SNS COLLEGE OF TECHNOLOGY

An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A' Grade

Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

Department of Computer Science and Engineering

Course Code & Title : 23ADO201 - Data Science Fundamentals

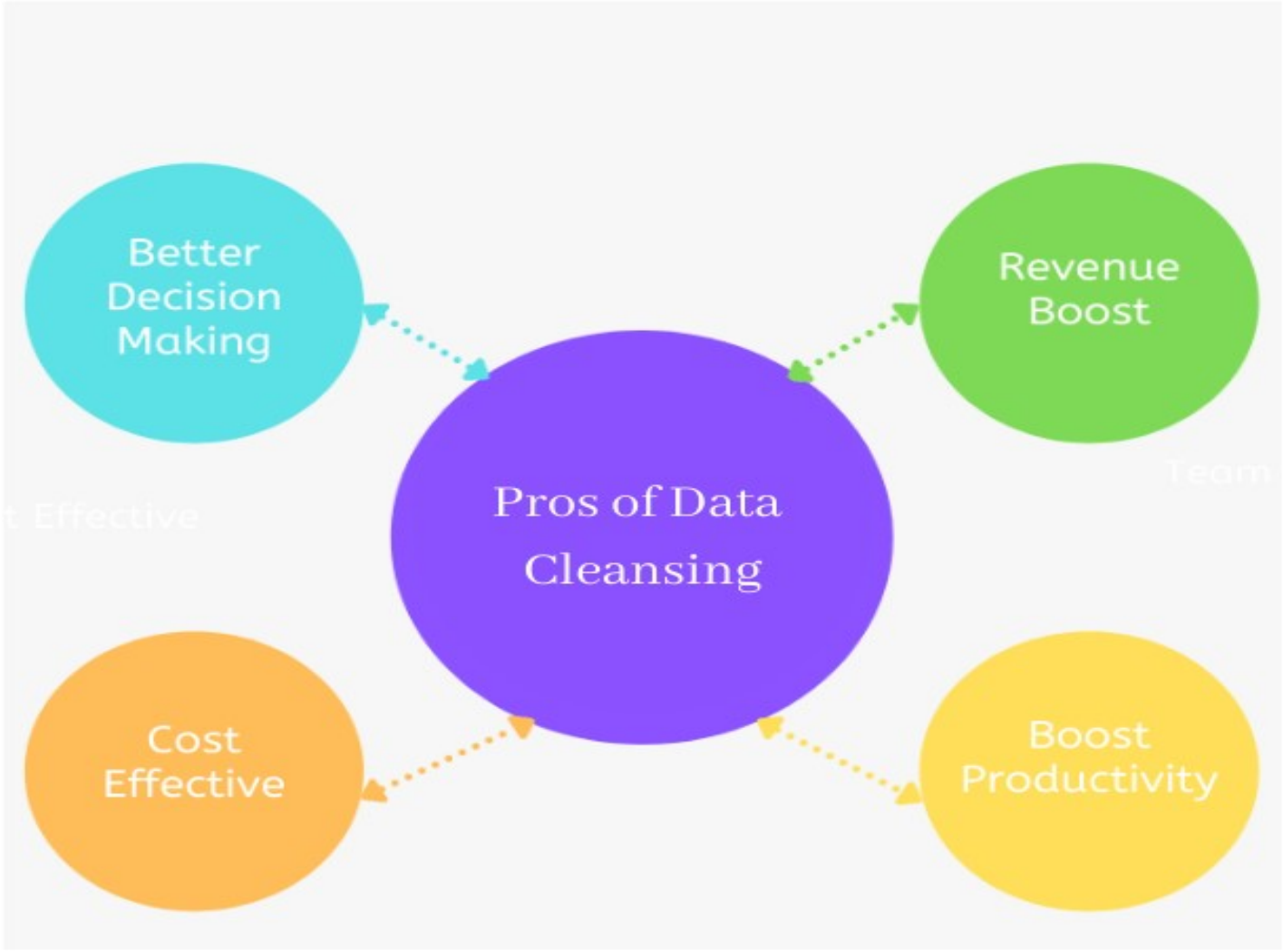
III YEAR / VI SEMESTER - EEE

Unit 1 - INTRODUCTION TO DATA SCIENCE

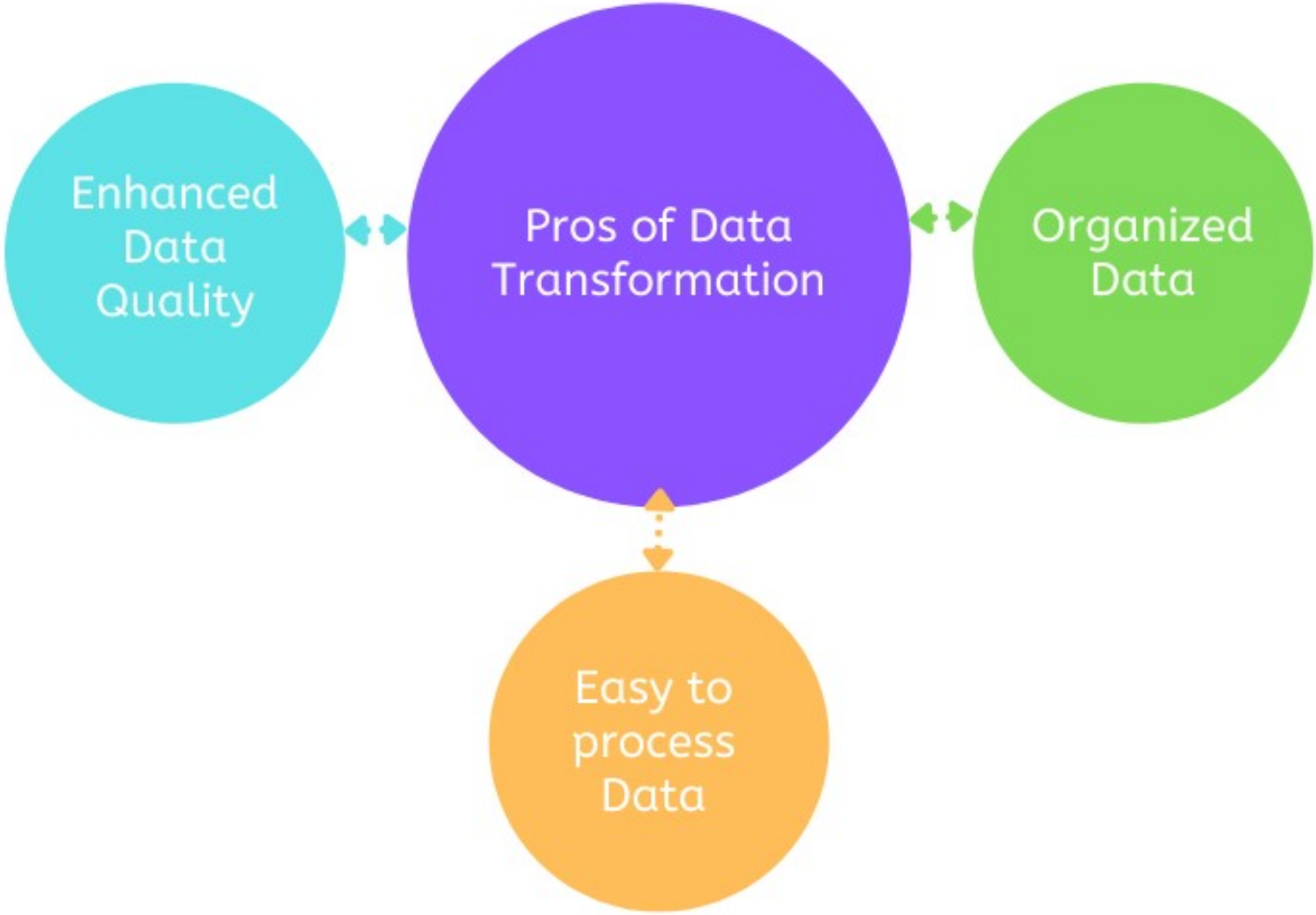
Topic : Cleansing, Integrating, and Transforming Data

