DATA MINING PRE-PROCESSING

Outline

1. Why data preprocessing?
2. Data cleaning
3. Data integration and transformation
4. Data reduction
5. Discretization and concept hierarchy generation
6. Summary

2

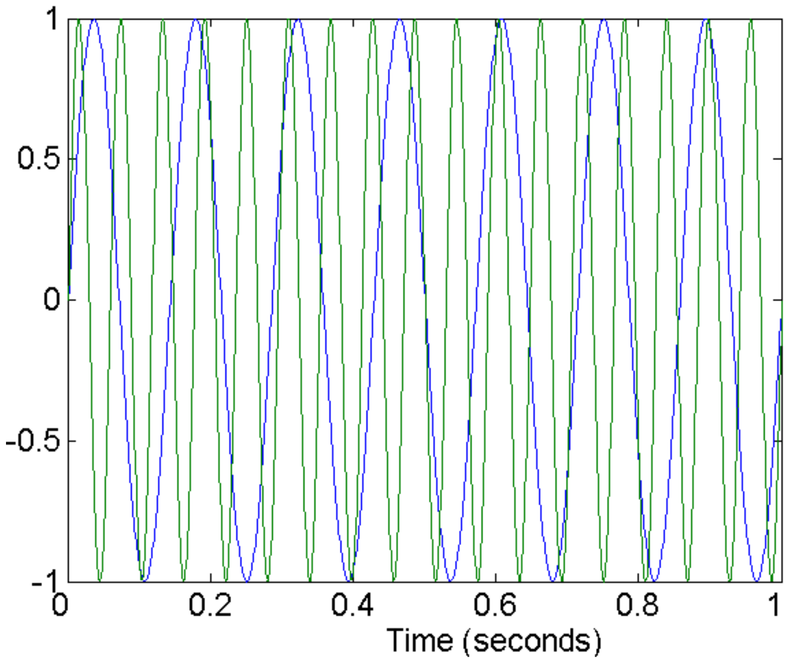
Data Cleaning

Data Quality

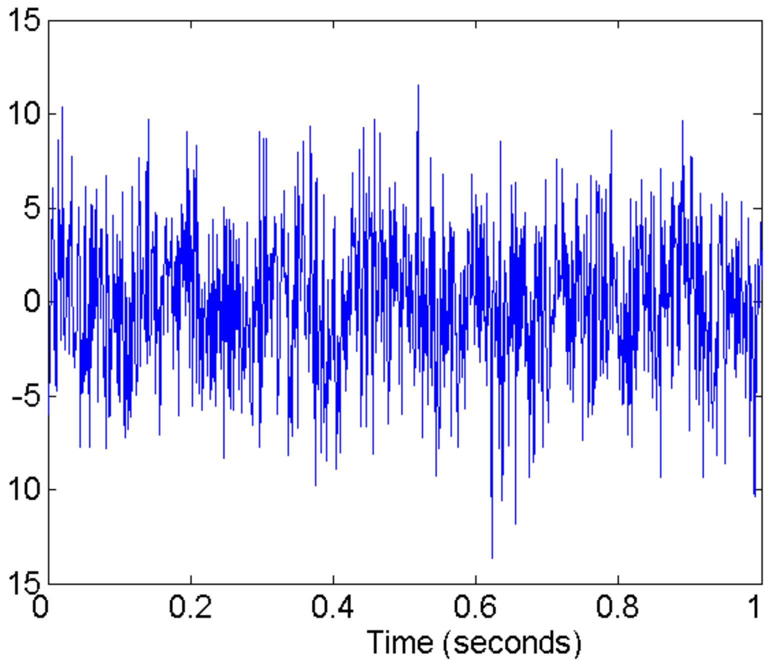
- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?
- Examples of data quality problems:
  - Noise and outliers
  - missing values
  - duplicate data

