Chapter -3

Data Preprocessing

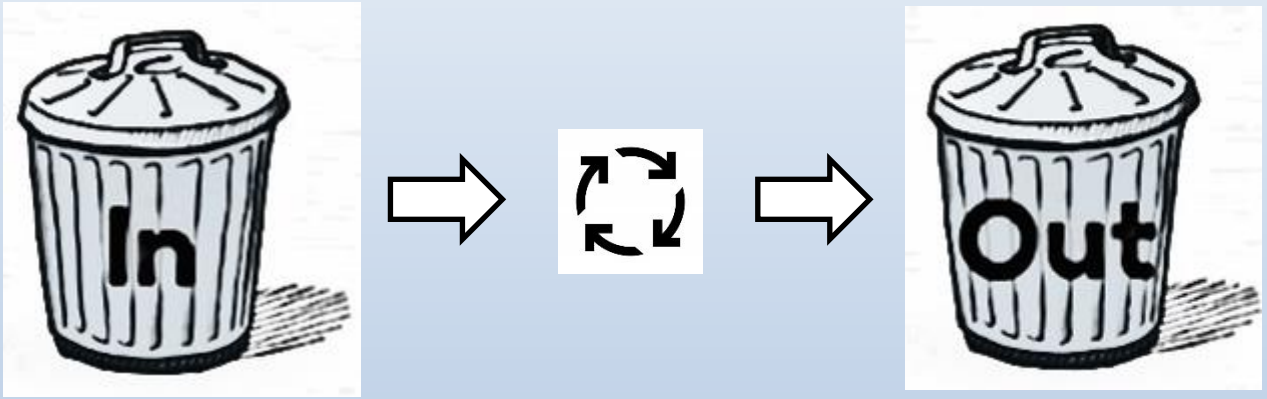
Why to preprocess data?

- Real world data are generally "**dirty**"
  - **Incomplete**: Missing attribute values, lack of certain attributes of interest, or containing only aggregate data.
    - **E.g.** Occupation=" "
  - **Noisy**: Containing errors or outliers.
    - **E.g.** Salary="abcxy"
  - **Inconsistent**: Containing similarity in codes or names.
    - **E.g.** "Gujarat" & "Gujrat" (Common mistakes like **spelling, grammar, articles**)

Why data preprocessing is important?

"No quality data, No quality results"

- It looks like **Garbage In Garbage Out (GIGO)**.



- Quality decisions must be based on **quality data**.
- Duplicate or missing data may cause incorrect or even misleading statistics.
- **Data preparation, cleaning and transformation are the majority task in data mining. (could be as high as 90%).**
- Data preprocessing **prepares** raw data for **further processing**.

Mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x$$

- Mean is the **average** of a dataset.
- To find the mean, calculate the sum of all the data and then divide by the total number of data.
- Example
  - ✓ Find out mean for **12, 15, 11, 11, 7, 13**

First, find the **sum of the data**.

12 + 15 +11 + 11 + 7 + 13 = **69**

Then **divide by the total number of data**.

69 / 6 = **11.5** ⇦ **Mean**

Median

- Median is the **middle number** in a dataset when the data is arranged in numerical order (Sorted Order).

If count is **Odd** then **middle number** is **Median**

If count is **Even** then take **average of middle two numbers** that is **Median**

Median - Odd (Cont..)

- Example
  - ✓ Find out Median for 12, 15, 11, 11, 7, 13, 15

In above example, count of data is **7**. (Odd)

First, arrange the **data** in **ascending order**.

7, 11, 11, 12, 13, 15, 15

Partitioning data into equal halfs

7, 11, 11, **12**, 13, 15, 15

12 ⇦ Median

Median - Even (Cont..)

- Example
  - ✓ Find out median for 12, 15, 11, 11, 7, 13

In above example, count of data is **6**. (Even)

First, arrange the **data** in **ascending order**.

7, 11, 11, 12, 13, 15

Calculate an **average** of the **two numbers** in the **middle**.

