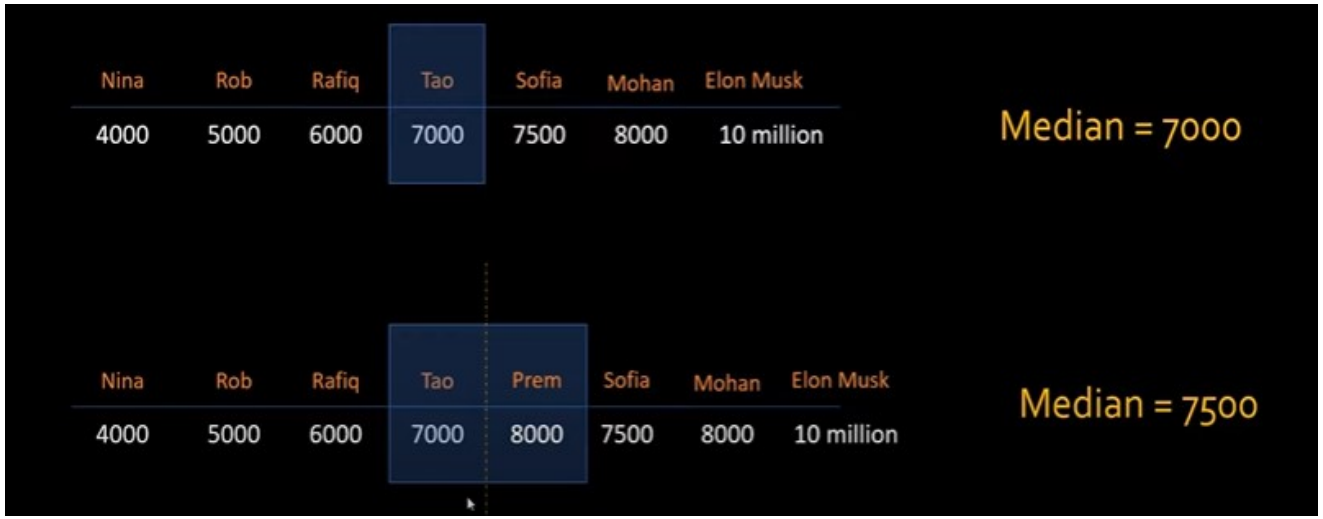
4.3 Different forms of Data(cont..)

a. Statistics to open a car showroom

Name	Monthly Income ($)
Rob	5000
Rafiq	6000
Nina	4000
Sofia	7500
Mohan	8000
Tao	7000
Elon Musk	10 million

4.3 Different forms of Data(cont..)

a. Statistics to open a car showroom



4.3 Different forms of Data:

b. Data Mining

- Data mining is the process of discovering patterns, relationships, or useful information from large datasets.
- It involves techniques such as clustering, classification, association rule mining, and anomaly detection.
- Data mining is a form of data analysis that focuses on finding valuable insights within data.
- Example: Market basket analysis: Identifying items frequently purchased together to inform store layout, product placement, and bundled deals.

4.3 Different forms of Data:

b. Data Mining

- **Personalized Recommendations:**

Analyzing customer purchase history and browsing behavior to suggest relevant products, like on Amazon or Netflix.

