



DATA ANALYTICS

Unit 1:Data Pre-processing

Mamatha.H.R

Department of Computer Science and
Engineering

DATA ANALYTICS

Unit 1:Data Pre-processing

Mamatha H R

Department of Computer Science and Engineering

- Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogeneous sources.
- Low-quality data will lead to low-quality analysis results.
- “How can the data be preprocessed in order to help improve the quality of the data and, consequently, of the analysis results?”
- How can the data be preprocessed so as to improve the efficiency and ease of the analysis process?”

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: 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?

- 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.”

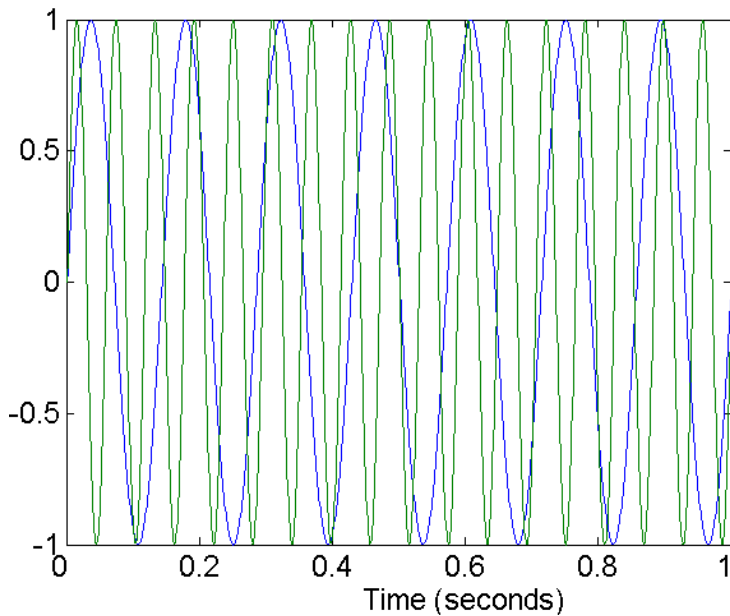
Thomas C. Redman, DM Review, August 2004

- 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

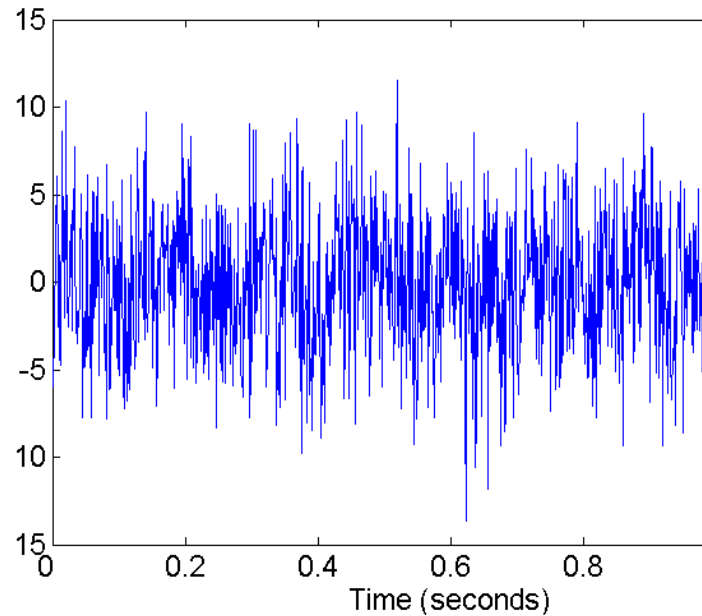
- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

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

- 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



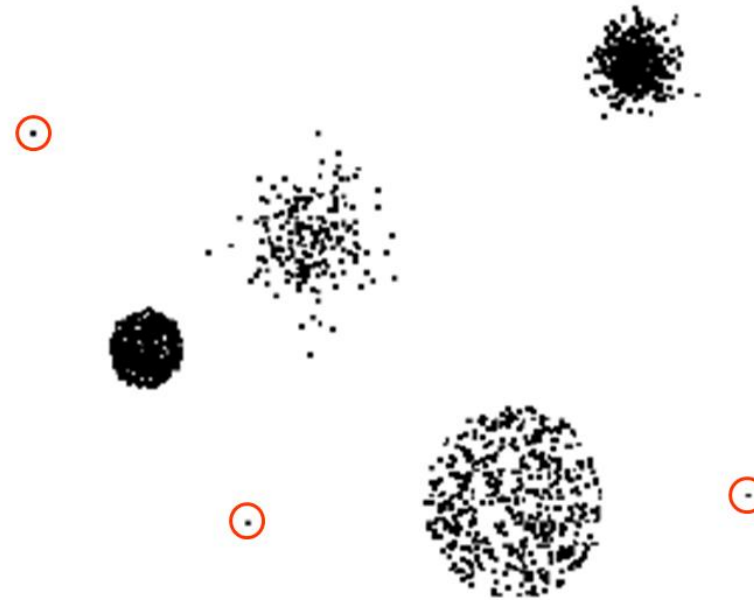
Two Sine Waves



Two Sine Waves + Noise

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



- 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 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

