

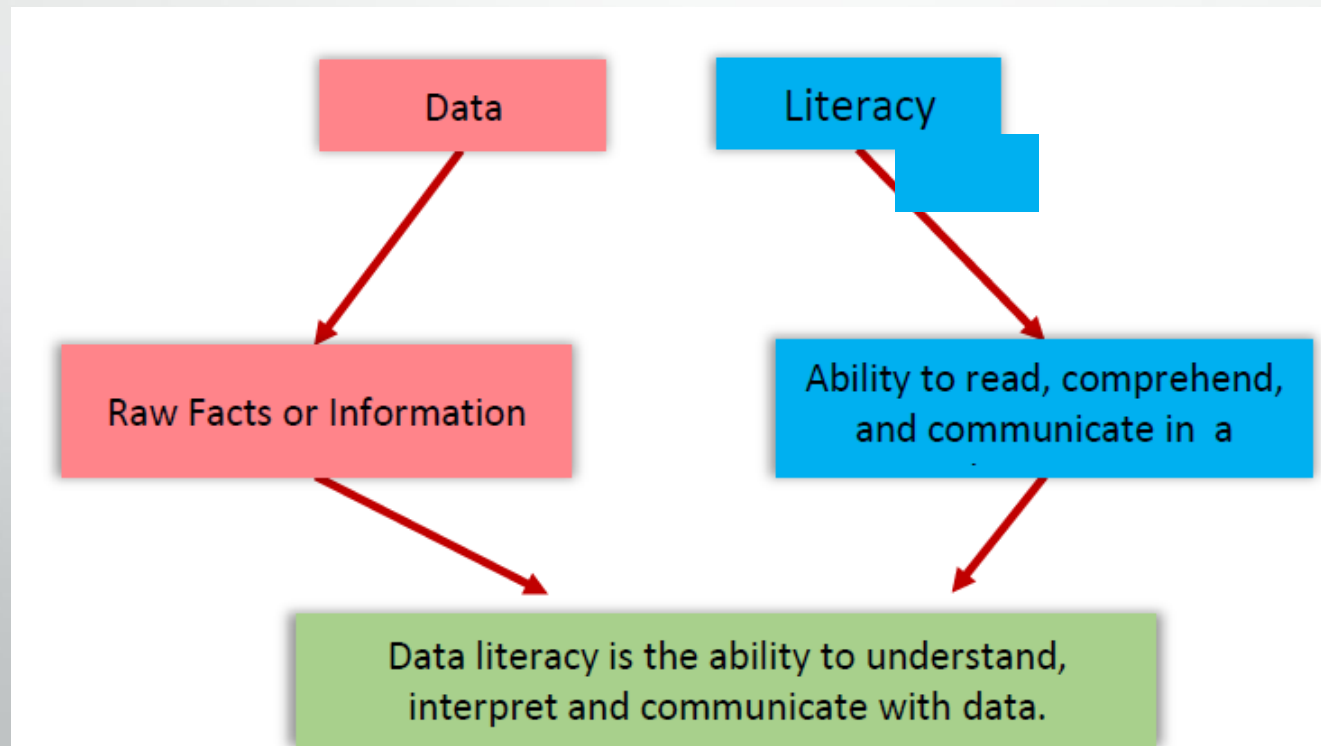


ARTIFICIAL INTELLIGENCE (SUBJECT CODE 417)

Unit 2 - Data Literacy

INTRODUCTION

- Data literacy means knowing how to understand, work with, and talk about data.
- It's about being able to collect, analyze, and show data in ways that make sense.

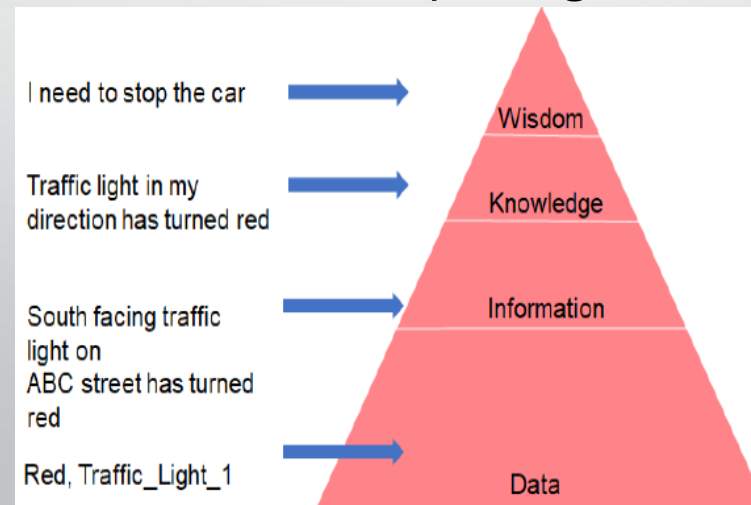
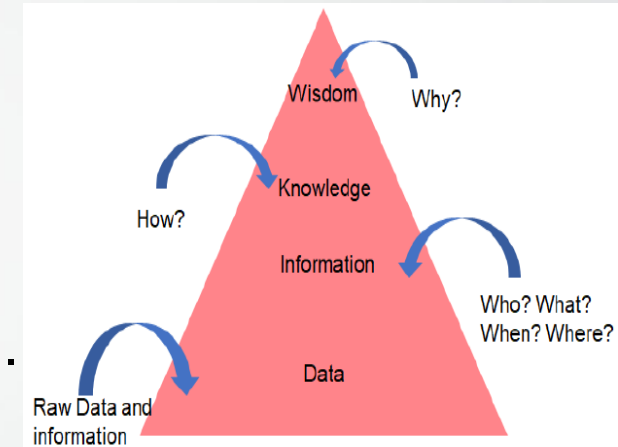


DATA PYRAMID

- Data Pyramid is made of different stages of working with data.

