Data Preprocessing

(Concepts and Techniques)

oleh
Jiawei Han
University of Illinois at Urbana-Champaign
www.cs.uiuc.edu/~hanj

Dimodifikasi oleh
Dr. Taufik Fuadi Abidin, S.Si., M.Tech)

Mengapa Diperlukan Data Preprocessing?

- Data *in the real world is dirty* (tidak sempurna)
  - *incomplete*: nilai atribut tidak lengkap, attribut yang seharusnya ada tidak ada, atau hanya data agrigasi yang tersedia (aggregate data)
    - e.g., Occupation = " ", Jenis_kelamin = " "
  - *noisy*: mengandung error atau outliers
    - e.g., Gaji = "-100.000"
  - *inconsistent*: terjadi perbedaan (*discrepancies*) dalam pengkodean dan nilai
    - e.g., Age="42" Birthday="03/07/1980"
    - e.g., Sebelumnya rating "1,2,3", sekarang "A, B, C"
    - e.g., Terjadi perbedaan pada data yang duplikat

Why Is Data Dirty?

- **Incomplete data** dapat terjadi karena
  - Pada saat dikumpulkan, nilai dari atribut tertentu tidak tersedia "*not applicable*"
  - Terjadi perbedaan pertimbangan sewaktu data dikumpulkan dengan sewaktu data dianalisa
  - Problem yang disebabkan oleh manusia/hardware/software
- **Noisy data (incorrect values)** dapat terjadi karena
  - Faulty data collection instruments (kesalahan pada alat)
  - Human atau komputer error pada saat entry data
  - Terjadi error pada saat dikirim (*errors in data transmission*)
- **Inconsistent data** dapat terjadi karena
  - Perbedaan sumber data (*different data sources*)
  - Pelanggaran ketergantungan fungsionalitas (*functional dependency violation*) e.g., modify some linked data
- Terjadinya **Duplikasi Record (Data)**

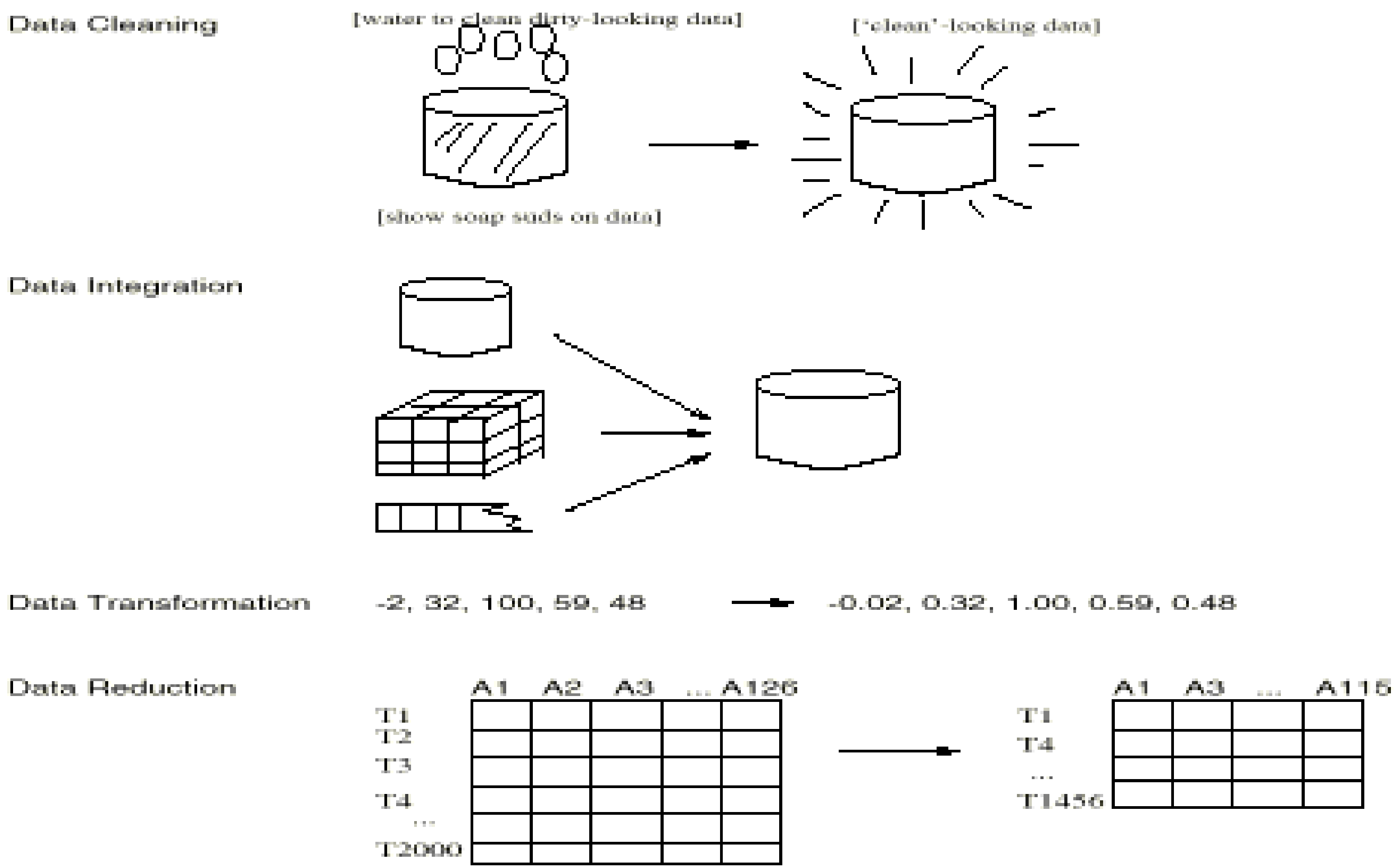
Mengapa Data Preprocessing Penting?

