DATA ANALYTICS

Unit 1:Data Preprocessing

Mamatha.H.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 1:Data Preprocessing

Mamatha H R

Department of Computer Science and Engineering

Why do we need Data preprocessing?

- **Data preprocessing is** an integral step in Machine Learning/Data mining as the quality of **data** and the useful information that **can** be derived from it directly affects the ability of our model to learn;

- Therefore, it **is** extremely important that **we preprocess** our **data** before feeding it into our model

Important Characteristics of Data

- Dimensionality (number of attributes)
  - High dimensional data brings a number of challenges
- Sparsity
  - Only presence counts
- Resolution
  - Patterns depend on the scale
- Size
  - Type of analysis may depend on size of data

Data Quality

- Poor data quality negatively affects many data processing efforts

"The most important point is that poor data quality is an unfolding disaster.

- Poor data quality costs the typical company at least ten percent (10%) of revenue; twenty percent (20%) is probably a better estimate."

Thomas C. Redman, DM Review, August 2004

- example: a classification model for detecting people who are loan risks is built using poor data
  - Some credit-worthy candidates are denied loans
  - More loans are given to individuals that default

Data Quality

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

- Examples of data quality problems:
  - Noise and outliers
  - Missing values
  - Duplicate data
  - Wrong data
  - Fake data

Noise

- For objects, noise is an extraneous object
- For attributes, noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen

