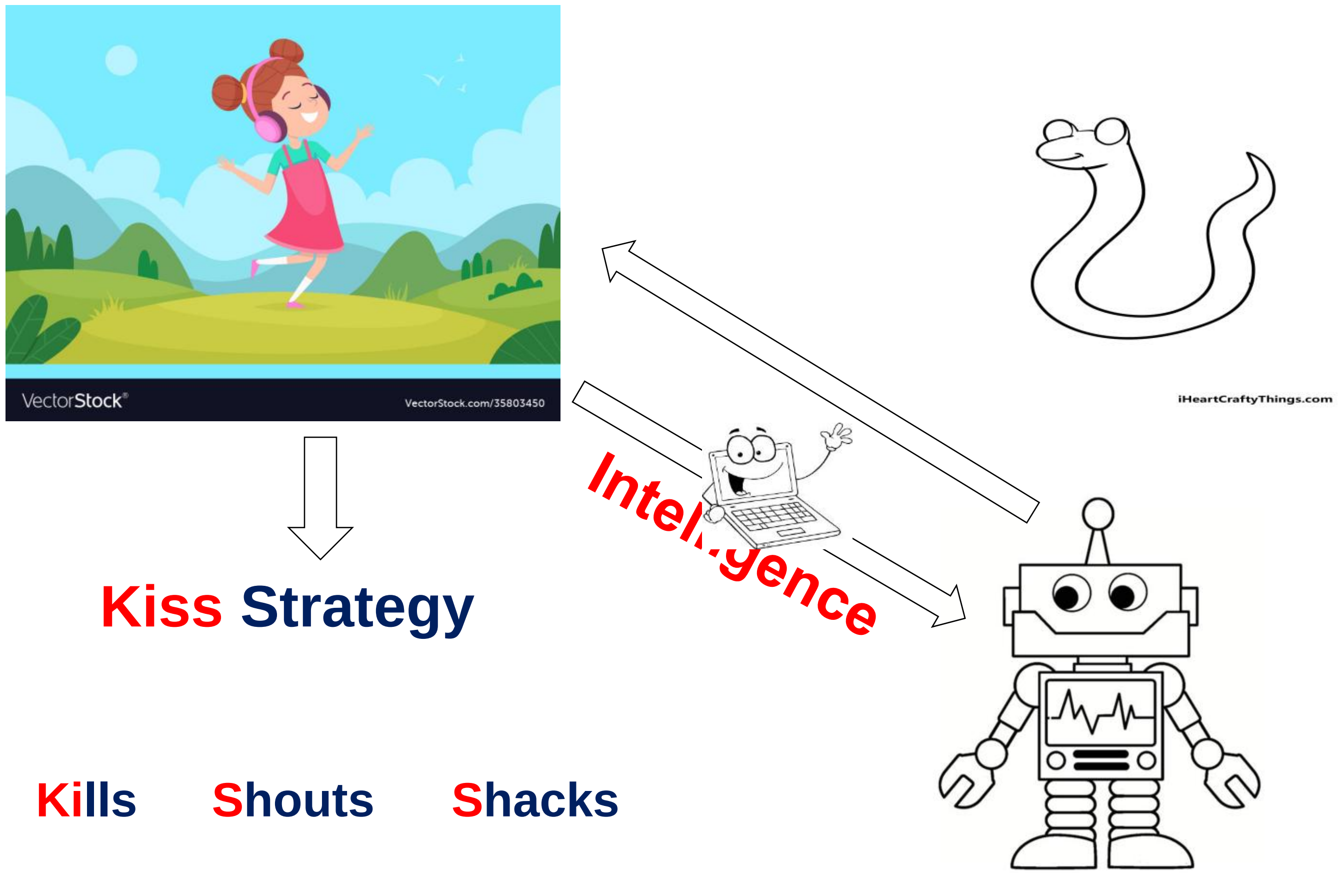
# Data Preprocessing

Dr. MALLIKHARJUNA RAO K
Assistant Professor
Data Science and Artificial Intelligence
IIIT Naya Raipur
mallikharjuna@iiitnr.edu.in