- No quality data, no quality mining results! (*Garbage in, garbage out*)
  - Keputusan yang baik harus berdasarkan data yang berkualitas pula (*Quality decisions must be based on quality data*)
    - e.g., duplicate or missing data may cause **incorrect** or even **misleading statistics**
  - Data warehouse membutuhkan gabungan data-data yang berkualitas
- Data extraction, cleaning, dan transformation merupakan bagian terpenting dari **data warehouse**

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 or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
  - Part of data reduction but with particular importance, especially for numerical data

Ilustrasi dari Beberapa Jenis Data Preprocessing



Data Summarization
Mengukur Nilai Tengah (Central Tendency)

- Mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

e.g: 4, 36, 45, 50, 75

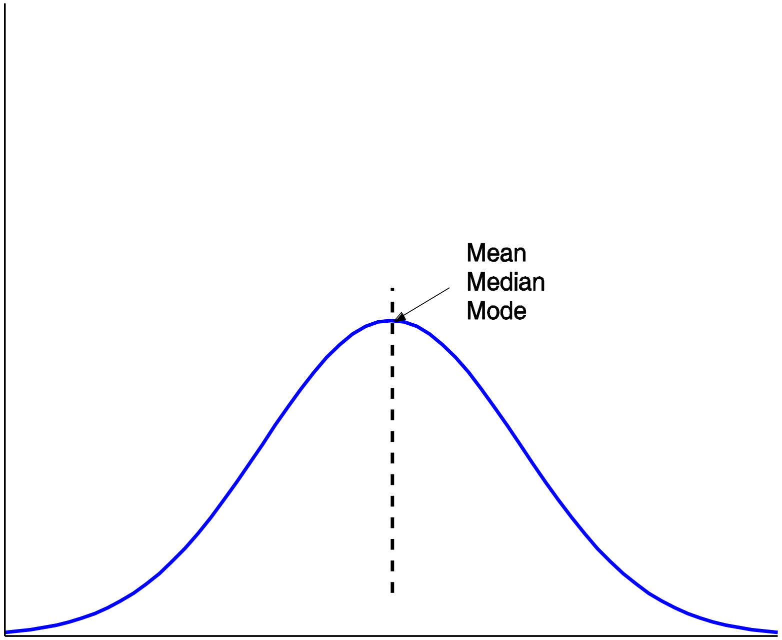
- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise e.g: 1, 5, 2, 8, 7