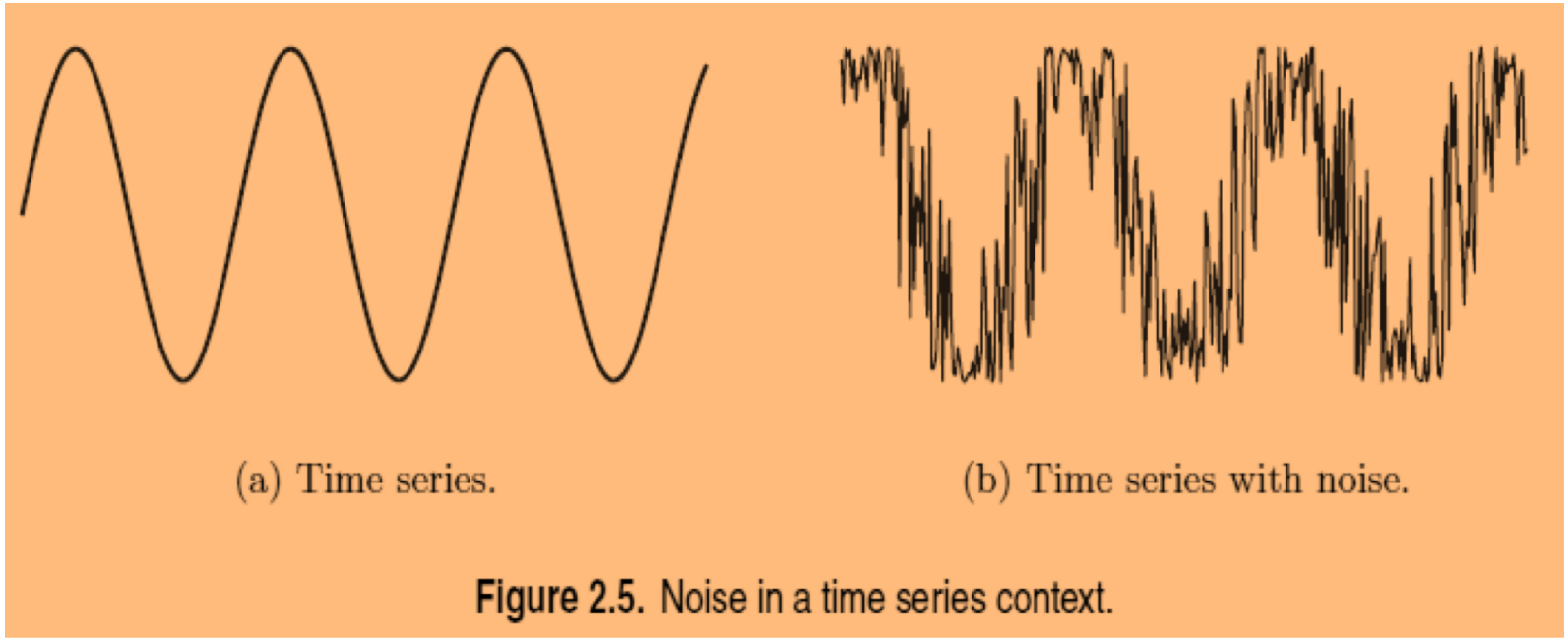
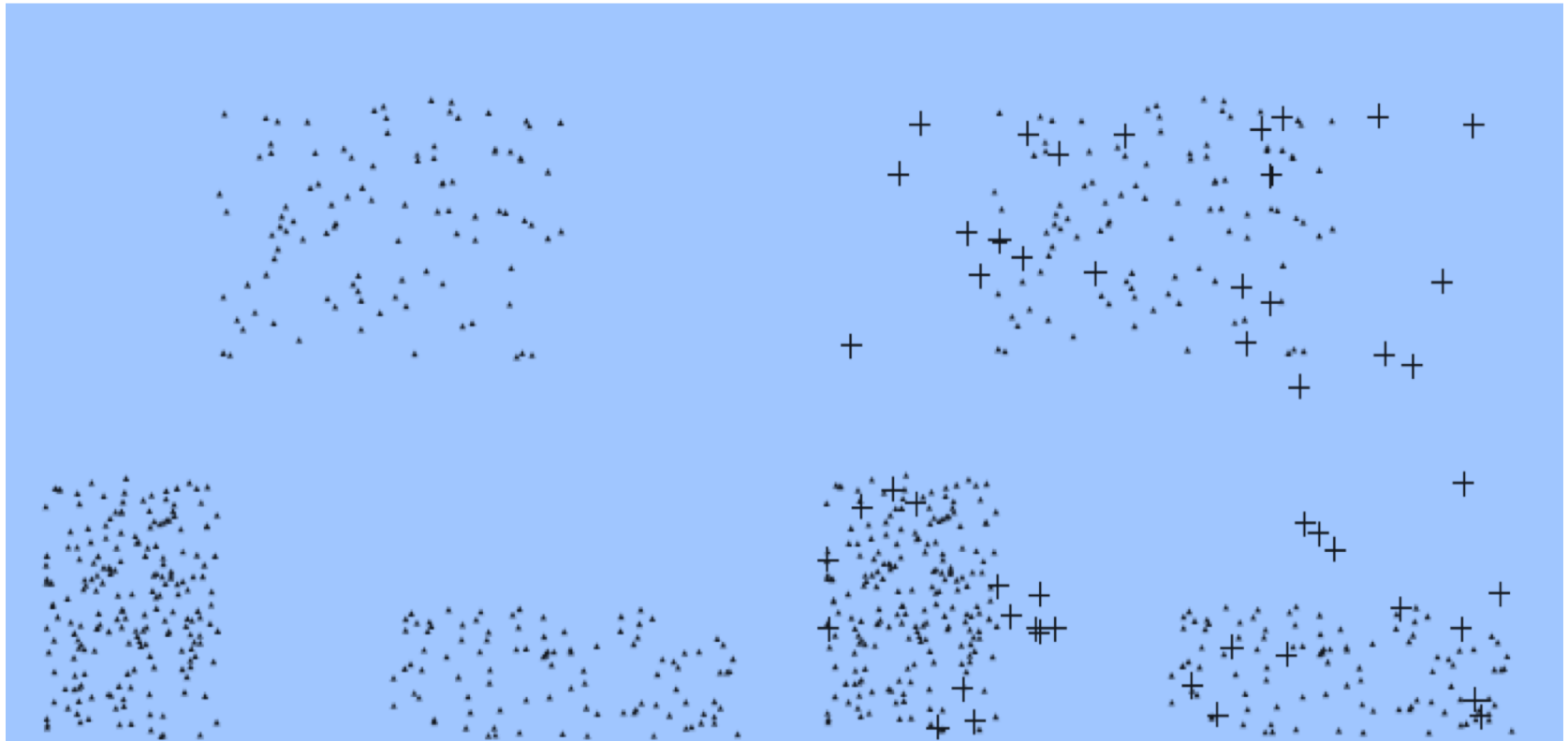


Figure 2.5. Noise in a time series context.