K
M
R

# ARTIFICIAL INTELLIGENCE



Intelligence

**Kiss Strategy**

**Kills** **Shouts** **Shacks**

Image source: Google images

# What is data?

- A **Raw fact**

- The data may be a
  - **Number**
  - **Text**
  - **Image**
  - **Audio**
  - **Video**

**⋮**

# Database Vs Data Warehouse Vs Data Mart

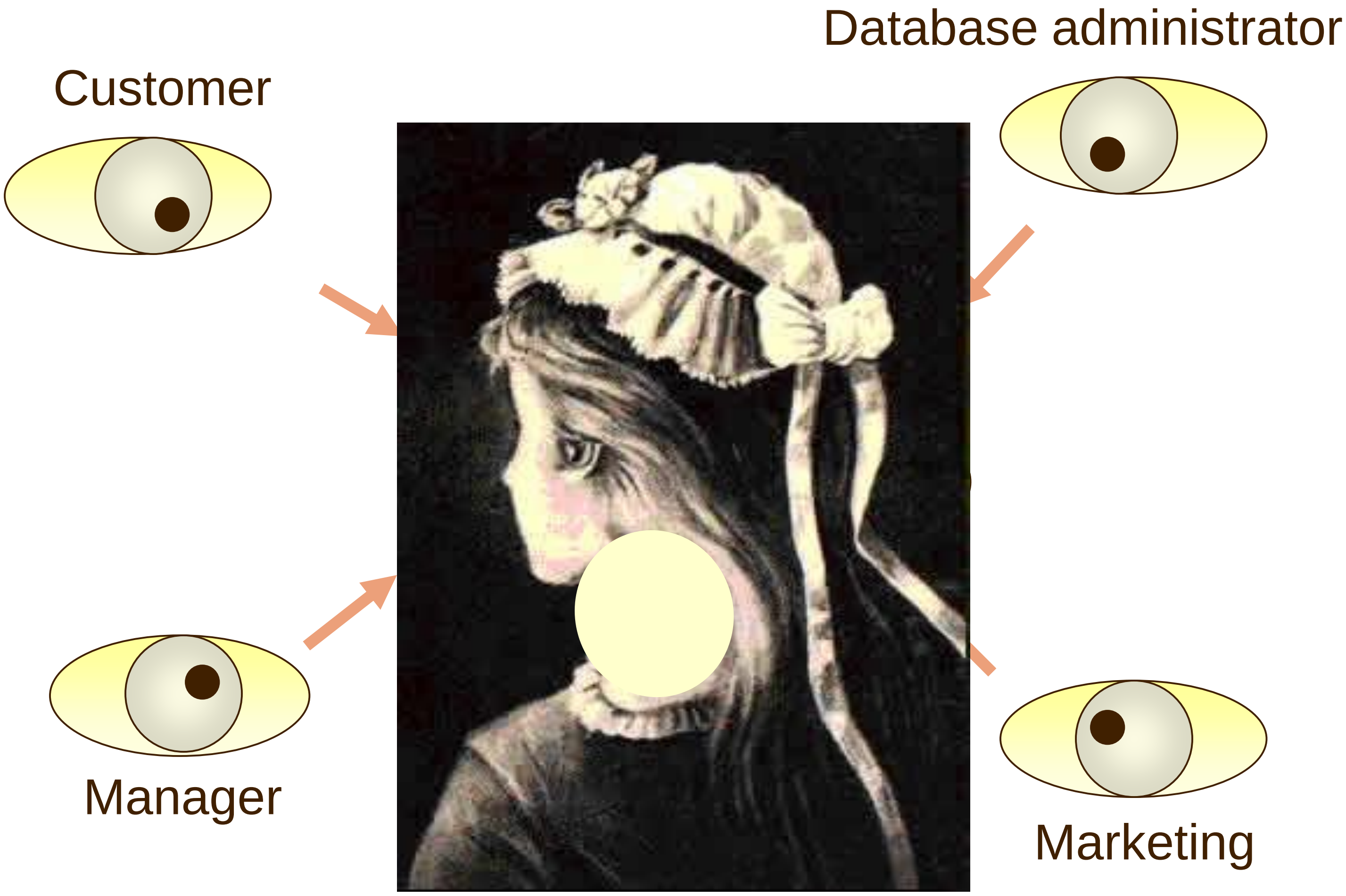
- **Data Base**
  - Detailed data
  - Utilizes current information
  - Information from one main source
  - Useful to perform day-to-day operations