7, 11, 11, 12, 13, 15

(11 + 12)/2 = **11.5** ⇦ **Median**

Mode

- The mode is the **number that occurs most often** within a set of numbers.
- Example

1

Find mode.

12, 15, 11, 11, 7, 13

11 ⇦ Mode (Unimodal)

2

Find mode.

12, 15, 11, 11, 7, 12, 13

11, 12 ⇦ Mode (Bimodal)

Mode (Cont..)

- Example

3

Find mode.

12, 12 15, 11, 11, 7, 13, 7

7, 11, 12 ⇦ Mode (Trimodal)

4

Find mode.

12, 15, 11, 10, 7, 14, 13

No Mode

Range

- The range of a set of data is the **difference** between the **largest and the smallest number in the set**.
- Example
  - ✓ Find range for given data 40, 30, 43, 48, 26, 50, 55, 40, 34, 42, 47, 50

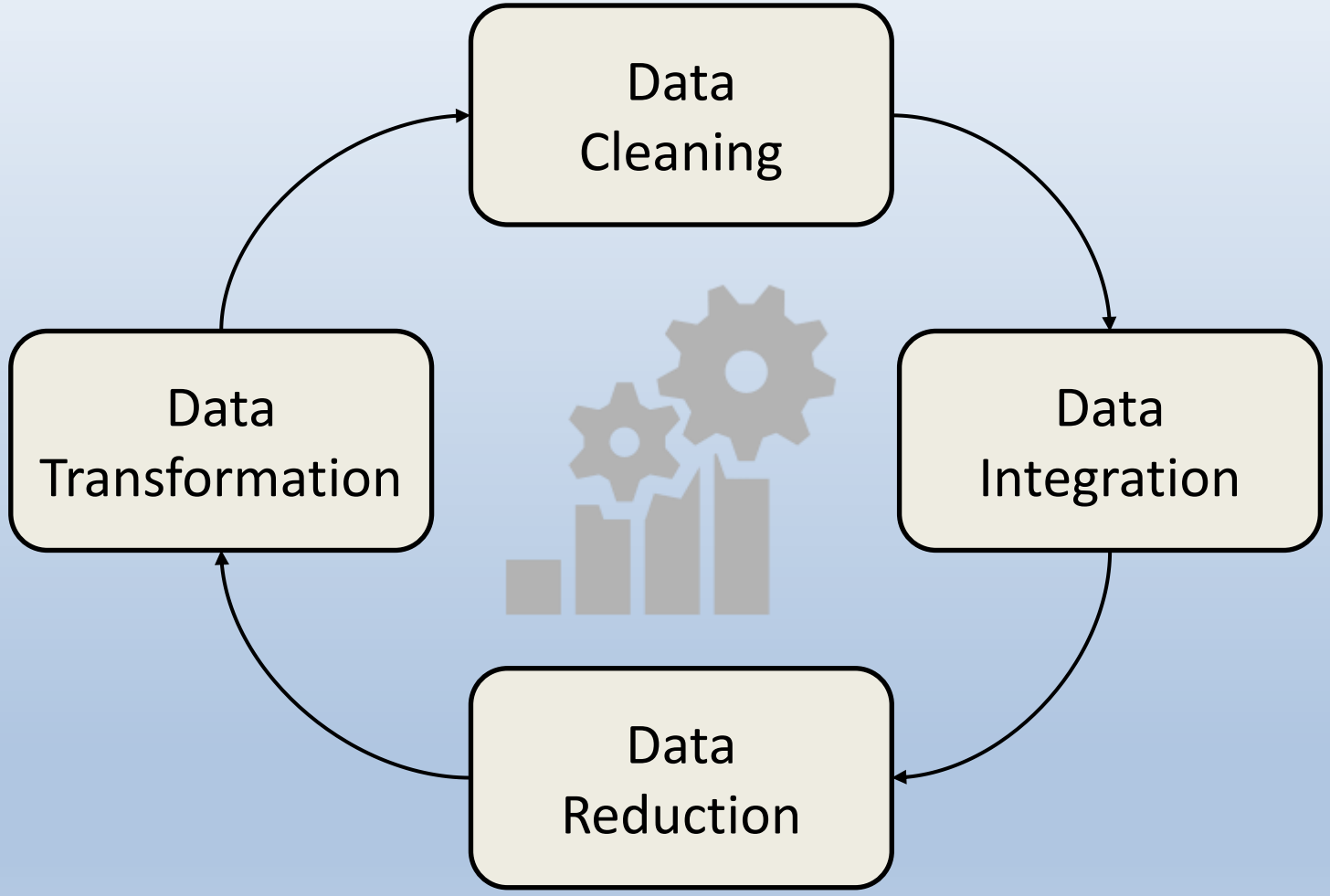
First, arrange the **data** in **ascending order**.