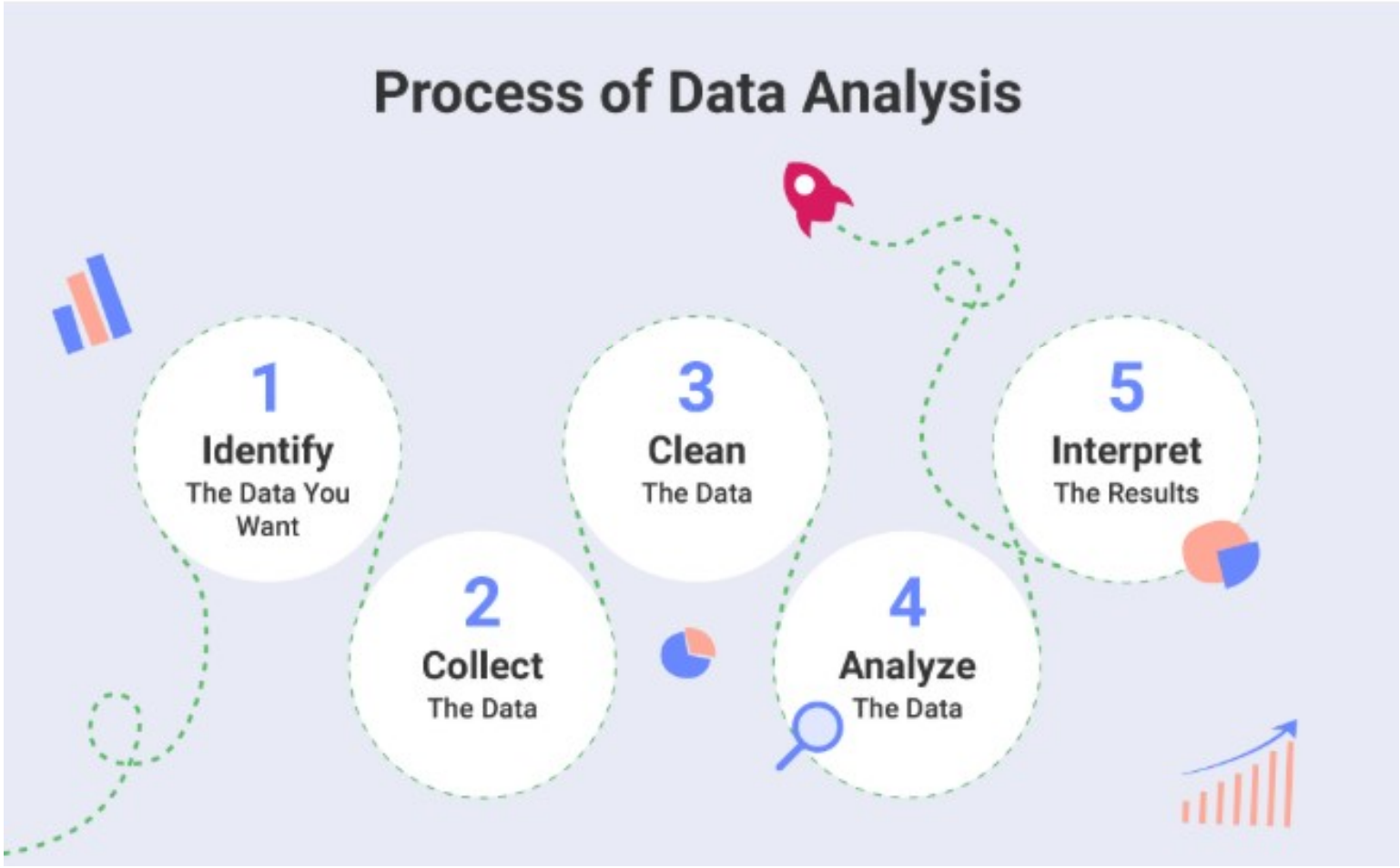
- **Targeted Marketing**: Using customer purchase data to send tailored promotions to specific customer segments.
- **Credit Scoring:** Analyzing financial data to determine creditworthiness for loan.