- **Data Warehouse**
  - Summarized data
  - Historical & current information
  - Information from various sources
  - Useful to perform business operations

- **Data Mart**
  - Condensed Summarized data
  - Internal department-based/specific information
  - Useful to perform internal business operations

# What you can see?

## Multiple problem viewpoints



Database administrator

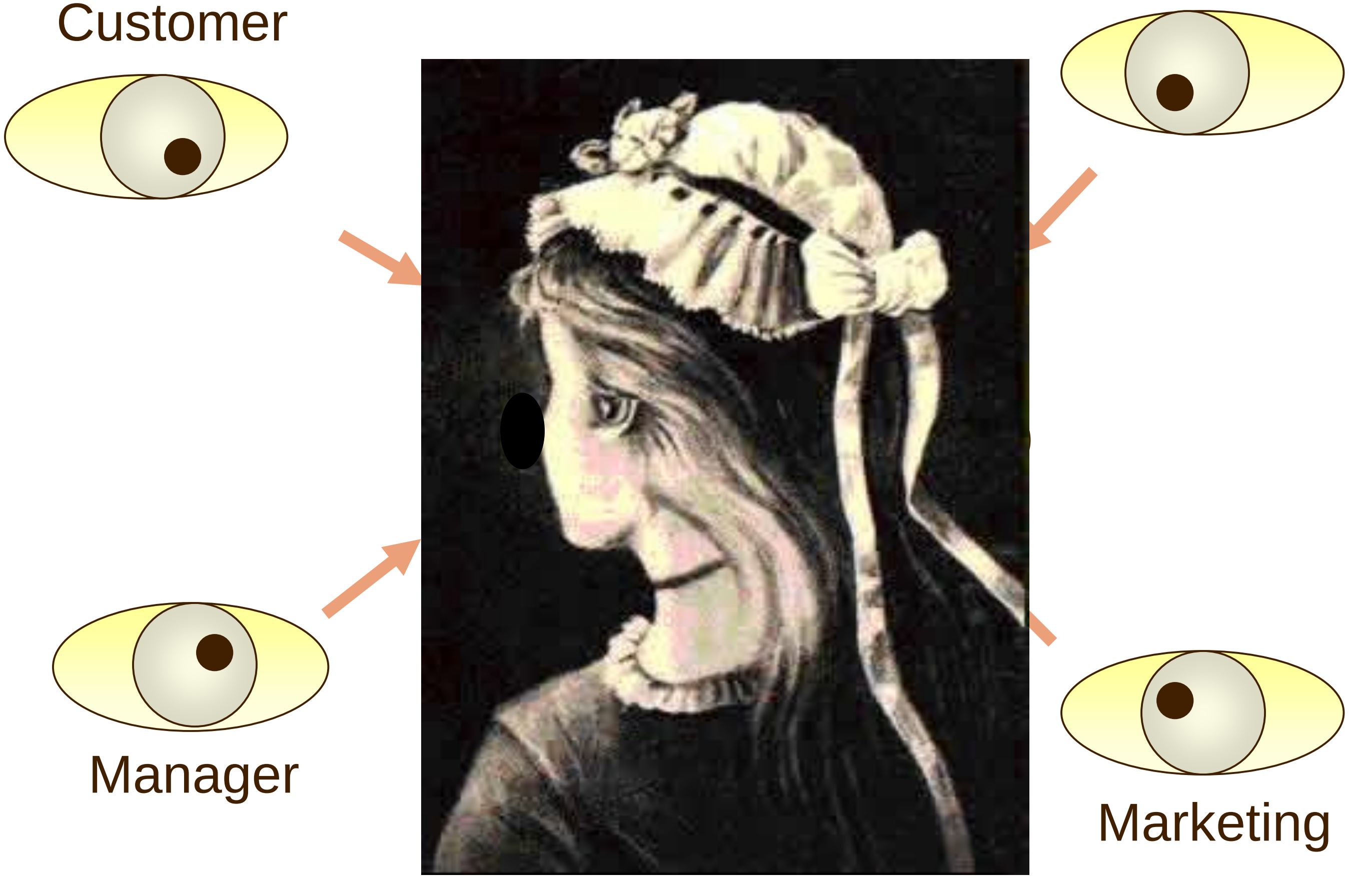
Customer