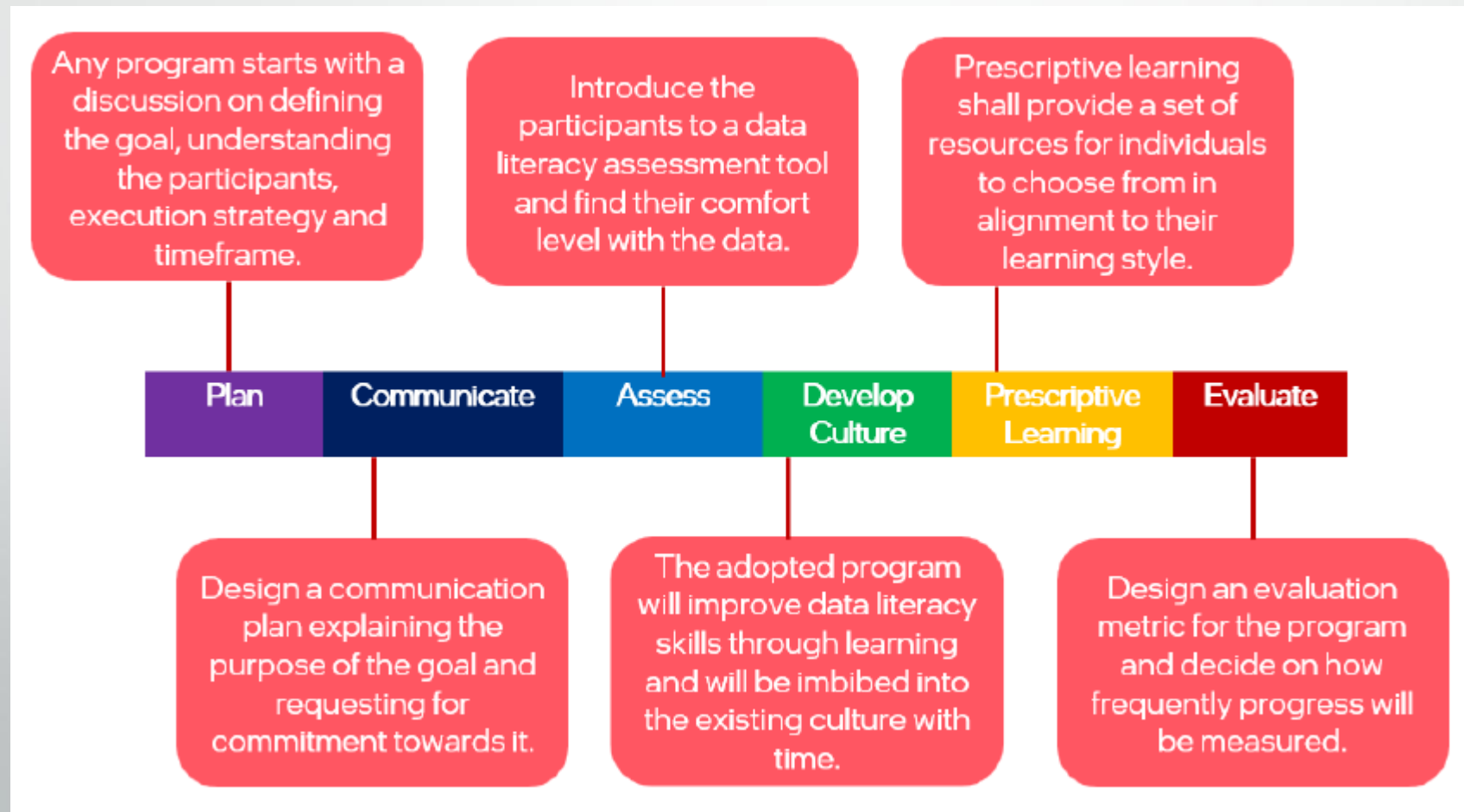
Moving up from the bottom

- Data is available in a raw form. Data in this form is not very useful.
- Data is processed to give us information about the world.
- Information about the world leads to knowledge of how things are happening.
- Wisdom allows us to understand why things are happening in a particular way.