4.3 Different forms of Data:

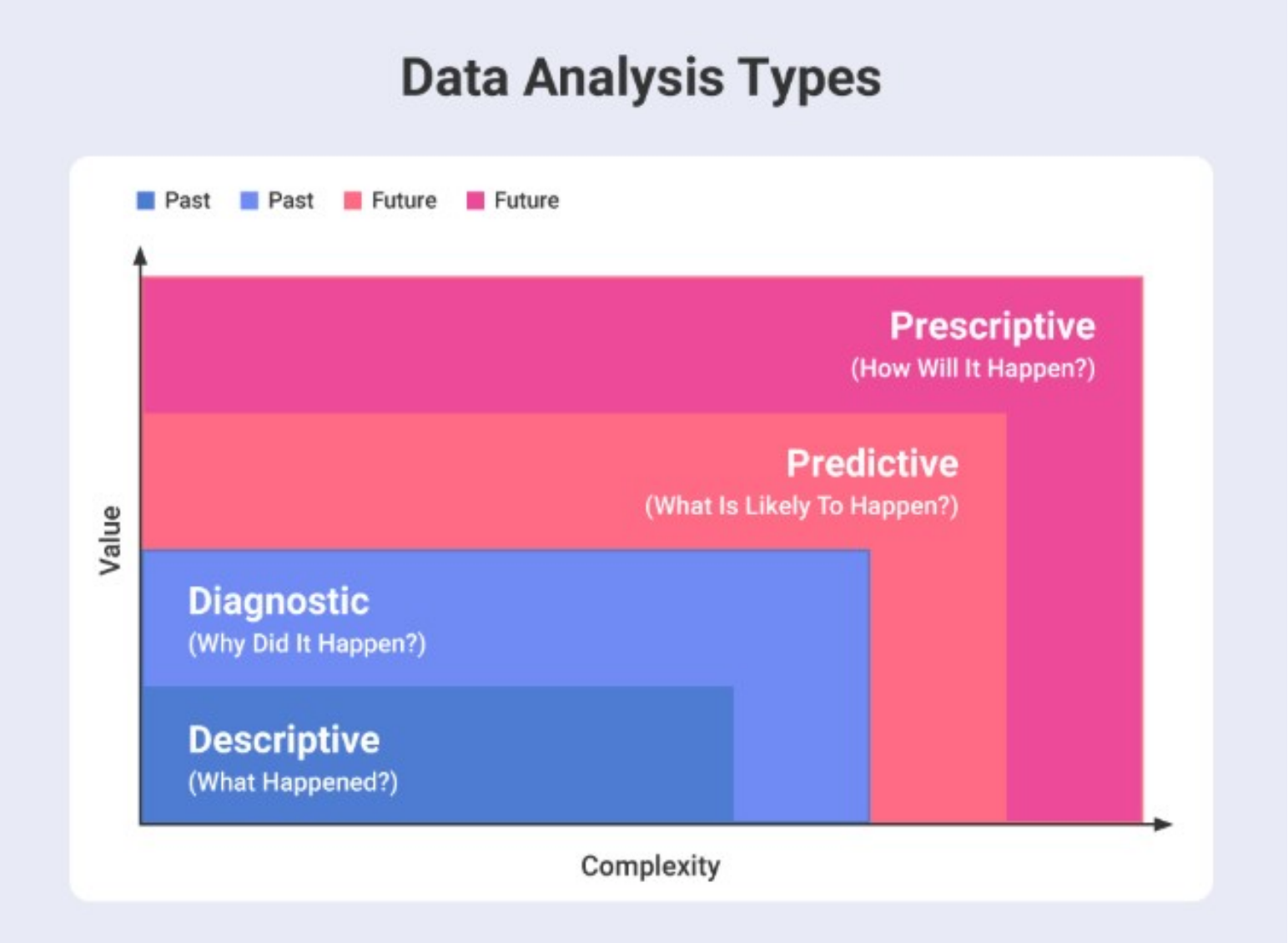
c. Data Analytics

- Data Analytics encompasses the entire process of examining, cleaning, transforming, and interpreting data to extract meaningful insights.
- It combines statistical analysis, data mining, and visualization to inform decision making.
- Data analytics is the broader practice of working with data to answer questions or make informed decisions.