Noise

- 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



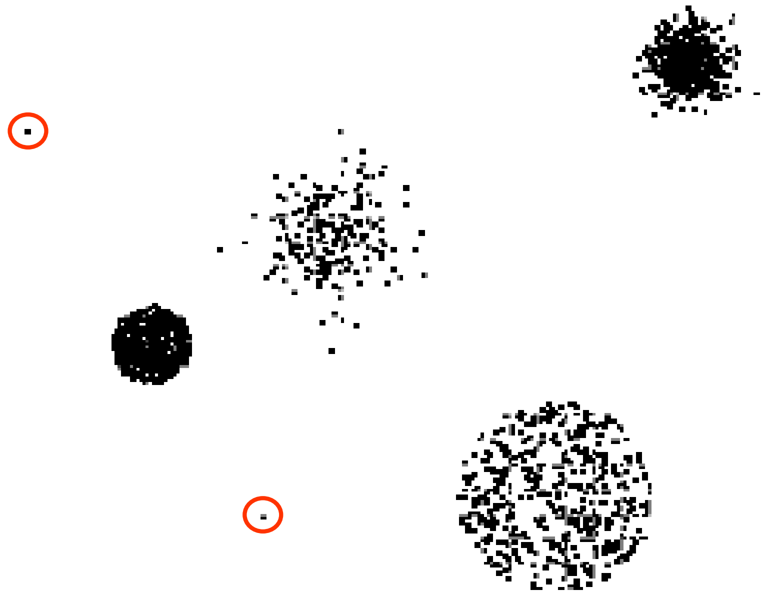
Two Sine Waves



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



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
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)

Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogenous sources
- Examples:
  - Same person with multiple email addresses
- Data cleaning
  - Process of dealing with duplicate data issues

Data Cleaning

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

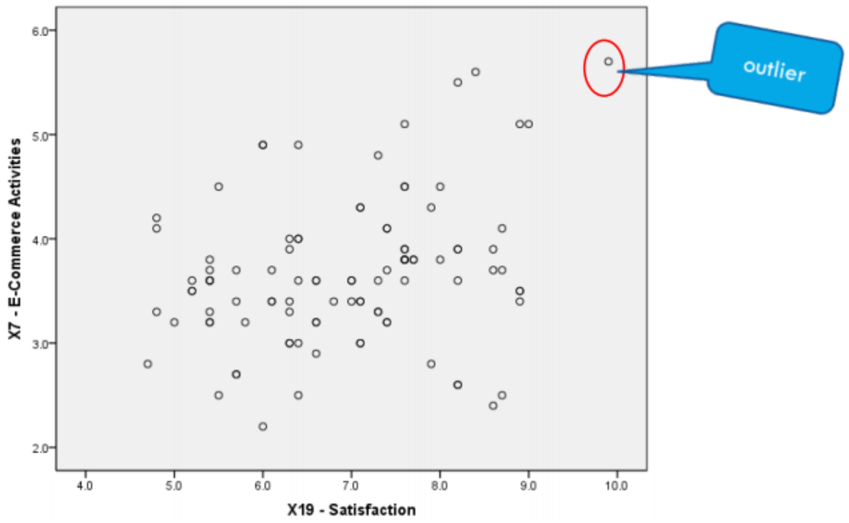
How to Handle Missing Data?

- ✓Ignore the tuple: usually done when class label is missing (assuming the task is classification—not effective in certain cases)
- ✓Fill in the missing value manually: tedious + infeasible?
- ✓Fill in it automatically with
  - Use a global constant to fill in the missing value: e.g., “unknown”, a new class?!
  - Use the attribute mean to fill in the missing value
  - Use the attribute mean for all samples of the same class to fill in the missing value: smarter
  - Use the most probable value to fill in the missing value: inference-based such as regression, Bayesian formula, decision tree

Outlier Detection

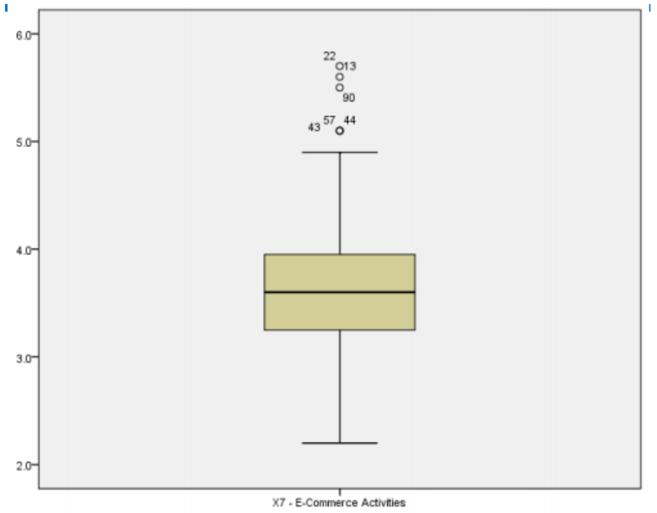
✓ Univariate outlier detection
  - scatter plot
  - box plot
  - standardized data

✓ Multivariate outlier detection



Univariate outlier: scatterplot

Univariate outlier: boxplot



Noisy Data

- Q: What is noise?
- A: Random error in a measured variable.
- Incorrect attribute values may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which requires data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

✓Binning method:

- first sort data and partition into (equi-depth) bins
- then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- used also for discretization (discussed later)

✓Clustering

- detect and remove outliers

✓Semi-automated method: combined computer and human inspection

- detect suspicious values and check manually

✓Regression

- smooth by fitting the data into regression functions

Simple Discretization Methods: Binning

- Equal-width (distance) partitioning:
  - It divides the range into *N* intervals of equal size: uniform grid
  - if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: *W* = (*B*-*A*)/*N.*
  - The most straightforward
  - But outliers may dominate presentation
  - Skewed data is not handled well.
- Equal-depth (frequency) partitioning:
  - It divides the range into *N* intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky.