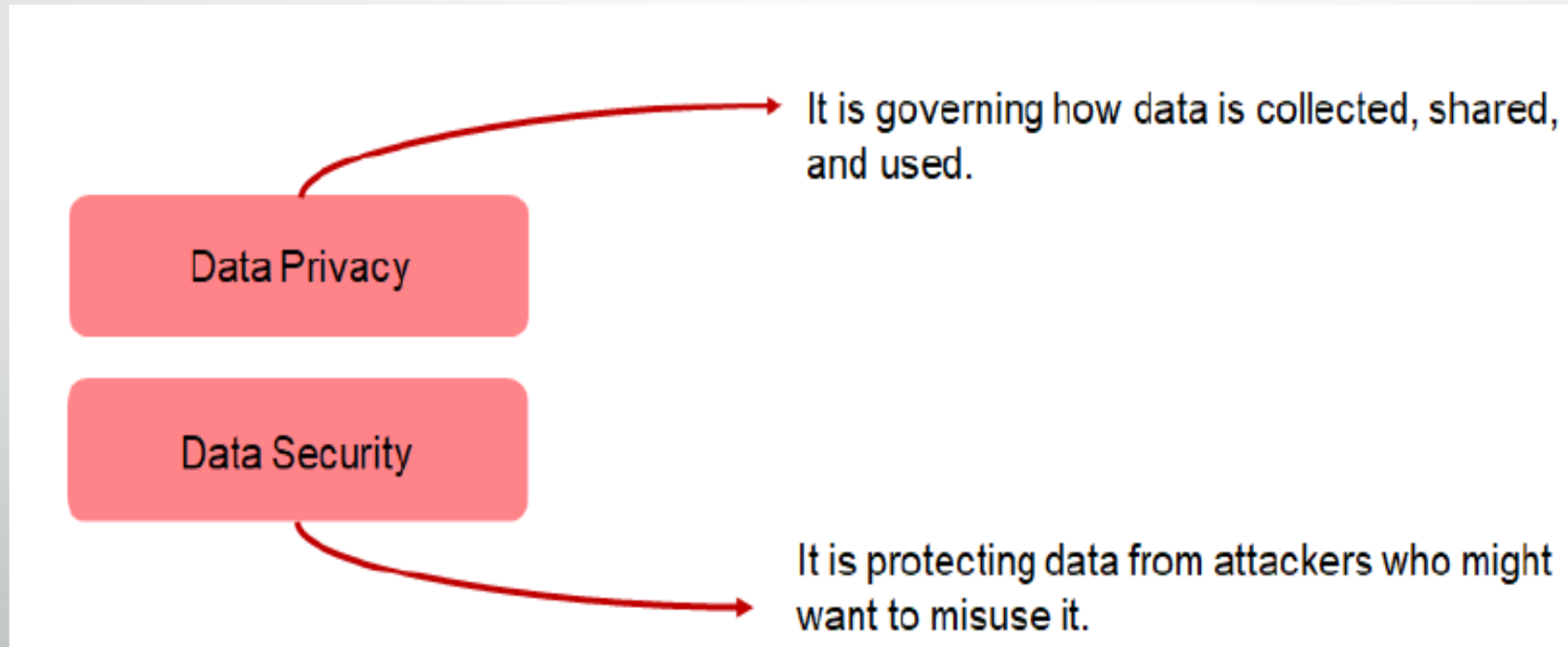
How to become Data Literate?

- **Data Literacy Process Framework** - The data literacy framework provides guidance on using data efficiently and with all levels of awareness. Data literacy framework is an iterative process.



What are Data Security and Privacy?

How are they related to AI?



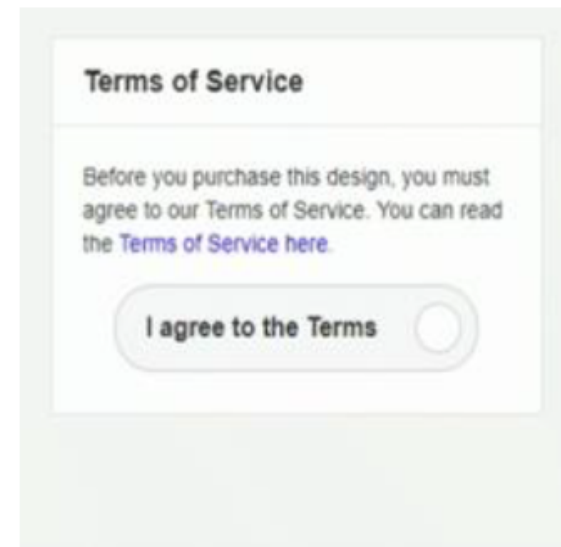
DATA PRIVACY

- Data privacy referred to as information privacy is concerned with the proper handling of sensitive data including personal data and other confidential data, such as certain financial data and intellectual property data, to meet regulatory requirements as well as protecting the confidentiality and immutability of the data.

Here are examples of two things which may compromise our data privacy



Downloaded an unverified mobile application



Accepted the Terms of Service without reading

Why is it important?



The following best practices can help you ensure data privacy

- Understanding what data, you have collected, how it is handled, and where it is stored.
- Necessary data required for a project should only be collected.
- User consent while data collection must be of utmost importance.

DATA SECURITY

- **Data security** is the practice of protecting digital information from unauthorized access, corruption, or theft throughout its entire lifecycle.

Why is it important?

- Due to the rising amount of data in the cloud there is an increased risk of cyber threats. The most appropriate step for such an amount of traffic being generated is how we control and protect the transfer of sensitive or personal information at every known place.
- The most possible reasons why data security is more important now are:
- Cyber-attacks affect all the people
- The fast-technological changes will boom cyber attacks

Best Practices for Cyber Security

- Cyber security involves protecting computers, servers, mobile devices, electronic systems, networks, and data from harmful attacks.

Do's

- Use strong, unique passwords with a mix of characters for each account.
- Activate Two-Factor Authentication (2FA) for added security.
- Download software from trusted sources and scan files before opening.
- Prioritize websites with "https://" for secure logins.
- Keep your browser, OS, and antivirus updated regularly.
- Adjust social media privacy settings for limited visibility to close contacts.
- Always lock your screen when away.
- Connect only with trusted individuals online.
- Use secure Wi-Fi networks.
- Report online bullying to a trusted adult immediately.