4.3 Different forms of Data(Cont..)

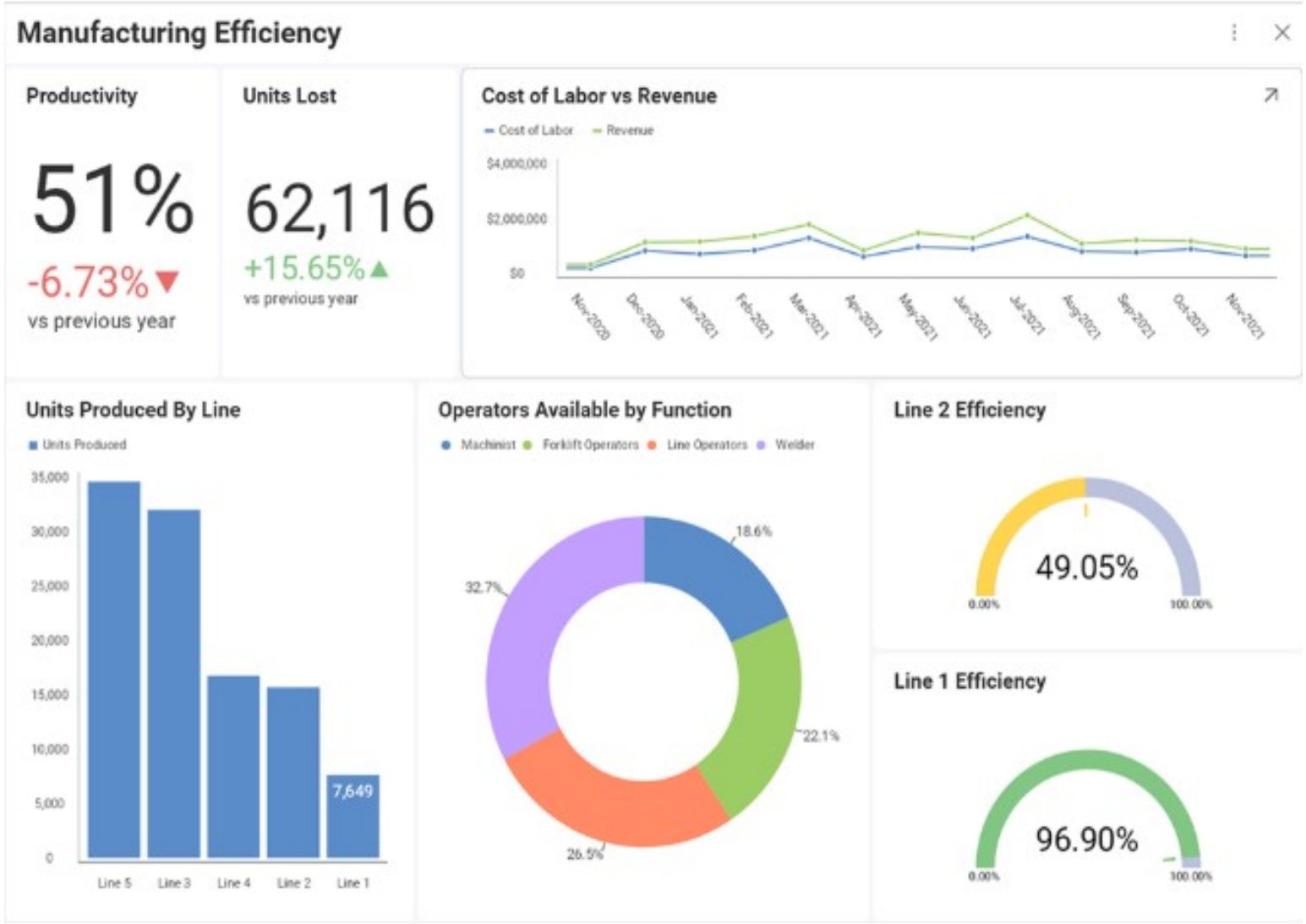


Types of data Analytics



Descriptive Analysis *(what happened)*

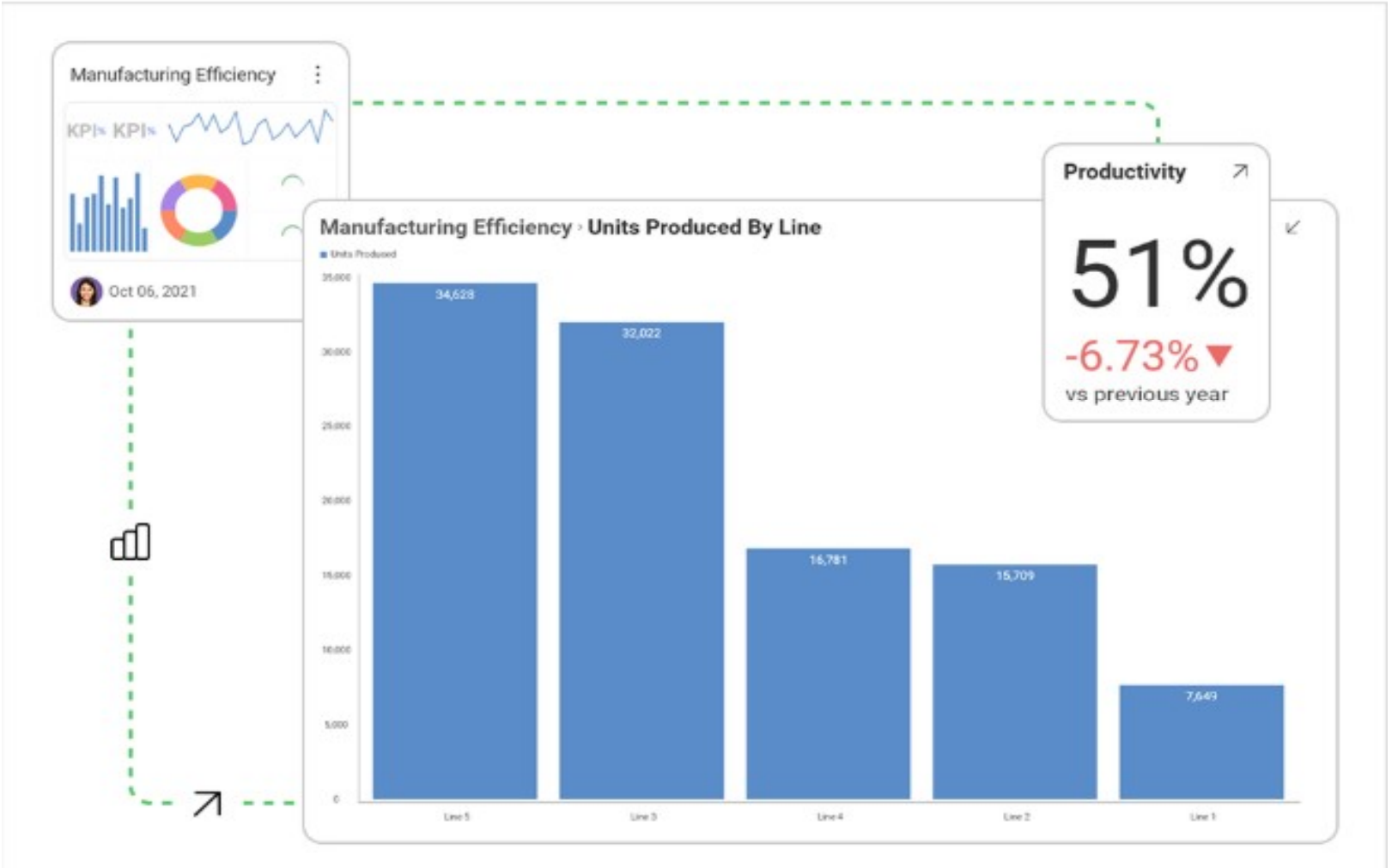
- The purpose of the descriptive type of data analysis is to answer the question of what happened.
- ***it's simply aimed at providing an easily digestible snapshot of what has happened in the past***.


Manufacturing Efficiency
Productivity
51%
-6.73%▼
vs previous year
Units Lost
62,116
+15.65%▲
vs previous year
Cost of Labor vs Revenue
Cost of Labor
Revenue
$4,000,000
$2,000,000
$0
Nov-2020
Dec-2020
Jan-2021
Feb-2021
Mar-2021
Apr-2021
May-2021
Jun-2021
Jul-2021
Aug-2021
Sep-2021
Oct-2021
Nov-2021
Units Produced By Line
Units Produced
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
7,649
Line 5
Line 3
Line 4
Line 2
Line 1
Operators Available by Function
Machinist
Forklift Operators
Line Operators
Welder
18.6%
22.1%
26.5%
32.7%
Line 2 Efficiency
49.05%
0.00%
100.00%
Line 1 Efficiency
96.90%
0.00%
100.00%

Diagnostic Analysis *(why did it happend)*

- Diagnostic analysis is one of the most powerful types of data analysis.
- ***gain a contextual understanding of why something has happened***.