26, 30, 34, 40, 40, 42, 43, 47, 48, 50, 50, 55

- In our example **largest number is 55**, and subtract the **smallest number is 26**.

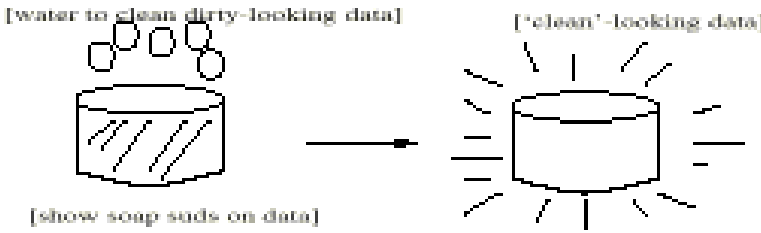
55 – 26 = 29⇦Range

Data Preprocessing Tasks

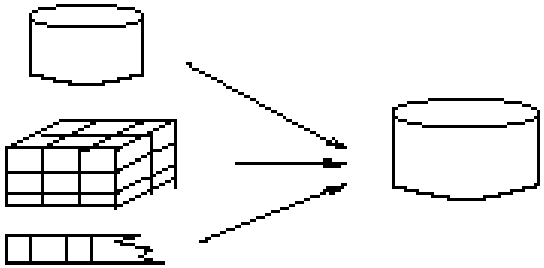


Forms of Data Preprocessing

Data Cleaning



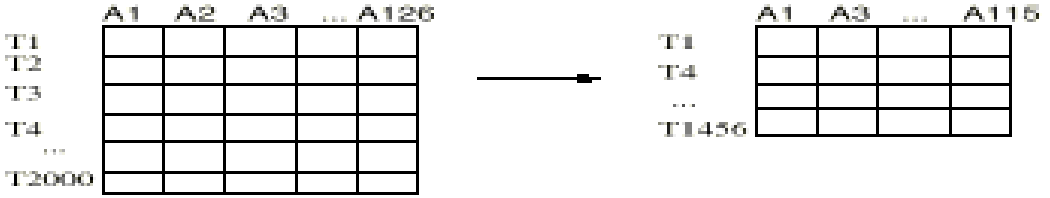
Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



1) Data Cleaning

1. **Fill in missing values**
    1. Ignore the tuple
    2. Fill missing value manually
    3. Fill in the missing value automatically
    4. Use a global constant to fill in the missing value
2. **Identify outliers and smooth out noisy data**
    1. Binning Method
    2. Clustering
3. **Correct inconsistent data**
4. **Resolve redundancy caused by data integration**

1) Fill missing values

Data Cleaning

- **Ignore the tuple (record/row)**:
  - Usually done when **class label is missing**.
  - **Example**
    - The task is to distinguish between two types of emails, “spam” and “non-spam” (Ham).
    - Spam & non-spam are called as class label.
    - If an email comes to you, in which class label is missing then it is discarded.
- **Fill missing value manually**:
  - Use the **attribute mean (average)** to **fill in the missing value** and **also use the attribute mean (average)** for **all samples belonging to the same class**.