Best Practices for Cyber Security

Don't's

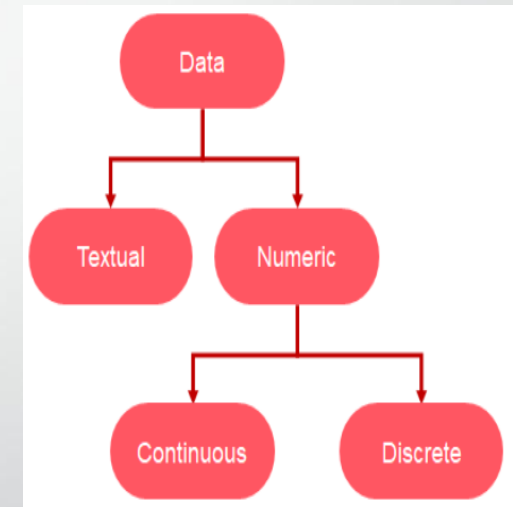
- Avoid sharing personal info like real name or phone number.
- Don't send pictures to strangers or post them on social media.
- Don't open emails or attachments from unknown sources.
- Ignore suspicious requests for personal info like bank account details.
- Keep passwords and security questions private.
- Don't copy copyrighted software without permission.
- Avoid cyberbullying or using offensive language online.

Acquiring Data, Processing, and Interpreting Data

- Types of data

Artificial Intelligence is crucial, with data serving as its foundation. We come across different types of information every day. Some common types of data include:

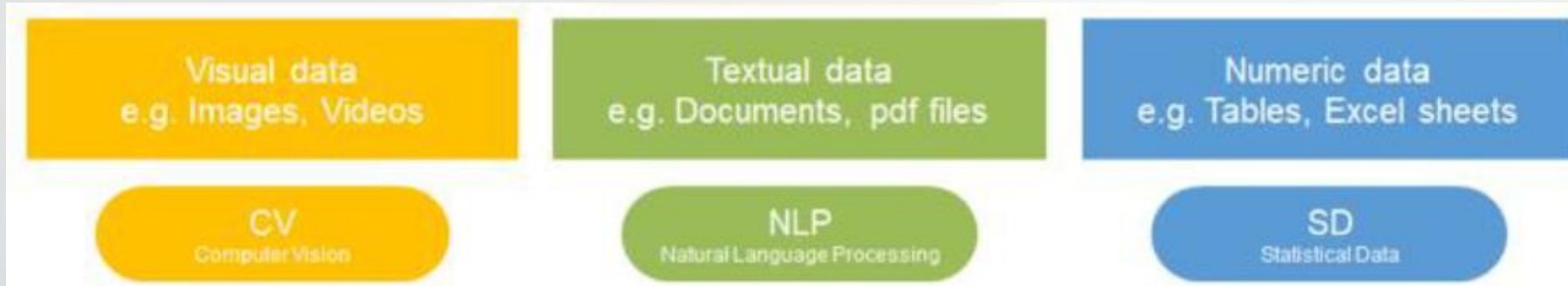
Textual Data (Qualitative Data)	Numeric Data (Quantitative Data)
• It is made up of words and phrases • It is used for Natural Language Processing (NLP) • Search queries on the internet are an example of textual data • Example: "Which is a good park nearby?"	• It is made up of numbers • It is used for Statistical Data • Any measurements, readings, or values would count as numeric data • Example: Cricket Score, Restaurant Bill



Numeric Data is further classified as:

1. **Continuous data** is numeric data that is continuous. E.g., height, weight, temperature, voltage
2. **Discrete data** is numeric data that contains only whole numbers and cannot be fractional E.g. the number of students in the class – it can only be a whole number, not in decimals

Types of Data used in three domains of AI:



Pick and Choose (Quantitative or Qualitative?)