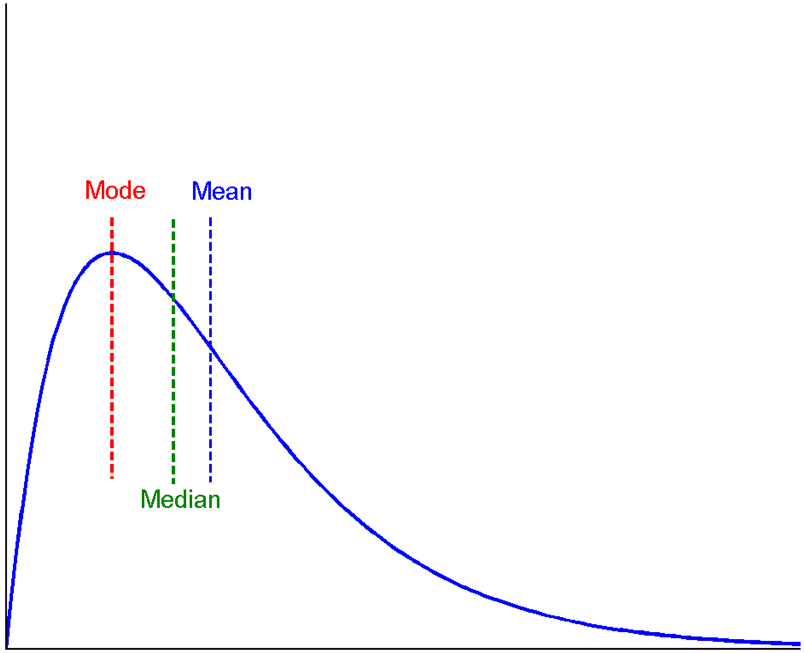
- Mode e.g: 1, 3, 6, 6, 6, 6, 7, 7, 12, 12, 17
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula:

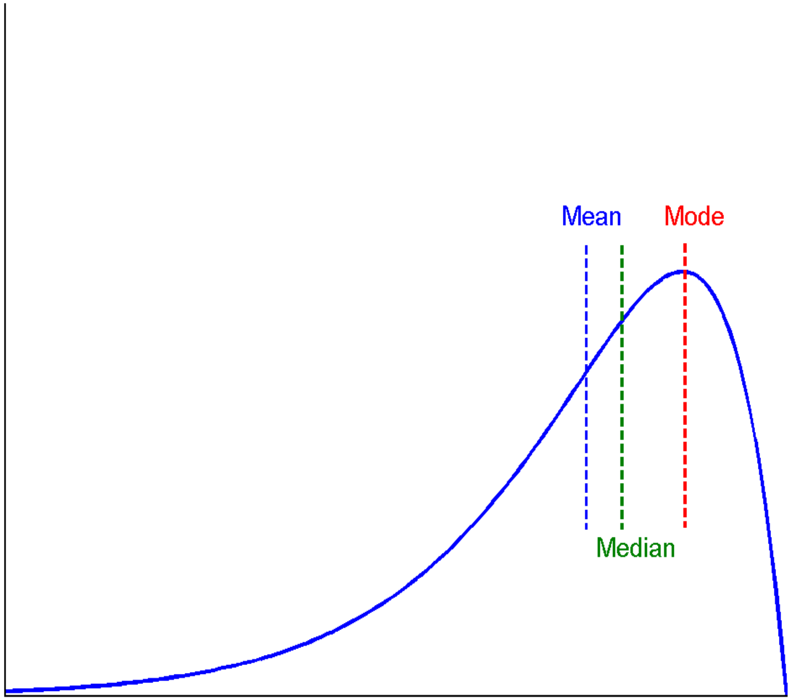
$$mean - mode = 3 \times (mean - median)$$

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data







Measuring the Dispersion of Data

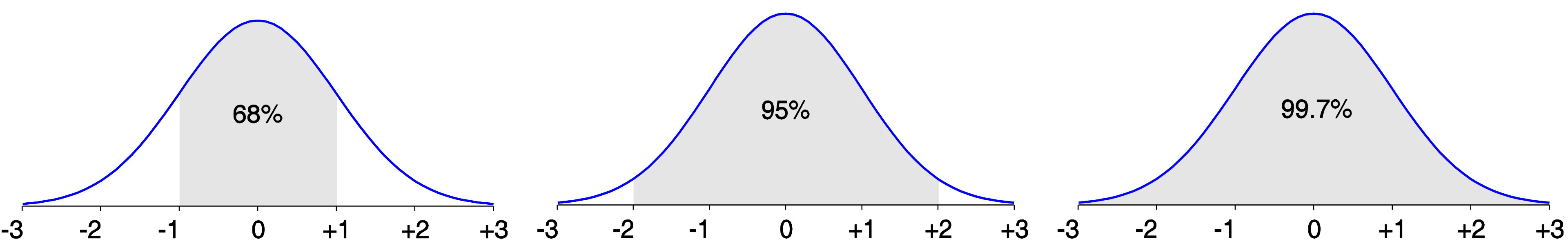
- Quartiles, outliers ex: 9, 14, 17, 19, 22, 32, 35, 42, 99
  - Quartiles: Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
  - Inter-quartile range: IQR = $Q_3 - Q_1$
  - Five number summary: min, Q_1, M, Q_3, max
  - Outlier: usually, a value higher/lower than 1.5 x IQR from median
- Variance and standard deviation
  - Variance: (algebraic, scalable computation)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^{\,2} - \mu^2$$