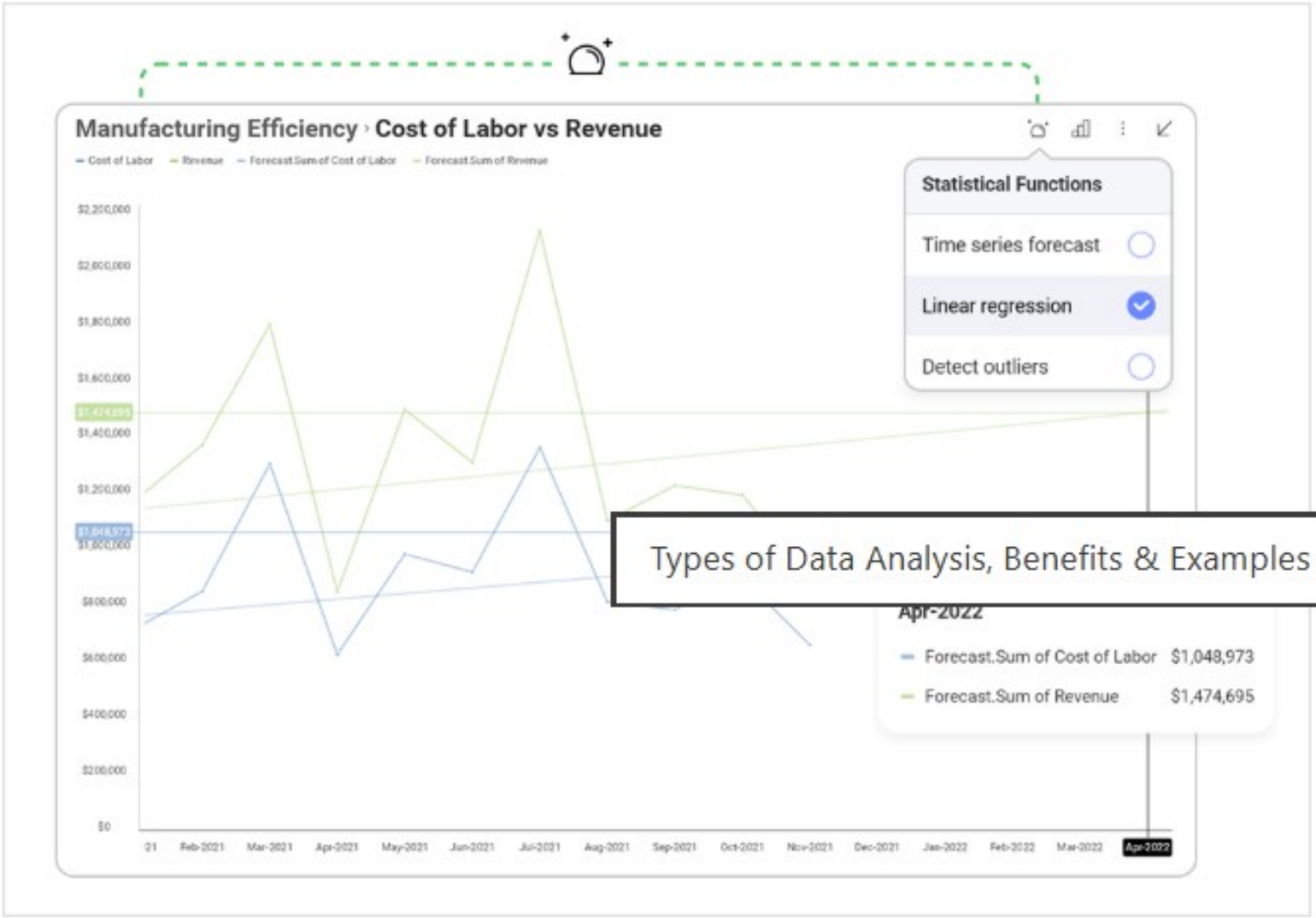
Diagnostic Analysis *(why did it happened)*



Predictive Analysis *(what is likely to happen)*

- As the name suggests, the predictive analysis predicts what is likely to happen in the future.
- ***estimates the likelihood of future events and outcomes***.

Predictive Analysis *(what is likely to happen)*



Prescriptive Analysis *(how will it happen)*

- The final and most advanced level type of data analysis is prescriptive. The prescriptive analysis combines the insights from all previous analysis in order to determine what should be done next. It shows you how you can best take advantage of the future outcomes that have been predicted and answer the question of how you will make it happen.
- The prescriptive analysis uses the full spectrum of complex **data science techniques, including advanced analytics**, and is the most difficult analysis to perform.

Data Analysis Examples

All kinds of businesses in every type of industry can harness the power of data analysis. These are some real-life examples of how 9 different industries are putting data to work:

Weather forecast – accumulates data about the current state of atmospheric conditions, particularly with temperature, humidity, and wind, and through the atmospheric process, data analysts determine what weather to expect.