Temperature



Gender



Shoe Size



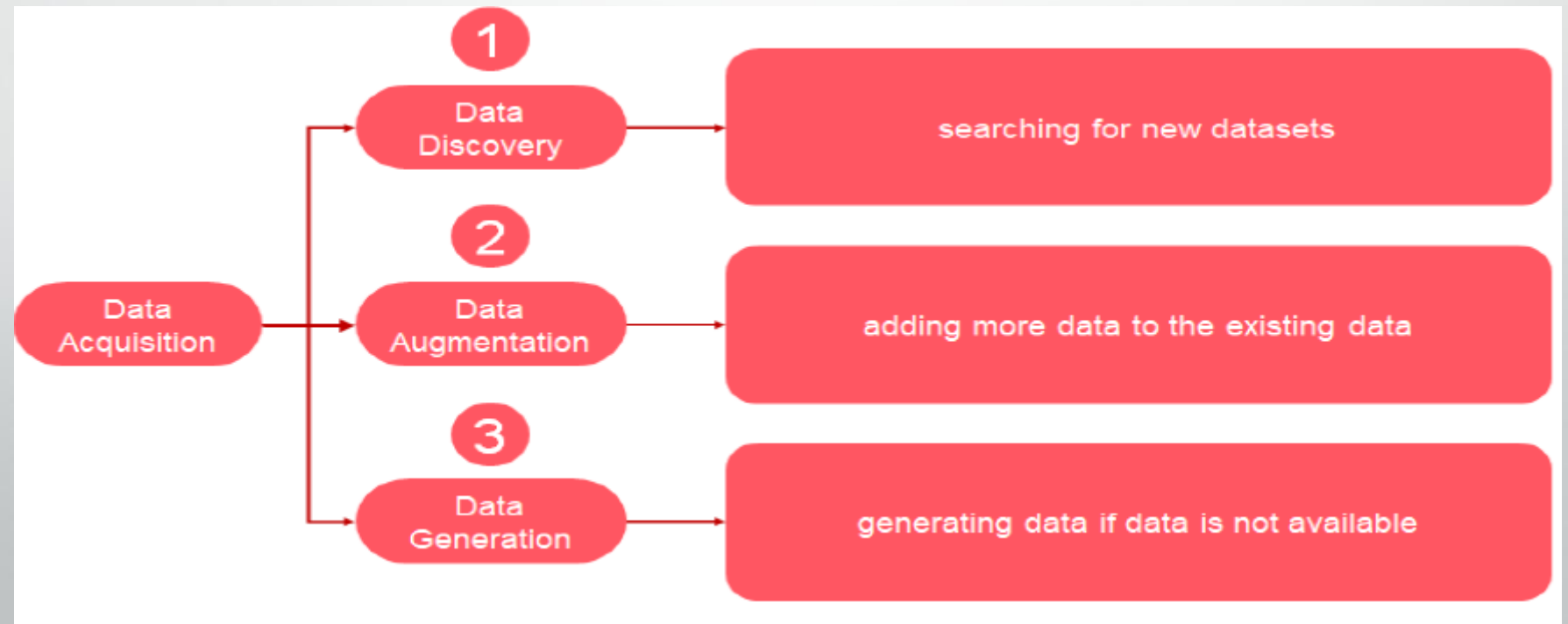
Favorite
Color



Weight of a
Person

Data Acquisition/Acquiring Data

- Data Acquisition, also known as acquiring data, refers to the procedure of gathering data. This involves searching for datasets suitable for training AI models. The process typically comprises three key steps:



1.Acquiring Data – Sample Data Discovery

- Let's say we want to collect data for making a CV model for a self-driving car
- We will require pictures of roads and the objects on roads
- We can search and download this data from the internet
- This process is called **data discovery**



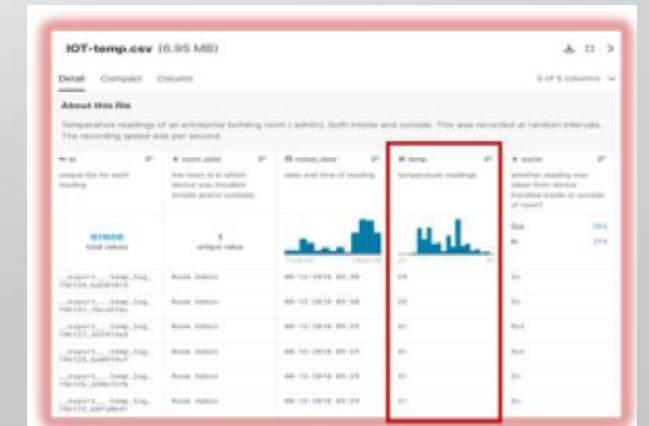
2. Acquiring Data – Sample Data Augmentation

- Data augmentation means increasing the amount of data by adding copies of existing data with small changes
- The image given here does not change, but we get data on the image by changing different parameters like color and brightness
- New data is added by slightly changing the existing data



3. Acquiring Data – Sample Data Generation

- Data generation refers to generating or recording data using sensors
- Recording temperature readings of a building is an example of data generation
- Recorded data is stored in a computer in a suitable form



Sources of Data

- Various Sources for Acquiring Data:
- **Primary Data Sources** — Some of the sources for primary data include surveys, interviews, experiments, etc. The data generated from the experiment is an example of primary data.
- Here is an excel sheet showing the data collected for students of a class.

Name	Height	Weight	Age	Residence	Favourite Hobby
John Doe	5' 3"	56kg	13	123 Main Street, New York, NY 10030	Football
[Student 2]					
[Student 3]					
[Student 4]					
[Student 5]					
[...]					

- **Secondary Data Sources**—Secondary data collection obtains information from external sources, rather than generating it personally. Some sources for secondary data collection include:



Best Practices for Acquiring Data

Checklist of factors that make data good or bad

Good Data

- ☐ Information is well structured
- ☐ It is accurate
- ☐ It is consistent
- ☐ It is cleanly presented
- ☐ Contains information which is relevant to our requirement

Bad Data

- ☐ Information is scattered
- ☐ Contains a lot of incorrect values
- ☐ Contains missing and duplicate values
- ☐ It is poorly presented
- ☐ Contains information which is not relevant to our requirement

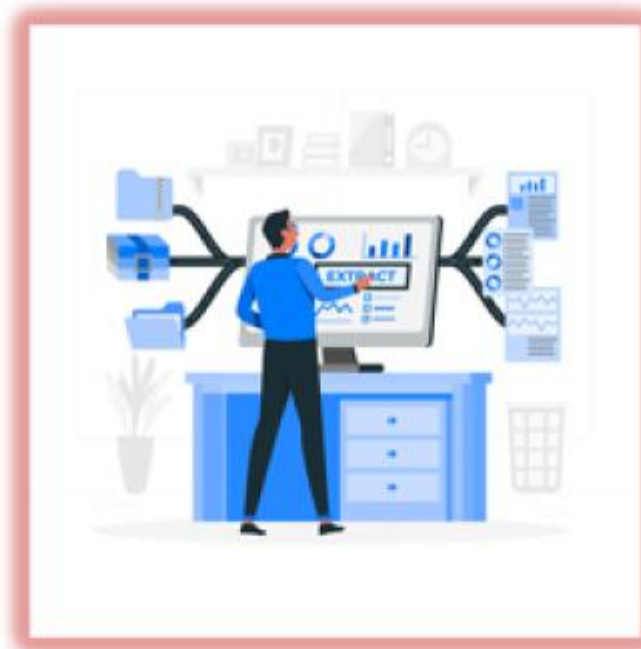
DATA ACQUISITION FROM WEBSITES

1

The process of **collecting data** from websites using software is called **Web Scraping**