K.KARTHIKEYAN AP/CSE,SNSCT

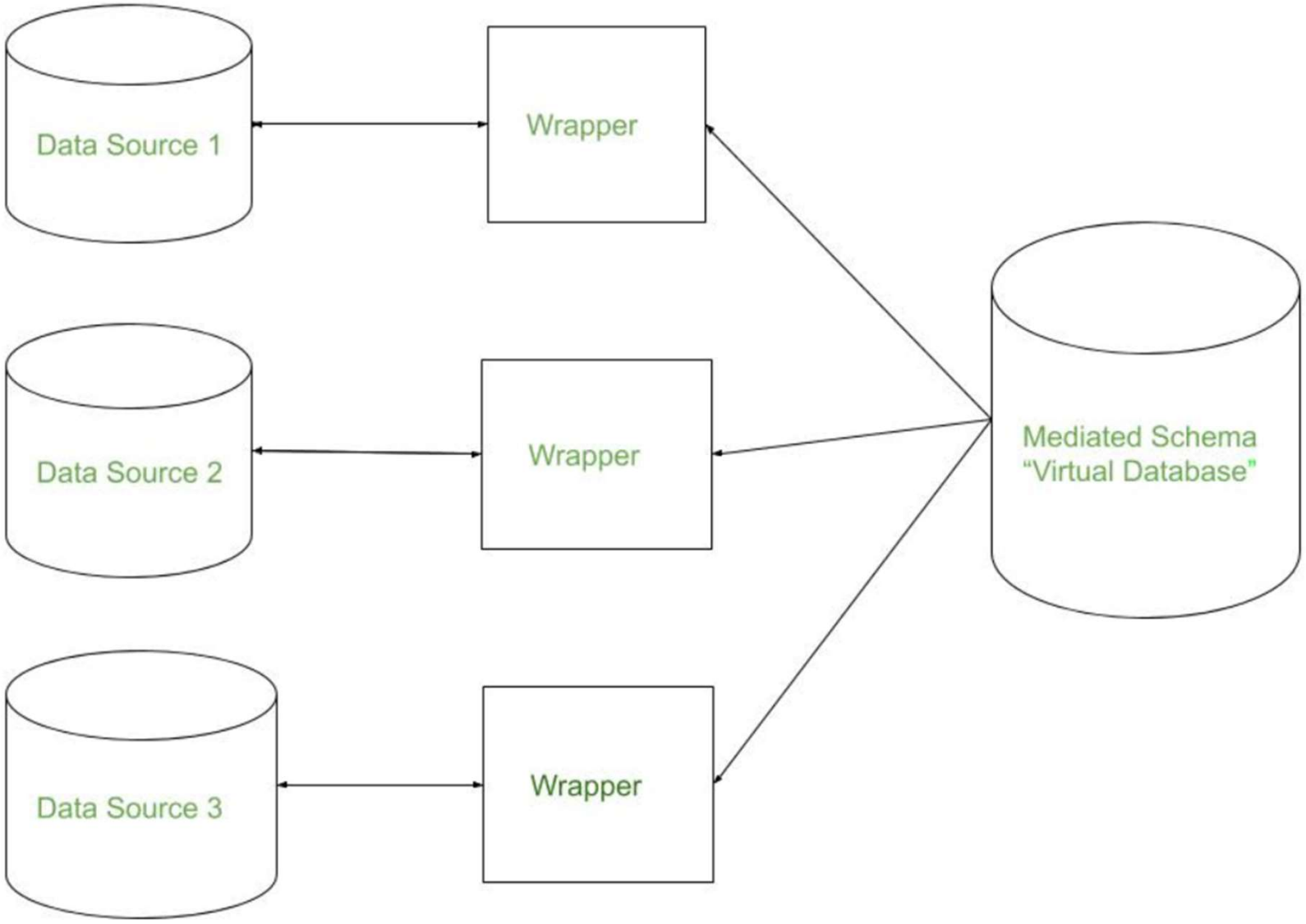
Recall in pros of Data cleansing