Manager

Marketing

# What you can see?

## Multiple problem viewpoints



Database administrator

Customer

Manager

Marketing

# What is Data Preprocessing?

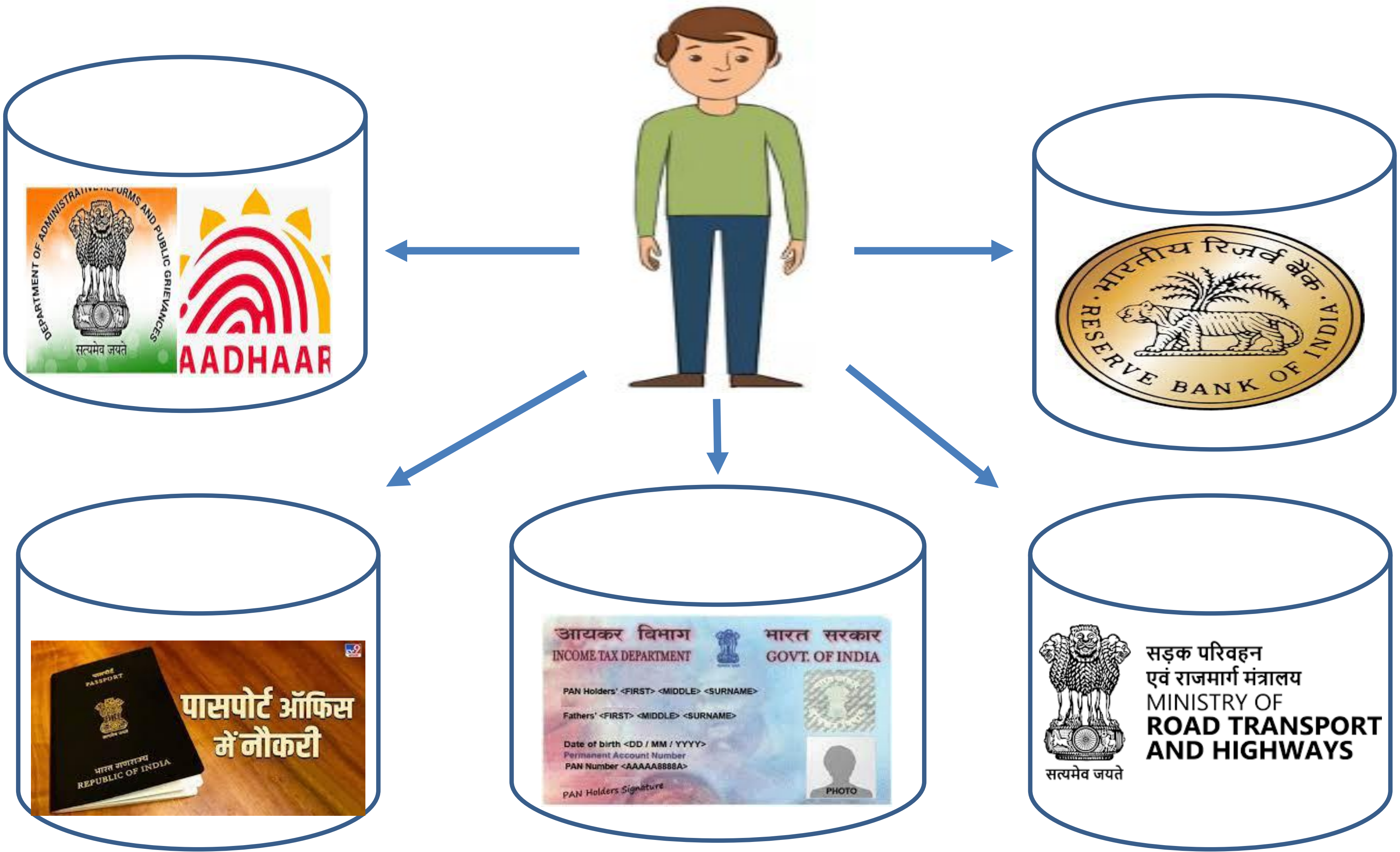
**Data Preprocessing** is the process of cleaning and engineering data in a way that it can be used as input to several important data science tasks such as data visualization, machine learning, deep learning, and data analytics.