Hotels – try to predict the number of guests for any given night to maximize occupancy and increase revenue.

Airlines – use predictive analytics to set ticket prices and determine aircraft maintenance requirements.

Data Analysis Examples

Marketing – marketing teams are using **data analysis to run targeted marketing campaigns** by segmenting audiences.

Finance – financial institutions can develop credit risk models, forecast financial market trends, and predict the impact of new policies, laws, and regulations on businesses and markets.

Manufacturing – use predictive analytics to monitor supplier performance, predict maintenance requirements and optimize production capacity.

Healthcare – hospitals, doctors, and other healthcare facilities can track the treatment of patients and determine patients who are at risk of developing diseases.

Security – protects businesses and individuals from hackers and cybercriminals.

4.3 Different forms of Data:

d. Statistics Data

- This term appears to refer to data that is relevant to statistical analysis.
- It can include datasets, variables, or information used in statistical studies and calculations.
- It’s not a distinct category of data but rather data that’s used within the field of statistics

4.3 Different forms of Data:

d. Statistics Vs. Data Mining Vs. Data Analytics

- There are comparisons or distinctions between different approaches to working with data.
- 1. Statistics Vs. Data Mining:
  - Statistics focuses on summarizing and drawing inferences from data while data mining focuses on discovering patterns and relationships within data. They serve different purposes but can complement each other in data analysis.
- 2. Data Analytics Vs. Data Science:
  - Data analytics is a broader term that encompasses data mining. Data mining is a subset of data analytics that specifically focuses on discovering patterns and knowledge from data.

4.3 Different forms of Data:

f. Data Analytics Vs. Data Science

- Data analytics is a component of data science.
- Data science is a multidisciplinary field that includes data analytics but also involves other areas such as machine learning, big data technologies, and domain expertise.
- Data science aims to extract valuable insights and knowledge from data to solve complex problems and make data-driven decision making.

4.4 Dataset for ML

Definition

- A dataset is a collection of data in which data is arranged in some order. A dataset can contain any data from a series of an array to a database table.

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

Types of Data

Data in datasets can be categorized into various types:

- 1. Numerical Data: Examples includes variables like house prices, temperatures, and numerical measurements that can take on a continuous range of values.
- 2. Categorical Data: This type comprises variables with distinct categories or labels, such as binary choices like Yes/No. True/False, or non-binary categories like colors (Blue/Green). These categories are typically not ordered

Types of Data

Data in datasets can be categorized into various types:

- 3. Ordinal Data: Similar to categorical data, ordinal data consists of categories, but these **categories have a meaningful order or ranking.** They can be measured based on comparison. Examples might include education levels (for example. High School, Bachelor's, Master's) or customer satisfaction ratings (for example, Very Dissatisfied. Dissatisfied, Neutral, Satisfied, Very Satisfied).
- **Note:** Real-world datasets can often be extensive and challenging to manage and process initially. To practice machine learning algorithms or conduct data analysis, using a simplified or dummy dataset is common and useful for learning and development.

Need for dataset

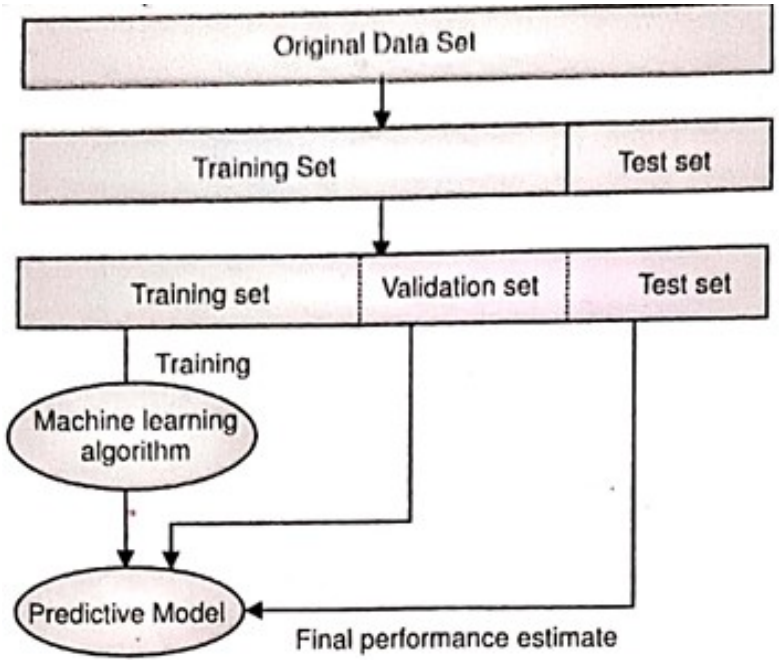
- **1. Training machine learning models:** Datasets are essential for training machine learning and artificial intelligence models. These models learn patterns and make predictions based on the data they are trained on. Without sufficient data, models may not generalize well to new, unseen examples
- **2. Data Collection and Preparation:** Collecting and preparing the dataset is a critical step in machine learning project development. High-quality, **well-organized data is necessary to ensure the accuracy and reliability of the models**. This process involves cleaning, formatting and sometimes augmenting the data to make it suitable for analysis and training.