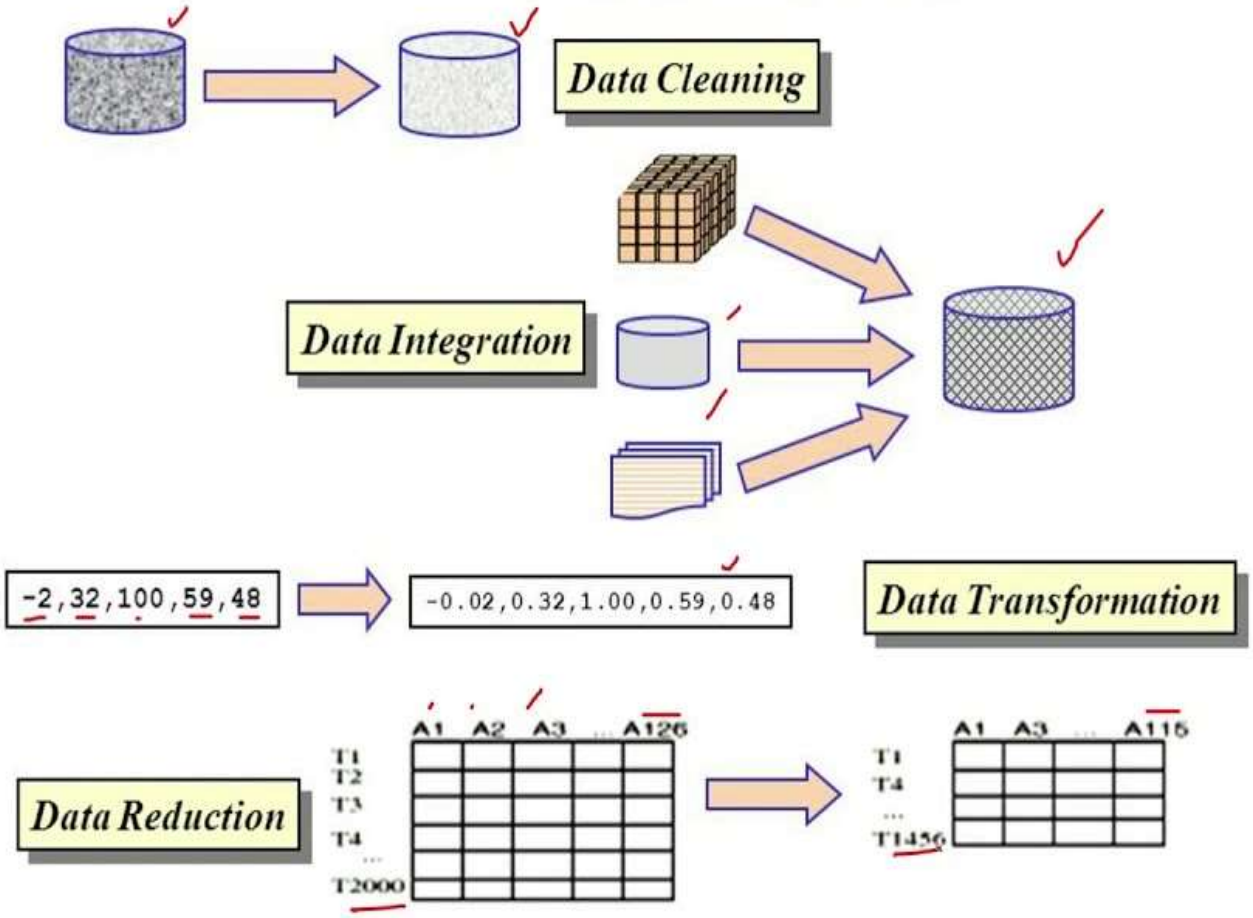
Empathy in pros of Data Transformation



Architecture of Virtual Database