  - Standard deviation *s (or σ)* is the square root of variance σ^2

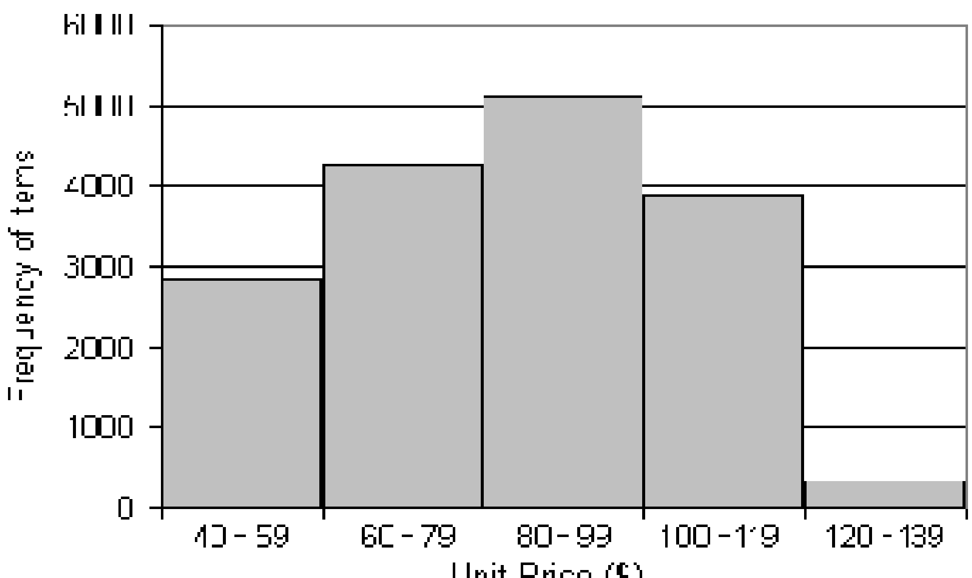
Properties of Normal Distribution Curve

- The normal (distribution) curve
  - From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From μ–2σ to μ+2σ: contains about 95% of it
  - From μ–3σ to μ+3σ: contains about 99.7% of it