3

While web scraping is not illegal, using data **without permission** is illegal



2

There are different **tools** that can help us collect data from websites

4

During data acquisition, we need to make sure that the data source **allows** data scraping

ETHICAL CONCERNS IN DATA ACQUISITION

Bias	Take steps to understand and avoid any preferences or partiality in data
Consent	Take necessary permissions before collecting or using an individual's data
Transparency	Explain how you intend to use the collected data and do not hide intentions
Anonymity	Protect the identity of the person who is the source of data
Accountability	Take responsibility for your actions in case of misuse of data

FEATURES OF DATA AND DATA PREPROCESSING

Usability of Data : There are three primary factors determining the usability of data:

1. Structure- Defines how data is stored.

Purchase ID	Last name	First name	Birthday	Country
1	Davidson	Michael	04/03/1986	United States
2	Vito	Jim	09/01/1994	United Kingdom
3	Johnson	Tom	23/08/1972	France
4	Lewis	Peter	18/10/1979	Germany
5	Koenig	Edward	13/05/1983	Argentina
6	Preston	Jack	16/06/1991	United States
7	Smith	David	11/03/1965	Canada
8	Brown	Luis	03/09/1997	Australia
9	Miller	Thomas	07/01/1980	Germany

Spreadsheet – Good structure

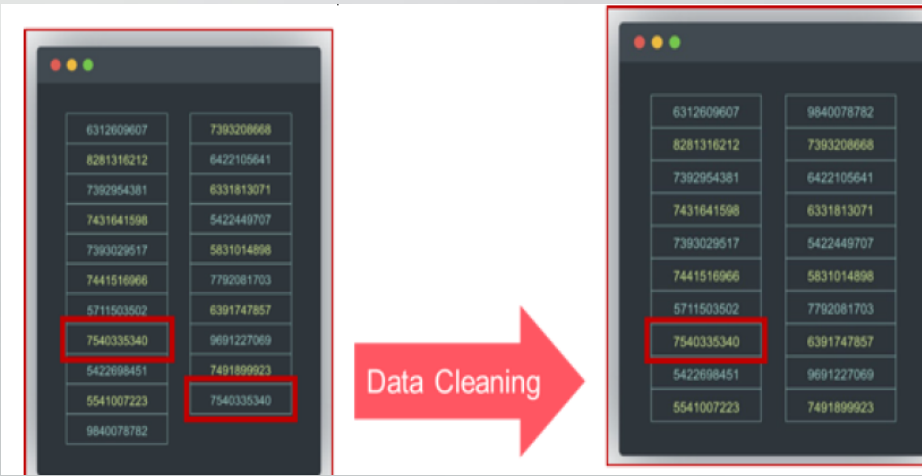
Data is stored in a sheet with the details of each individual stored according to a set of rules.

Micheal Davison lives in United States. He was born on 04/03/1986. Jim Vito lives in United Kingdom. He was born on 09/01/1994. Tom Johnson lives in France. He was born on 23/08/1972.

Text document – Poor structure

Data is stored in a text document with no set of organizing rules.

2. Cleanliness- Clean data is free from duplicates, missing values, outliers, and other anomalies that may affect its reliability and usefulness for analysis. In this particular example, duplicate values are removed after cleaning the data.



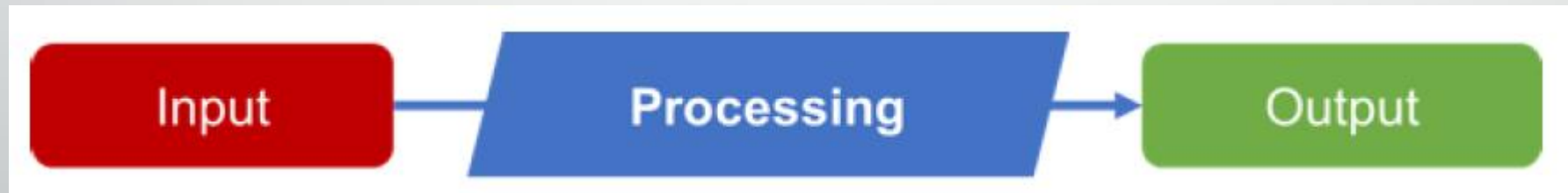
3. Accuracy- Accuracy indicates how well the data matches real-world values, ensuring reliability. Accurate data closely reflects actual values without errors, enhancing the quality and trustworthiness of the dataset. In this particular example, we are comparing data gathered from measuring the length of a small box in centimeters.



Kaggle assigns a usability score to the data sets that are present on the website based on scores given by the users of that data.