# Contents

- **Why preprocess the data?**
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

# Why Preprocess the data?

PERSON



# Major Tasks in Data Preprocessing

1. Data Cleaning
2. Data Integration
3. Data Selection
4. Data Reduction
5. Data Mining

# DATA CLEANING

- DATA IS AVAILABLE IN THE REAL WORLD IN THE FORM OF **DIRTY.**
- **DIRTY MEANS**
  - **INCOMPLETE**
  - **NOISY**
  - **INCONSISTENT**

# INCOMPLETE DATA

- lacking certain attributes of interest
- containing only aggregate data
- “Not applicable” values
- Time gap
- Human/hardware/software problems

- **MISSING THE ATTRIBUTE VALUES WHICH ARE INTERESTED.**

Ex: I would like to know the **PASS GRADE** of a student. But **no Marks** are available.

# NOISY DATA

- Faulty data collection instruments
- Human or computer error at data entry
- Errors in data transmission

- CONTAINING **ERRORS** OR **OUTLIERS.**
- **Expecting the data in one format but in the real world existed in another format.**
  - Ex: Marks available in PERCENTAGE rather than CGPA

# INCONSISTENT DATA

- Different data sources
- Functional dependency violation

- **DISCREPANCIES IN CODES OR NAMES WHICH ARE INTERESTED.**
  - **Ex:**

    **1) Date Format**

    - **India: DD/MM/YYYY**
    - **Foreign Countries: MM/DD/YYYY**

    **2) Ranking**

    - **Earlier used 1, 2, 3,…**
    - **Now A, B, C, ….**

# DATA INTEGRATION

- DATA COLLECTED FROM **MULTIPLE SOURCES.**
- **TEMPERATURE PREDICTION**
  - **Fahrenheit (${}^{o}$F)**
  - **Celsius (${}^{o}$C)**
  - **Kelvin (K)**
- We have different formats to measure.
- Converted into **unique representation.**

# Data Selection

- **Selection of Appropriate Data/ Attributes to retrieve interested patterns.**
  - **Ex: I have sales Data Base. If I try to get MARKS VALUES. It doesn't Provide.**

# DATA REDUCTION

- **The large volumes of data is reduced into different smaller chunks.**
- **MULTIDIMENSIONAL DATA INTO**
  - **1D Data**
  - **2D Data**
  - **3D Data**

# DATA REDUCTION

- 1 year into Q1, Q2, Q3, Q4
- Q1 into M1, M2, M3
- Q2 into M4, M5, M6
- Q3 into M7, M8, M9
- Q4 into M10, M11, M12

into W1, W2, W3, W4

into D1, D2, ….D7

# Why Is Data Preprocessing Important?

- **No quality data, no quality mining results!**
  - Quality decisions must be based on quality data
    - Ex: duplicate or missing data may cause incorrect or even misleading statistics.
  - Data warehouse needs **consistent integration of quality data**
- Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse

# Multi-Dimensional Measure of Data Quality

- A well-accepted multidimensional view:
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Value added
  - Interpretability
  - Accessibility
- Broad categories:
  - Intrinsic, contextual, representational, and accessibility