Need for dataset

- **3. Model Performance:** Model Performance: The quality of the dataset directly impacts the performance of machine learning models. If the dataset is noisy. incomplete, or biased it can lead to inaccurate predictions and unreliable results. Proper data pre-processing and cleaning are essential to address these issues.
- **4. Model Training:** During the Development of a machine learning project, developers rely on datasets lo train and fine-tune models. The training dataset is used to teach the model patterns and relationships within the data, enabling it lo make predictions

Need for dataset

- **5. Evaluation and Testing:**
- Evaluation and Testing: Datasets are divided into two main parts: the training dataset and the testing dataset. The training dataset is used to train the model, while the test dataset is used to evaluate its performance. This separations helps assess how well the model generalizes to new, unseen data.



Data Cleaning: Missing Data, Outliers

- Data cleaning is an essential step in the data preparation process to ensure that your dataset is accurate and reliable. Two common aspects of data cleaning are handling missing data and identifying and dealing with outliers.
- **Missing data** is data that is not present in a dataset. This can happen for a variety of reasons, such as data that was not collected, data that was lost, Or data that was not entered into the dataset.
- **Outliers** are data points that are far from the rest of the data, They can be caused by errors in data collection, incorrect data entry, or natural variations in the data.
- **Missing data and outliers** can both impact the accuracy of machine learning models Missing data can cause a model to be less accurate because it has fewer data to learn from. Outliers can cause a model to be less accurate because they can skew the data and cause the model to learn from incorrect data.

Data Cleaning: Missing Data, Outliers

- **Missing Data:**
  - **Identification:** The first step is identifying missing values in your dataset. Missing data can occur for various reasons, such as data entry errors. equipment malfunctions, or survey non-responses .
  - **Handling Strategies:** There are several approaches to handling missing data.
  - **Imputation:** Replace missing values with estimated or calculated values. This can be done using mean, median, mode. or more advanced methods like regression imputation. **Deletion:** Remove rows or columns with missing data. This should be done carefully, as it can lead to information loss and bias if not handled appropriately.
  - **Data Collection:** In some cases, you might need lo collect missing data through additional surveys. experiments, or data sources.
  - **Advanced Techniques:** Depending on the context, more advanced techniques like predictive modeling can be used to impute missing values.

Data Cleaning: Missing Data, Outliers

- **Outliers:**
  - **Identification:** Outliers are data points that deviate significantly from majority of the data. They can skew statistical analysis and model results.
  - **Handling Strategies:** Outliers can be dealt in various ways.
  - **Detection:** Use statistical methods like z-scores, box plots, or scatter plots to identify outliers.
  - **Transformation:** Apply mathematical transformations to the data to reduce the impact of outliers, such as log transformations.
  - **Winsorization:** Cap extreme values by replacing them with values at a specified percentile (for example, replacing values above the 95th percentile with the 95th percentile value).
  - **Removal:** In some cases, Outliers may be influential or erroneous and should be removed from the dataset. However, be cautious when removing outliers and consider the potential impact on your analysis.
  - **Model-Based Approaches:** Some machine learning algorithms are robust to outliers. Using such models can mitigate the impact of outliers on predictions.

Thank You