1) Fill missing values (Cont..)

Data Cleaning

- **Fill in the missing value automatically**:
  - **Predict** the **missing value** by using a **learning algorithm**:
    - Consider the attribute with the missing value as a dependent variable and run a learning algorithm (usually Naive Bayes or Decision tree) to predict the missing value.
- **Use a global constant to fill in the missing value**
  - Replace **all missing attribute values** by the same constant such as a label like ***"Unknown"***.

2) Identify outliers and smooth out noisy data

Data Cleaning

1. **Binning method**
2. **Clustering**

1) Binning method

- Data binning or **bucketing** is a data pre-processing technique used to **reduce the effects of minor observation errors**.
- The original data values which fall in a given small interval called **as a bin** are **replaced by a value which represents that interval**, often called the central value.
- **Steps of Binning method**
  1. **Sort the attribute values** and **partition** them into **bins**.
  2. Then smooth by **bin means, bin median** or **bin boundaries**.

Binning method - Example

- Given data: **4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34**
- Step: 1
- Partition into **equal-depth [n=4]**:

  Bin 1: 4, 8, 9, 15

  Bin 2: 21, 21, 24, 25

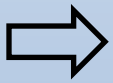
  Bin 3: 26, 28, 29, 34

- Step: 2
  - Smoothing by **bin means**:

    Bin 1: 9, 9, 9, 9

    Bin 2: 23, 23, 23, 23

    Bin 3: 29, 29, 29, 29



(4 + 8 + 9 + 15)/4 = **9**

(21 + 21 + 24 + 25)/4 = **23**

(26 + 28 + 29 + 34)/4 = **29**

Binning method - Example (Cont..)

- Given data: **4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34**
- Step: 1
- Partition into **equal-depth [n=4]**:

  Bin 1: 4, 8, 9, 15

  Bin 2: 21, 21, 24, 25

  Bin 3: 26, 28, 29, 34

- Step: 2
  - Smoothing by **bin boundaries**:

    Bin 1: **4**, **4**, **4**, **15**

    Bin 2: **21**, **21**, **25**, **25**

    Bin 3: **26**, **26**, **26**, **34**

1) Binning method (Cont..)

- Binning method is a **top-down splitting technique** based on a **specified number of bins**.
- It is also used as **discretization method** for data reduction and concept hierarchy generation.
- For example, attribute values can be discretized (separated) by applying equal-width or equal-frequency binning, and then replacing each value by the bin mean or median.
- It can be applied **recursively to the resulting partitions** to **generate concept hierarchies**.
- It **does not use class information**, therefore it is an **unsupervised discretization technique**.

2) Clustering

- Clustering is a process of **partitioning a set of data** (or objects) into a **set of meaningful sub-classes**, called clusters.
- It enables the abstraction of **large amounts data** by forming **meaningful groups or categories of objects**.
- In clustering, objects in the same cluster are similar to each other and those in different clusters are dissimilar.
- **Example**
  - Library (Group of Books based on different categories)
  - Cloths (By size S, M, L, XL, XXL etc.)

3) Correct inconsistent data

Data Cleaning

- If you have inconsistencies in your data, it can cause major problems later on.
- But with larger datasets, it can be difficult to find all of the inconsistencies.
- **It contains similarity in codes or names.**
- We can manually solve common mistakes like spelling, grammar, articles or use other tools for it.

4) Resolve redundancy caused by data integration

Data Cleaning

- Data redundancy occurs in database systems **which have a field that is repeated in two or more tables**.
- When customer data is duplicated and attached with each product bought, then redundancy of data is known as **inconsistency**.
- So, the entity "customer" **might appear with different values**.
- Database **normalization** prevents redundancy and makes the best possible usage of storage.
- The proper use of **foreign keys** can minimize data redundancy and reduce the chance of destructive anomalies appearing.