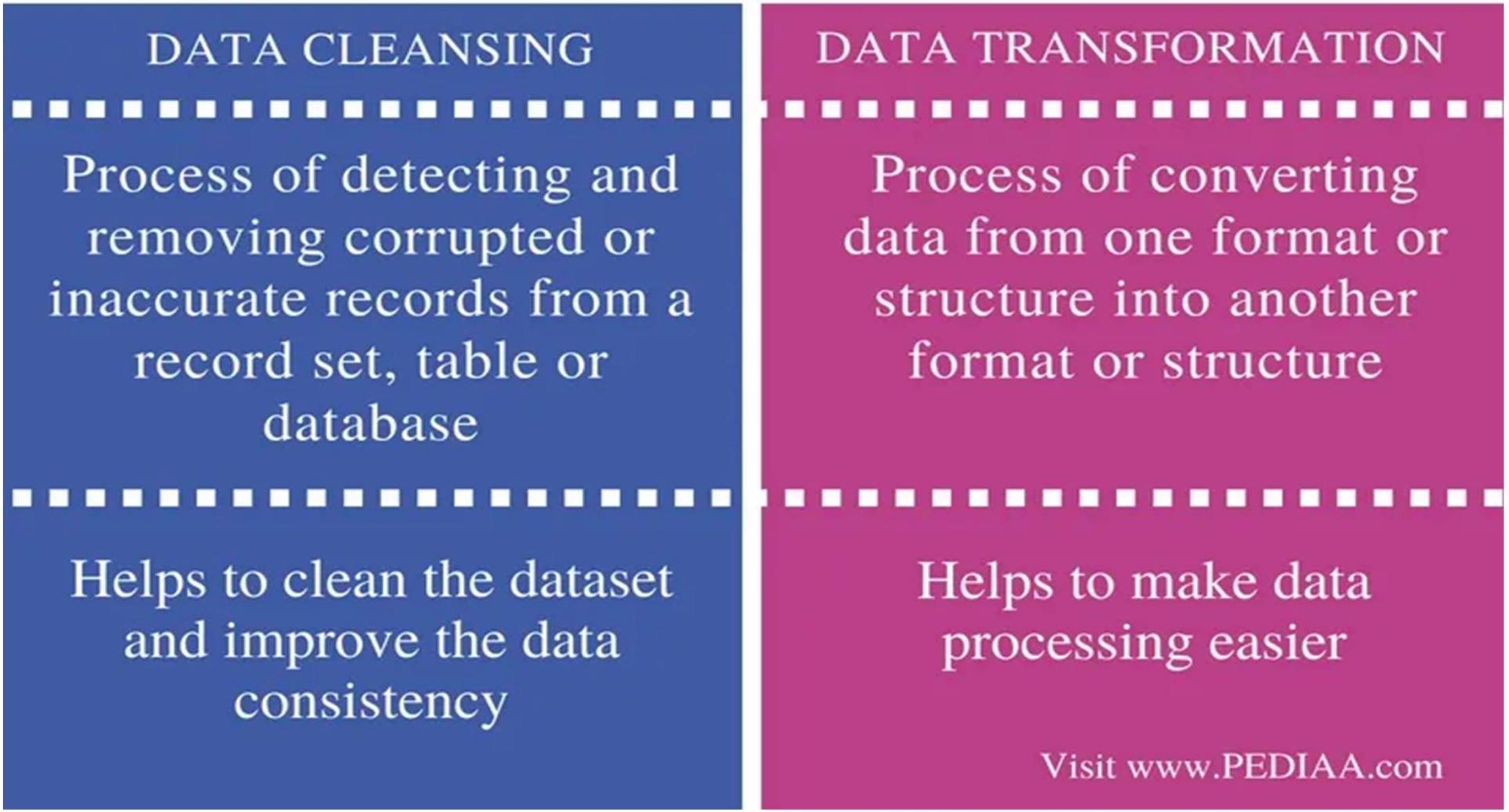


Forms of data preprocessing



Shahzad ALi [shahzad.ali@ue.edu.pk]