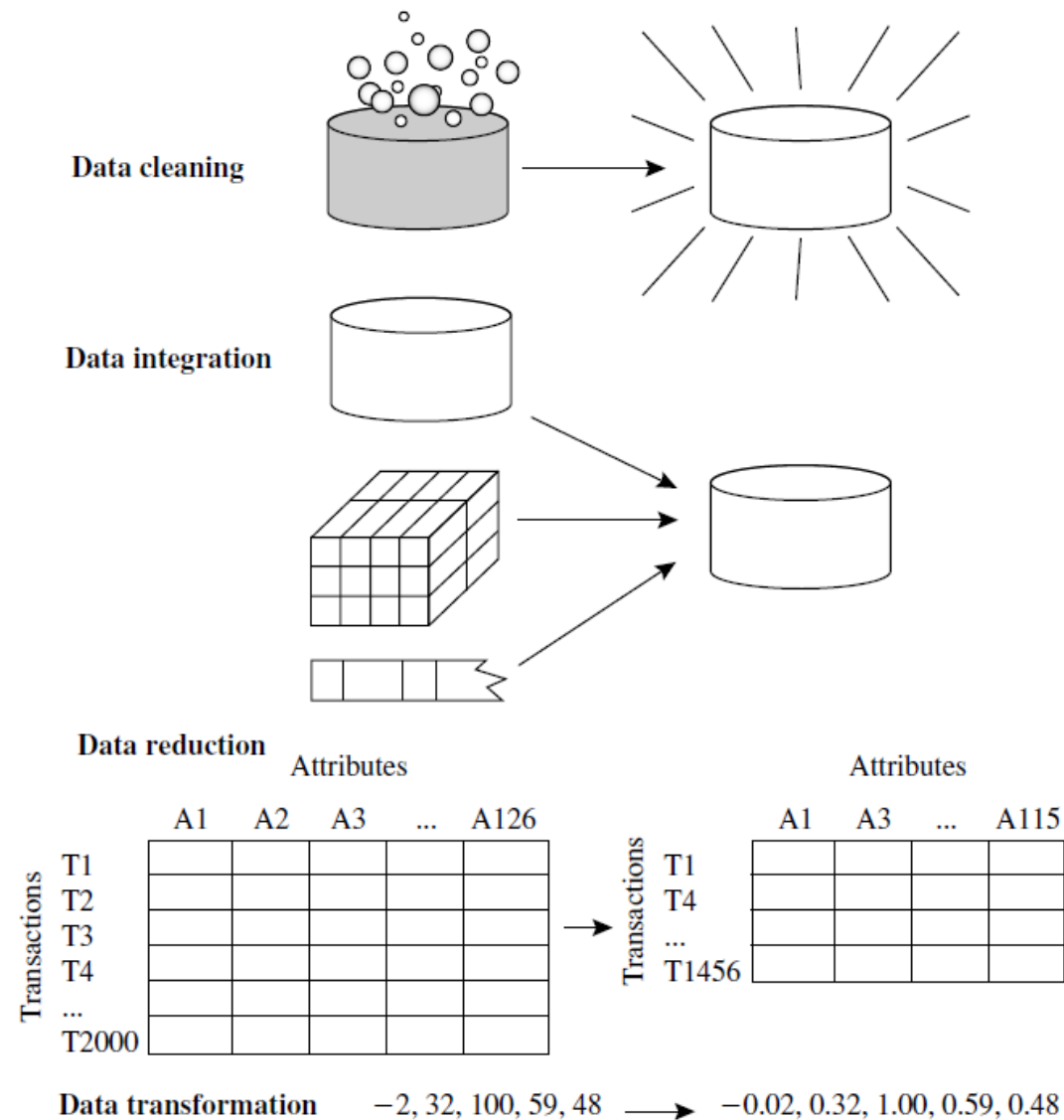
- Data set may include data objects that are duplicates, or almost duplicates of one another
 - Major issue when merging data from heterogeneous sources
- Examples:
 - Same person with multiple email addresses
- Data cleaning
 - Process of dealing with duplicate data issues
- When should duplicate data not be removed?

- Similarity 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

Major Tasks in Data Preprocessing

- **Data cleaning**
 - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
 - Integration of multiple databases, data cubes, or files

- **Data reduction**
 - Dimensionality reduction
 - Numerosity reduction
 - Data compression
- **Data transformation and data discretization**
 - Normalization
 - Concept hierarchy generation



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

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