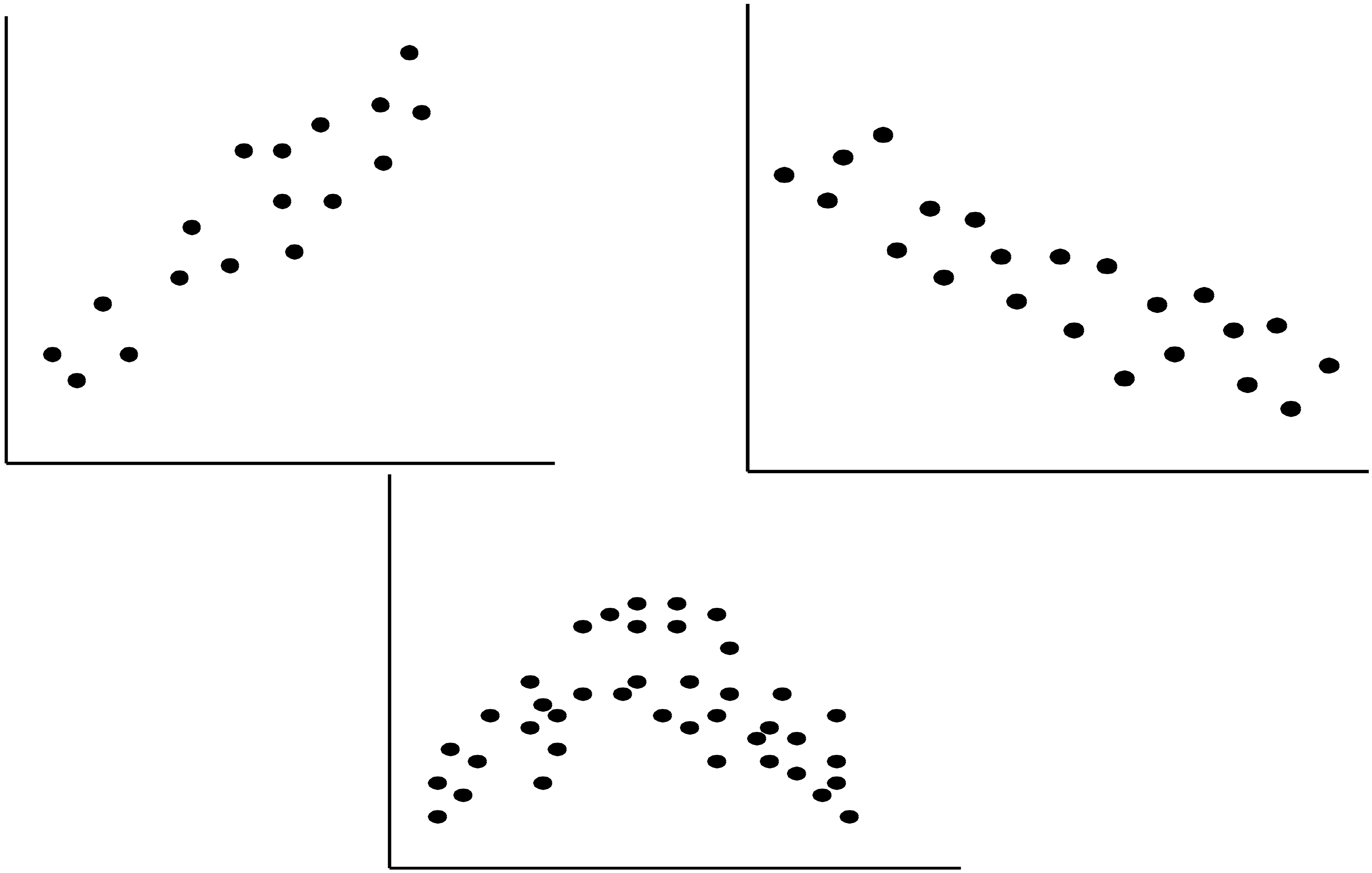


Histogram Analysis

- Graph displays of basic statistical class descriptions
  - Frequency histograms
    - A univariate graphical method
    - Consists of a set of rectangles that reflect the counts or frequencies of the classes present in the given data



Positively and Negatively Correlated Data



Data Cleaning

- Importance
  - "Data cleaning is one of the three biggest problems in data warehousing"—Ralph Kimball
  - "Data cleaning is the number one problem in data warehousing"—DCI survey
- 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 disebabkan oleh:
  - 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

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably.
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., "unknown", a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: hasil dari decision tree (klasifikasi)

Noisy Data

- **Incorrect attribute** dapat disebabkan oleh
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
- Other data problems which requires data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

Bagaimana Mengatasi Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Simple Discretization Methods: Binning

- Equal-width (distance) partitioning
  - 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
  - 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 equal-frequency (equi-depth) bins:

  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34

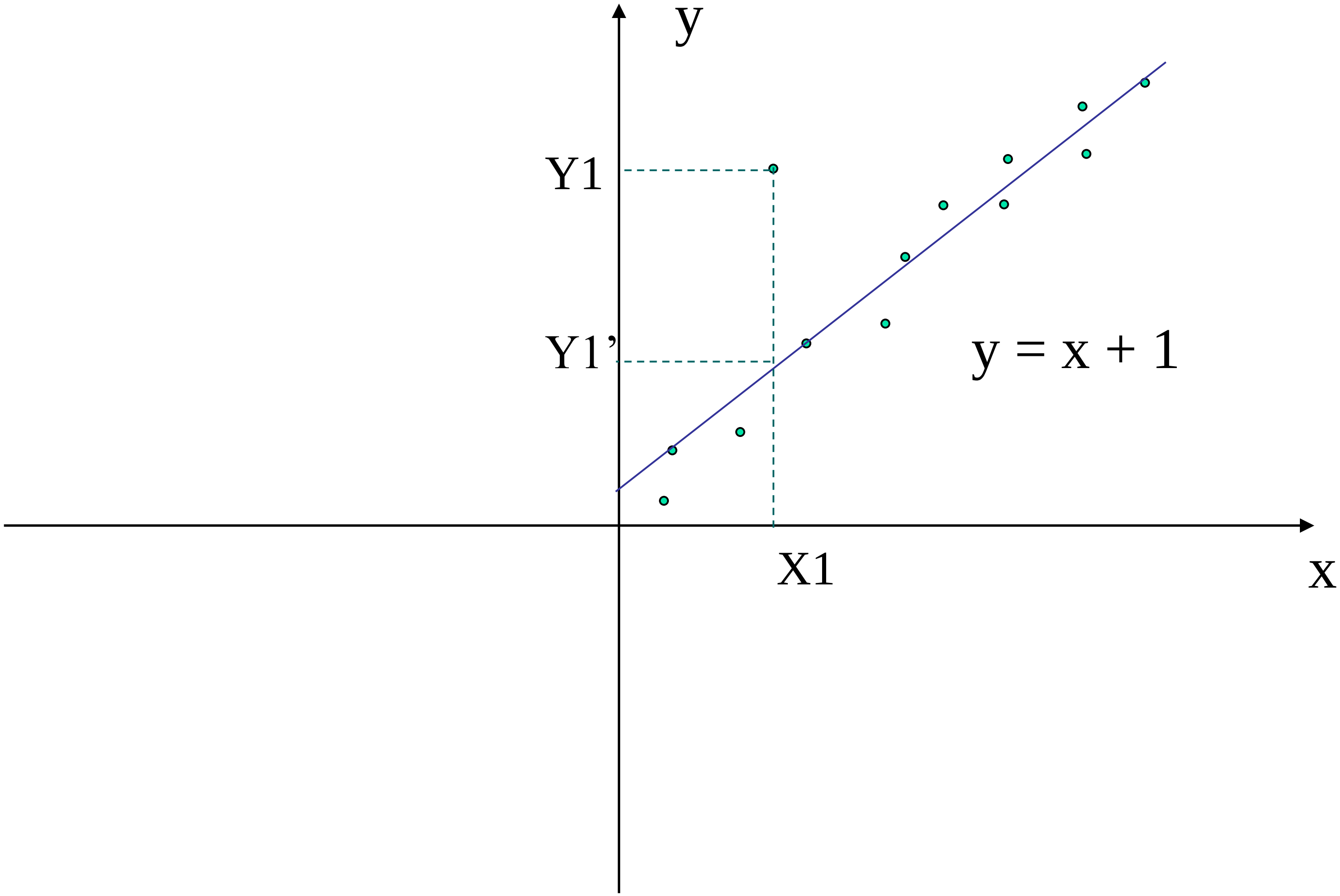
* Smoothing by bin means:

  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29

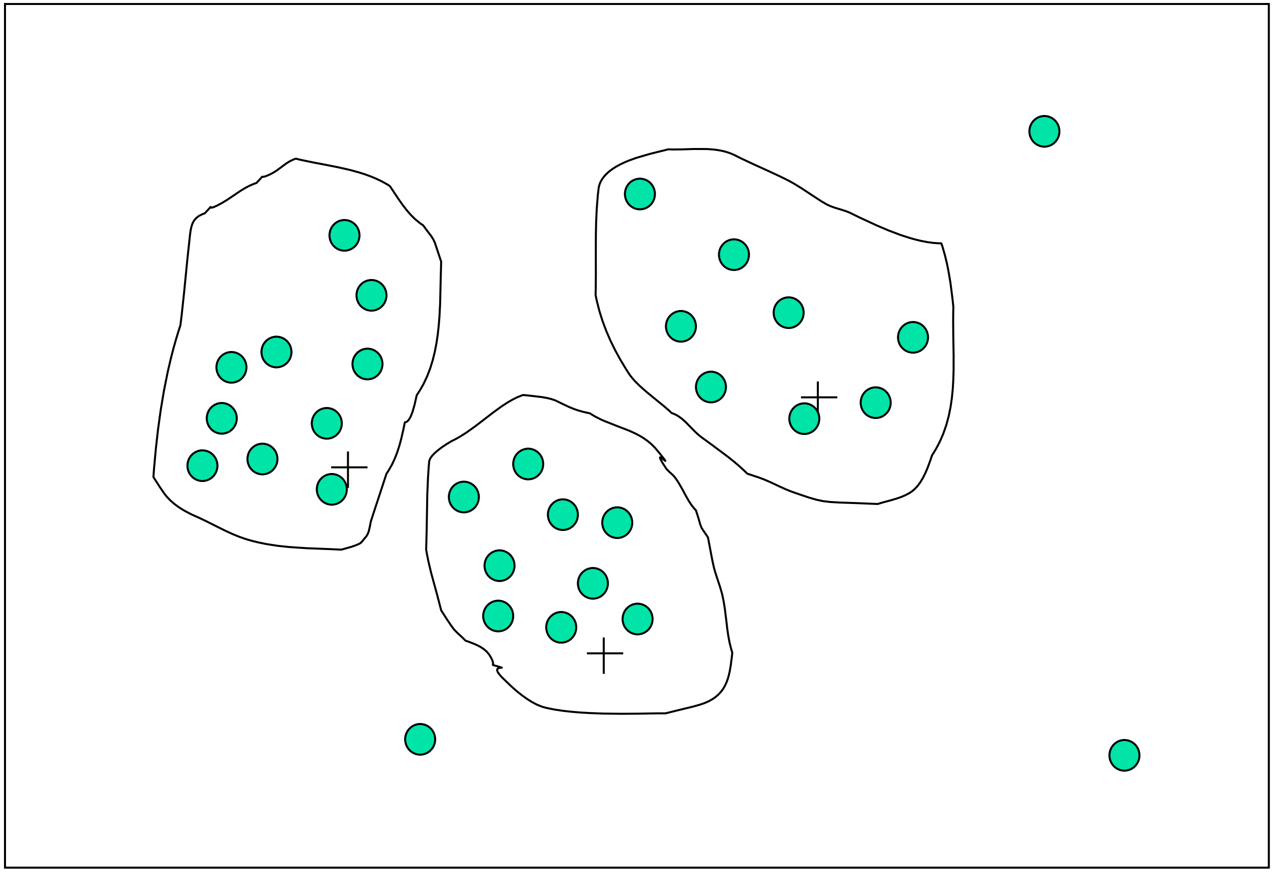
* Smoothing by bin boundaries:

  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Regresi



Pengelompokan Data



Data Integration

- Data integration:
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id $\equiv$ B.cust-#
  - Integrate metadata from different sources
- Entity identification problem:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Mengatasi Redudansi saat Data Integrasi

- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data:* One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by *correlation analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Analisa Korelasi untuk Data Numerik

- Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum (A - \bar{A})(B - \bar{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum (AB) - n\bar{A}\bar{B}}{(n-1)\sigma_A \sigma_B}$$

where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and Σ(AB) is the sum of the AB cross-product.

- If $r_{A,B} > 0$, A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.
- $r_{A,B} = 0$: independent; $r_{A,B} < 0$: negatively correlated

Transformasi Data

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Attribute/feature construction
  - New attributes constructed from the given ones

Data Transformation: Normalization

- Min-max normalization: to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73,600 - 12,000}{98,000 - 12,000}(1.0 - 0) + 0 = 0.716$
- Z-score normalization (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73,600 - 54,000}{16,000} = 1.225$
- Normalization by decimal scaling

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max(|ν'|) < 1

Referensi

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. Communications of ACM, 42:73-78, 1999
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003
- T. Dasu, T. Johnson, S. Muthukrishnan, V. Shkapenyuk. Mining Database Structure; Or, How to Build a Data Quality Browser. SIGMOD'02.
- H.V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Technical Committee on Data Engineering, 20(4), December 1997
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- E. Rahm and H. H. Do. Data Cleaning: Problems and Current Approaches. *IEEE Bulletin of the Technical Committee on Data Engineering. Vol.23, No.4*
- V. Raman and J. Hellerstein. Potters Wheel: An Interactive Framework for Data Cleaning and Transformation, VLDB'2001
- T. Redman. Data Quality: Management and Technology. Bantam Books, 1992
- Y. Wand and R. Wang. Anchoring data quality dimensions ontological foundations. Communications of ACM, 39:86-95, 1996
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. IEEE Trans. Knowledge and Data Engineering, 7:623-640, 1995