DATA CLEANSING VERSUS DATA TRANSFORMATION




END-TO-END
DATA CLEANSING
1
Identifying & Deleting Duplicate Data
2
Fixing Incomplete & Irrelevant Data
3
Inserting Missing Details
4
Auditing and Aggregation
5
Matching & Correlation
6
Interlinking & Consolidation
7
Identifying Key Variables
8
Suggesting New Variable
9
Address Data Sanitization
10
Case Correction & Conversion

Data Transformation and Cleaning Pipelines

1. Understanding Data Transformation
2. Building Data Transformation Pipelines
3. Data Quality Assurance
4. Handling Incremental Data

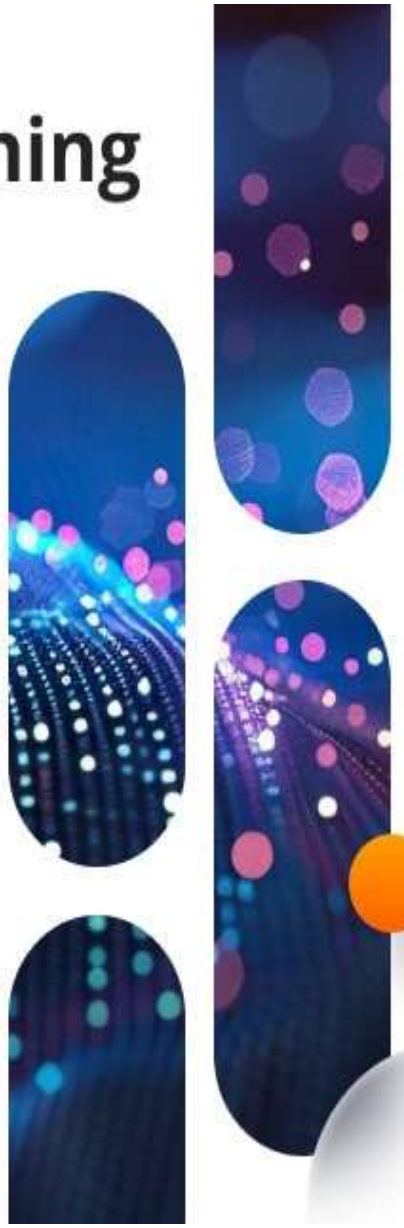
The difference between **data cleaning** and **data transformation**

data cleaning

removes data that does not belong in your dataset

data transformation

converts data from one format or structure into another



Activity - Fix & Fit Data

🧩 **Step 1: Clean the Data (8–10 min)**

Give students **one small messy table** with:

- Missing values
- Duplicate rows
- Spelling mistakes

Task:

✓ Find and correct the errors

🔄 **Step 2: Make it Ready (8–10 min)**

Using the same table, ask students to:

- Convert Yes/No → 1/0
- Group numbers (Low / Medium / High)

🏁 **Quick Discussion (5 min)**

Ask:

- Why is clean data important?
- How did transforming help?

MIND MAP


Goals & Objectives
Market Analysis
Innovation
Competitive Advantage
SWOT Analysis
• Strengths
• Weaknesses
• Opportunities
• Threats
Business Strategy
Team & Leadership
Financial Planning
Risk Management
Action Plan
Vision & Mission
Growth & Expansion

ASSESSMENT

MCQs: Cleansing, Integrating, and Transforming Data

1. Data cleansing mainly focuses on:

a) Data visualization

b) Improving data quality

c) Data storage

d) Model evaluation

Answer: b

---

2. Which of the following is an example of a data quality issue?

a) Normalized data

b) Duplicate records

c) Encoded values

d) Aggregated data

Answer: b

REFERENCE BOOKS

1.Allen B. Downey, “Think Stats: Exploratory Data Analysis in Python”, Green Tea Press, 2014.
2.Sanjeev J. Wagh, Manisha S. Bhende, Anuradha D. Thakare, “Fundamentals of Data Science”, CRC Press, 2022.
3.Chirag Shah, “A Hands-On Introduction to Data Science”, Cambridge University Press, 2020.
4.Vineet Raina, Srinath Krishnamurthy, “Building an Effective Data Science Practice: A Framework to Bootstrap and Manage a Successful Data Science Practice”, A press, 2021.


Thank
You