FEATURES OF DATA

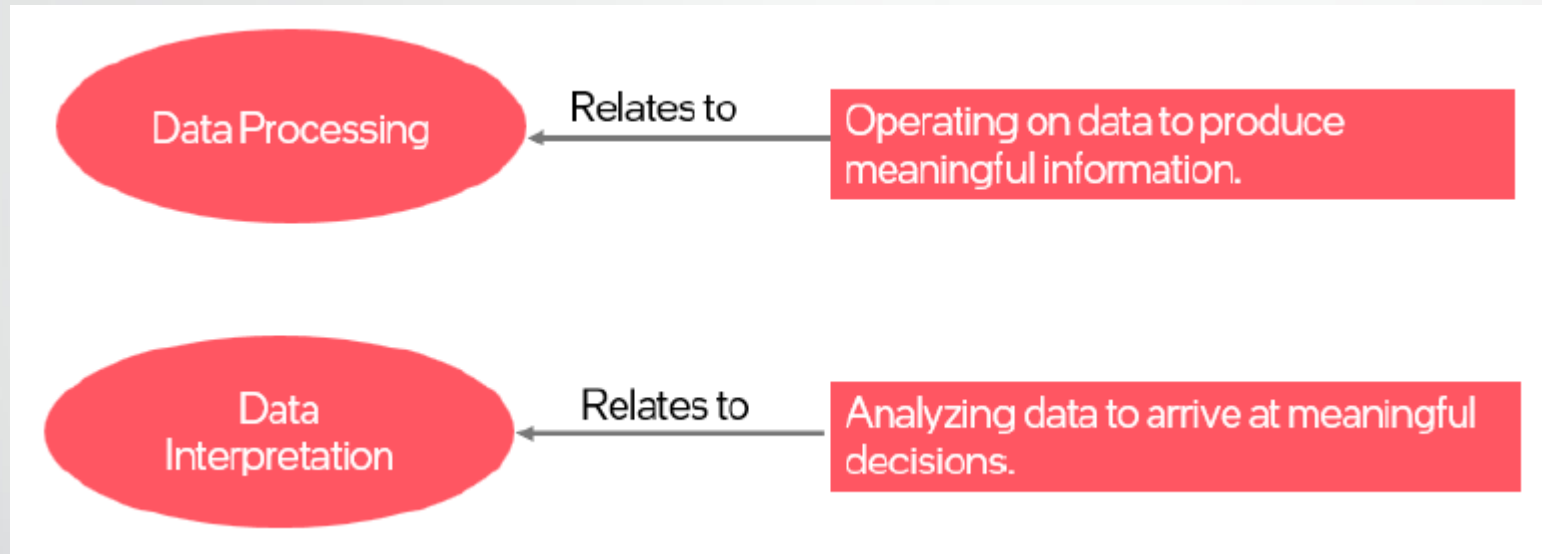
- Data features are the characteristics or properties of the data. They describe each piece of information in a dataset. For example, in a table of student records, features could include things like the student's name, age, or grade. In a photo dataset, features might be the colors present in each image. These features help us understand and analyze the data.
- In AI models, we need two types of features: independent and dependent.



Independent features are the input to the model—they're the information we provide to make predictions.

Dependent features, on the other hand, are the outputs or results of the model—they're what we're trying to predict.

Data Processing and Data Interpretation



Data Processing

Data processing helps computers understand raw data.

Use of computers to perform different operations on data is included under data processing.

Data Interpretation

It is the process of making sense out of data that has been processed.

The interpretation of data helps us answer critical questions using data.

- **Acquire Data-** Acquiring data is to collect data from various data sources.
- **Data Processing-** After raw data is collected, data is processed to derive meaningful information from it.
- **Data Analysis** – Data analysis is to examine each component of the data in order to draw conclusions.
- **Data Interpretation** – It is to be able to explain what these findings/conclusions mean in a given context.
- **Data Presentation-** In this step, you select, organize, and group ideas and evidence in a logical way.



Methods of Data Interpretation

- Based on the two types of data, there are two ways to interpret data-
 - Quantitative Data Interpretation
 - Qualitative Data Interpretation
- **Qualitative Data Interpretation**
- Qualitative data tells us about the emotions and feelings of people
- Qualitative data interpretation is focused on insights and motivations of people
- **Data Collection Methods – Qualitative Data Interpretation**
- Record keeping: This method uses existing reliable documents and other similar sources of information as the data source. It is similar to going to a library.
- Observation: In this method, the participant – their behavior and emotions – are observed carefully
- Case Studies: In this method, data is collected from case studies.
- Focus groups: In this method, data is collected from a group discussion on relevant topic.
- Longitudinal Studies: This data collection method is performed on the same data source repeatedly over an extended period.
- One-to-One Interviews: In this method, data is collected using a one-to-one interview.

- **5 Steps to Qualitative Data Analysis**

- Collect Data
- Organize
- Set a code to the Data Collected
- Analyze your data
- Reporting

Quantitative Data Interpretation

- Quantitative data interpretation is made on numerical data
- It helps us answer questions like “when,” “how many,” and “how often”
- For example – (how many) numbers of likes on the Instagram post

Data Collection Methods -Quantitative Data Interpretation

- **Interviews:** Quantitative interviews play a key role in collecting information.
- **Polls:** A poll is a type of survey that asks simple questions to respondents. Polls are usually limited to one question.
- **Observations:** Quantitative data can be collected through observations in a particular time period
- **Longitudinal Studies:** A type of study conducted over a long time
- **Survey:** Surveys can be conducted for a large number of people to collect quantitative data.