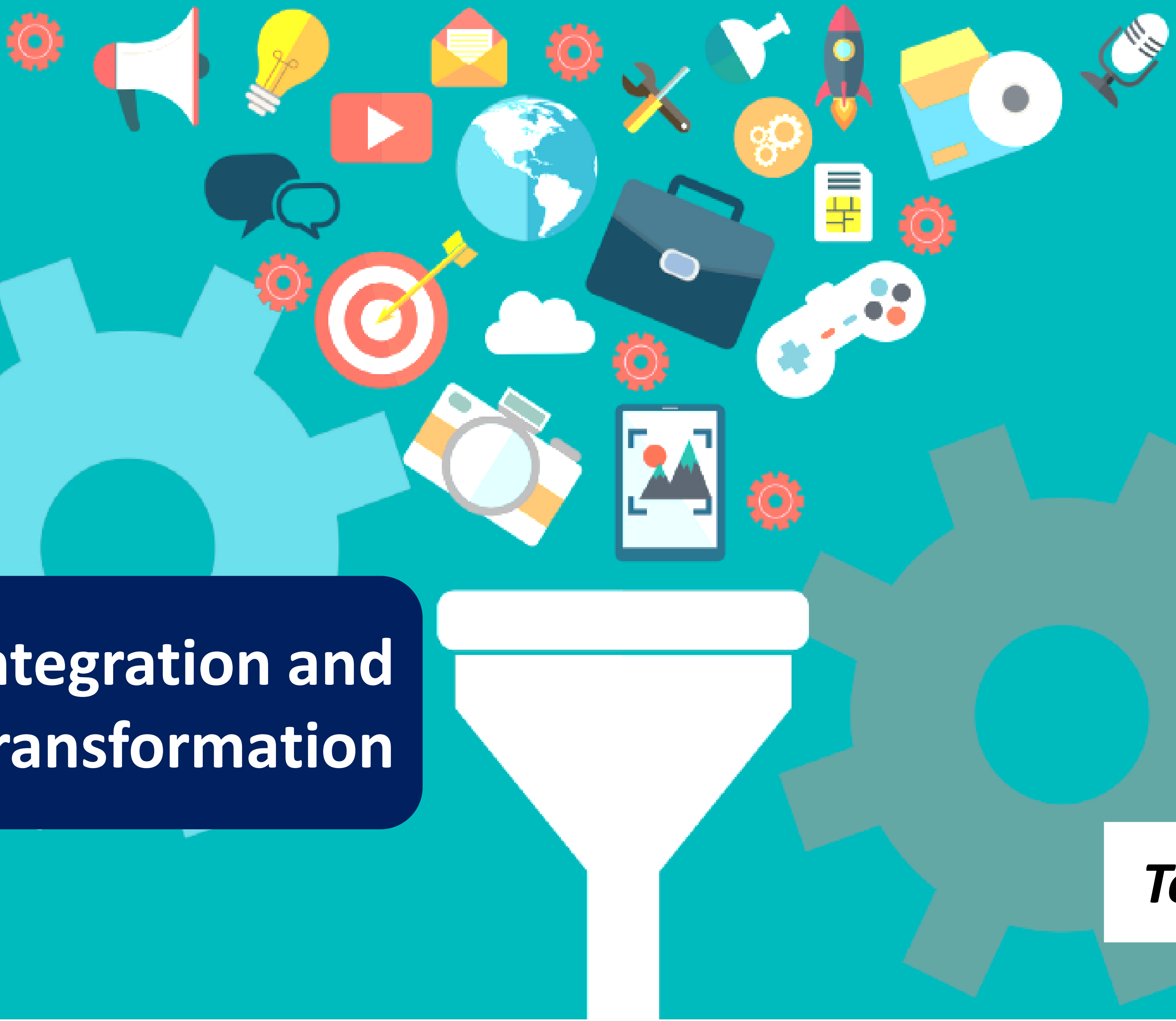
Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34

- Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
- Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

How to Handle Inconsistent Data?

- Manual correction using external references
- Semi-automatic using various tools
  - To detect violation of known functional dependencies and data constraints
  - To correct redundant data



3 Data Integration and Transformation

To be continued