Noise



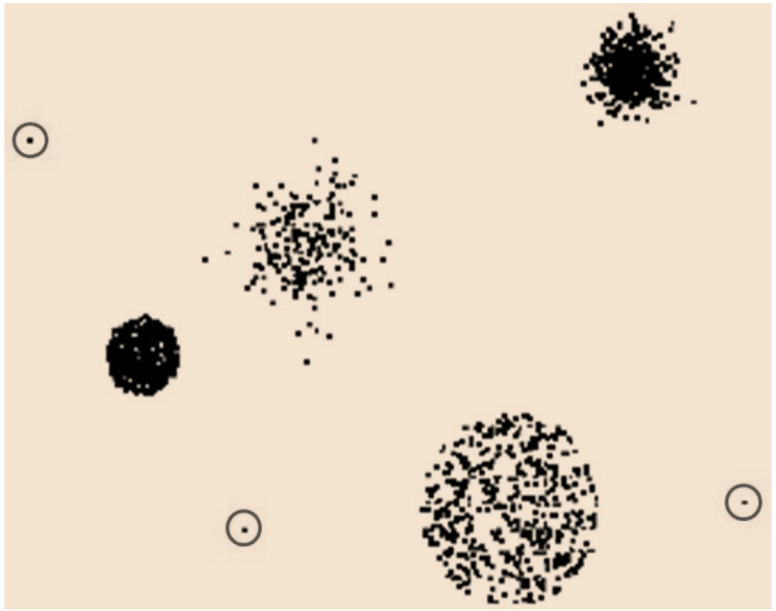
(a) Three groups of points.

(b) With noise points (+) added.

Figure 2.6. Noise in a spatial context.

Outliers

- ***Outliers*** are data objects with characteristics that are considerably different than most of the other data objects in the data set
  - **Case 1:** Outliers are noise that interferes with data analysis
  - **Case 2:** Outliers are the goal of our analysis
    - Credit card fraud
    - Intrusion detection



Missing Values

- Reasons for missing values
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)

- Handling missing values
  - Eliminate data objects or variables
  - Estimate missing values
    - Example: time series of temperature
    - Example: census results
  - Ignore the missing value during analysis

Missing Values …

- Missing completely at random (MCAR)
  - Missingness of a value is independent of attributes
  - Fill in values based on the attribute
  - Analysis may be unbiased overall
- Missing at Random (MAR)
  - Missingness is related to other variables
  - Fill in values based other values
  - Almost always produces a bias in the analysis
- Missing Not at Random (MNAR)
  - Missingness is related to unobserved measurements
  - Informative or non-ignorable missingness
- Not possible to know the situation from the data

Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources

- Examples:
  - Same person with multiple email addresses

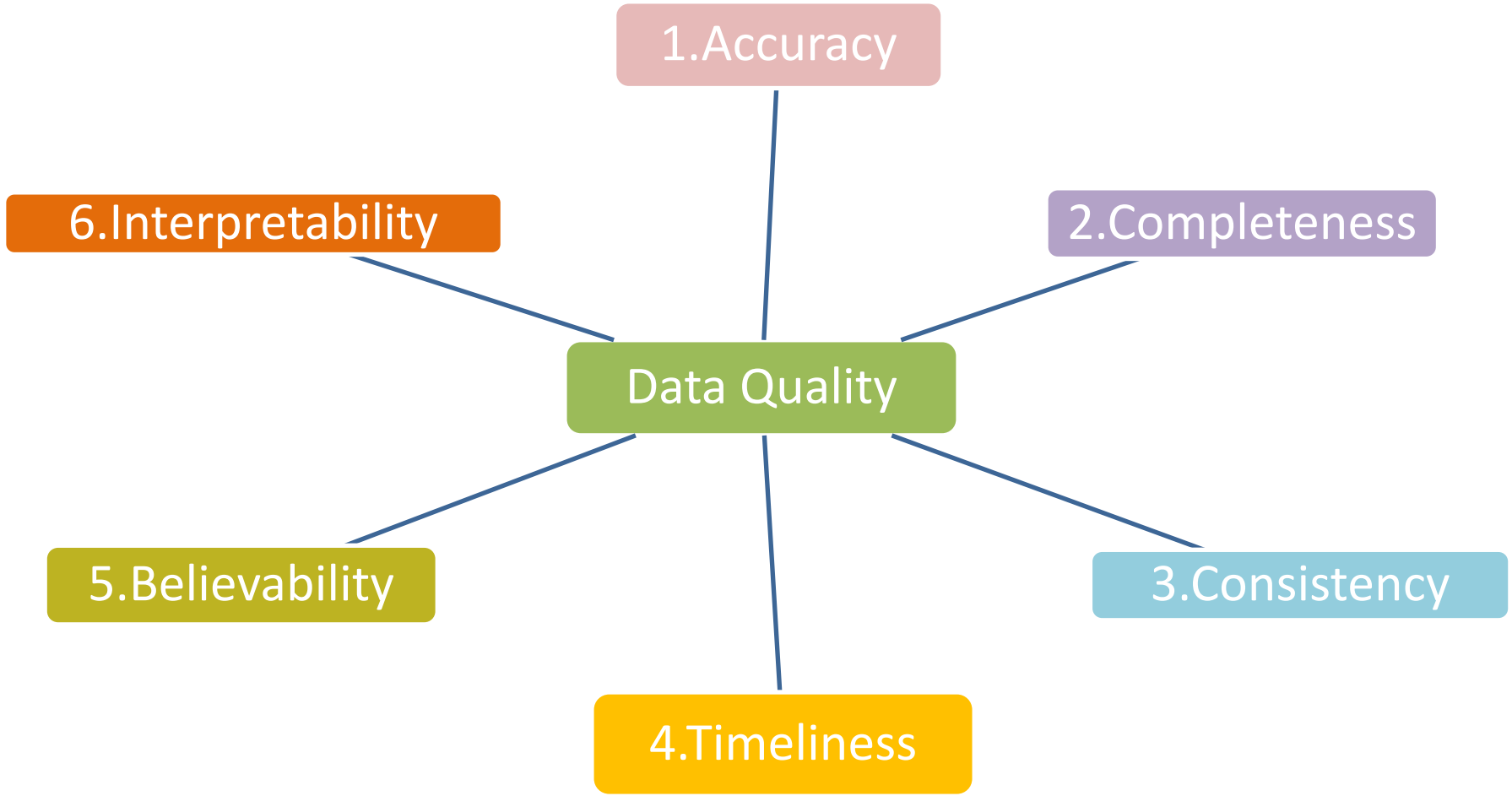
- Data cleaning
  - Process of dealing with duplicate data issues

- When should duplicate data not be removed?

Similarity and Dissimilarity Measures

- Similarity measure
  - Numerical measure of how alike two data objects are.
  - Is higher when objects are more alike.
  - Often falls in the range [0,1]
- Dissimilarity measure
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- Proximity refers to a similarity or dissimilarity

A Multidimensional view of data quality

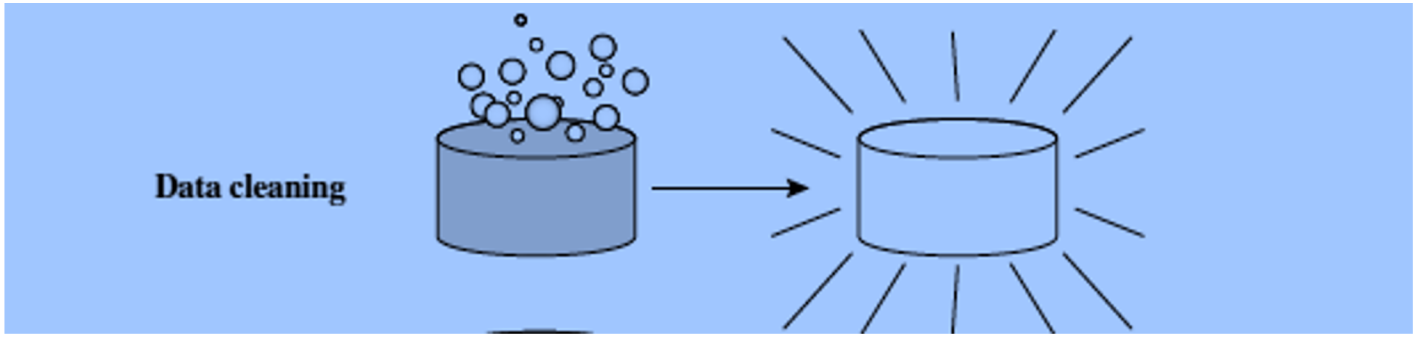


Data Quality: Why Preprocess the Data?

- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

Major Tasks in Data Preprocessing

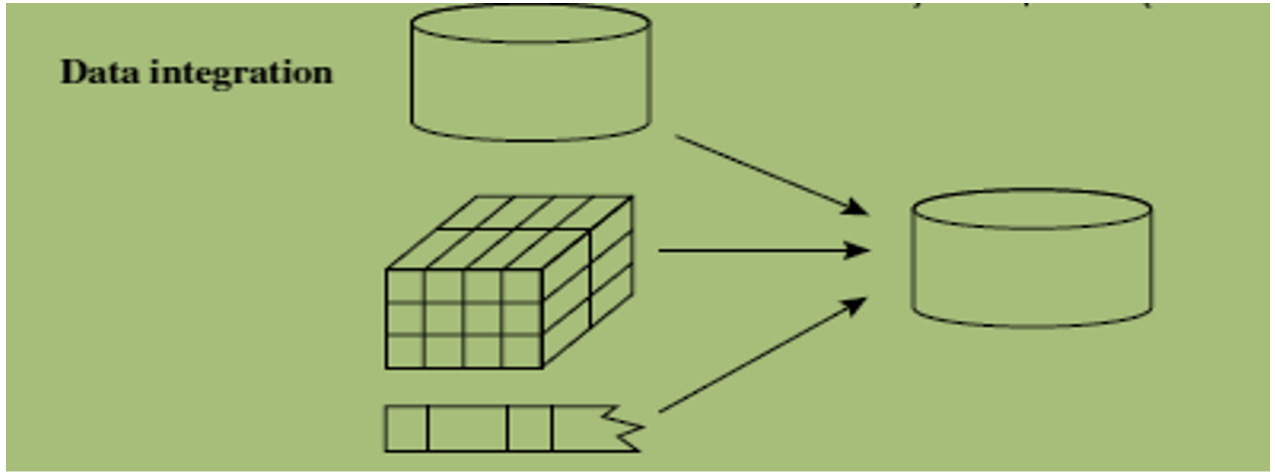
1. **Data cleaning**
   - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies



Major Tasks in Data Preprocessing

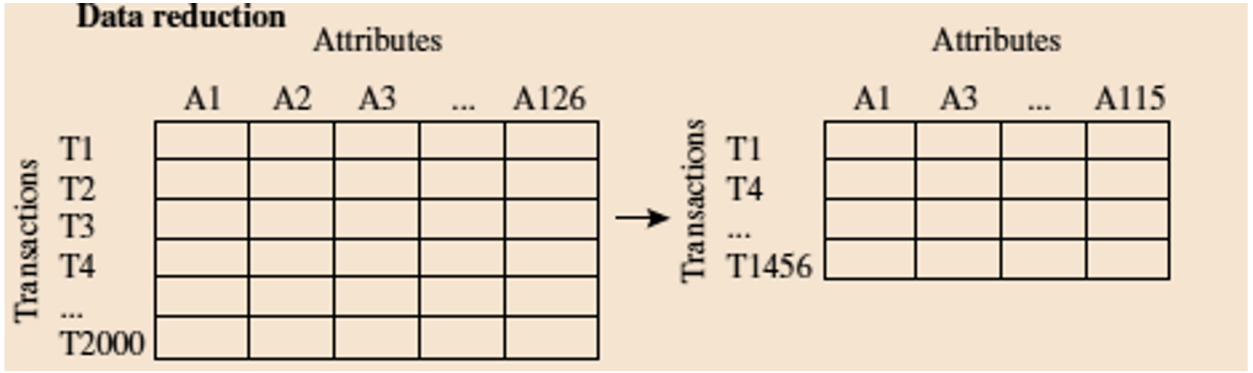
2. Data integration

- Integration of multiple databases, data cubes, or files



Major Tasks in Data Preprocessing

3. **Data reduction**
   - Dimensionality reduction
   - Numerosity reduction
   - Data compression



Major Tasks in Data Preprocessing

4. **Data transformation and data discretization**
   - Normalization
   - Concept hierarchy generation

Data transformation $-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

Summary :Data Preprocessing

PES UNIVERSITY ONLINE

Data preprocessing

- Data cleaning
- Data integration
- Data reduction
- Data transformation and data discretization

Exercise

- Mention the important characteristics of the data.
- Why we need to pre-process the data?
- Explain the process of data preprocessing.

References

Text Book:

- Data Mining: Concepts and Techniques by Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems, 3rd Edition.
- Introduction to Data Mining by Tan, Steinbach, Kumar, 2nd Edition

THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834