4 Steps to Quantitative Data Analysis

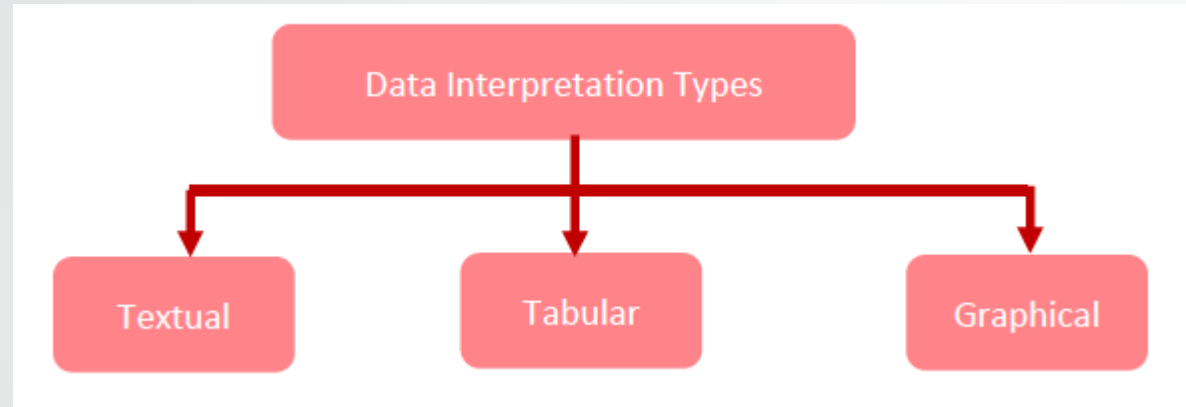
- Relate measurement scales with variables
- Connect descriptive statistics with data
- Decide a measurement scale
- Represent data in an appropriate format

Qualitative & Quantitative Data Interpretation

Qualitative Data Interpretation	Quantitative Data Interpretation
Categorical	Numerical
Provides insights into feelings and emotions	Provides insights into quantity
Answers how and why	Answers when, how many or how often
Methods – Interviews, Focus Groups	Methods – Assessment, Tests, Polls, Surveys
Example question – Why do students like attending online classes?	Example question – How many students like attending online classes?

Types of Data Interpretation

- There are three ways in which data can be presented:



Textual DI

- The data is mentioned in the text form, usually in a paragraph.
- Used when the data is not large and can be easily comprehended by reading.
- Textual presentation is not suitable for large data.

In the Science Olympiad class of 45 Students, 3 students obtained the perfect score of 50. 10 students got a score of 45 and above, 15 students got a score of 40 and above, 8 students got a score of 30 and above, 6 students got a score of 20 and above and 3 got 19 and below.



More than 60% of students scored more than 80% Marks in Olympiad!

Tabular DI

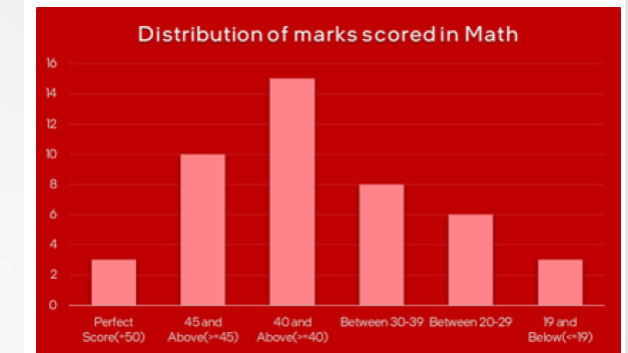
- Data is represented systematically in the form of rows and columns.
- Title of the Table (Item of Expenditure) contains the description of the table content.
- Column Headings (Year; Salary; Fuel and Transport; Bonus; Interest on Loans; Taxes) contains the description of information contained in columns.

Year	Item of Expenditure				
	Salary	Fuel and Transport	Bonus	Interest on Loans	Taxes
1998	288	98	3.00	23.4	83
1999	342	112	2.52	32.5	108
2000	324	101	3.84	41.6	74
2001	336	133	3.68	36.4	88
2002	420	142	3.96	49.4	98

Graphical DI

Bar Graphs

In a Bar Graph, data is represented using vertical and horizontal bars.

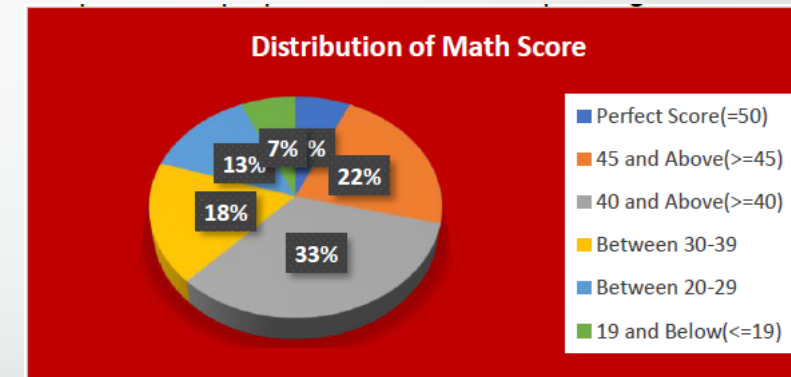


Pie Charts

Pie Charts have the shape of a pie and each slice of the pie represents the portion of the entire pie allocated to each category

It is a circular chart divided into various sections (think of a cake cut into slices)

Each section of the pie chart is proportional to the corresponding value



Line Graphs

A line graph is created by connecting various data points. It shows the change in quantity over time.

