

# Data Warehousing

## 資料倉儲

### Data Preprocessing: Integration and the ETL process

992DW03

MI4

Tue. 8,9 (15:10-17:00) L413

Min-Yuh Day

戴敏育

Assistant Professor

專任助理教授

Dept. of Information Management, Tamkang University

淡江大學 資訊管理學系

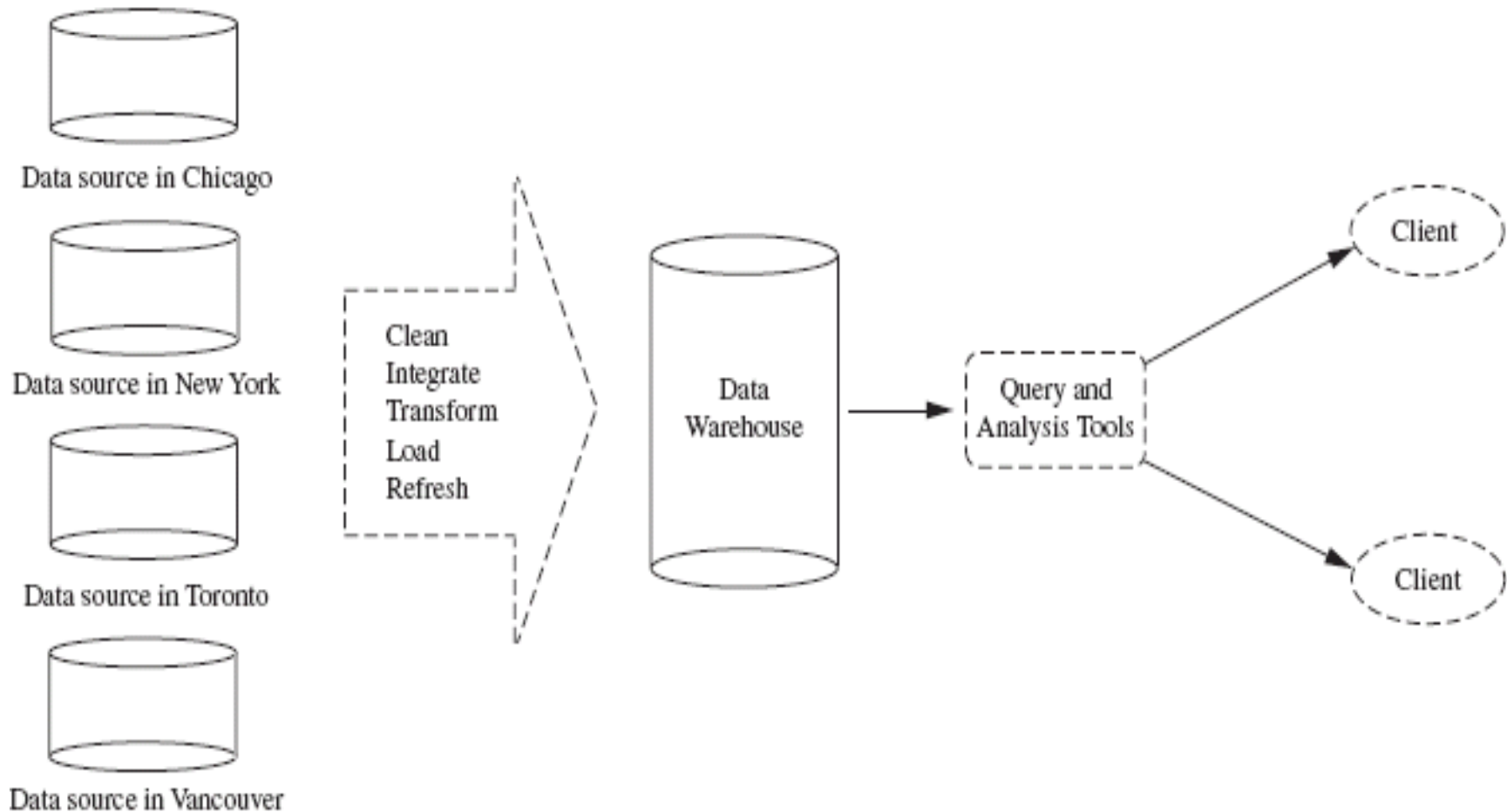
<http://mail.im.tku.edu.tw/~myday/>

2011-03-01

# Syllabus

- 1 100/02/15 Introduction to Data Warehousing
- 2 100/02/22 Data Warehousing, Data Mining, and Business Intelligence
- 3 100/03/01 Data Preprocessing: Integration and the ETL process
- 4 100/03/08 Data Warehouse and OLAP Technology
- 5 100/03/15 Data Cube Computation and Data Generation
- 6 100/03/22 Association Analysis
- 7 100/03/29 Classification and Prediction
- 8 100/04/05 (放假一天) (民族掃墓節)
- 9 100/04/12 Cluster Analysis
- 10 100/04/19 Mid Term Exam (期中考試週)
- 11 100/04/26 Sequence Data Mining
- 12 100/05/03 Social Network Analysis and Link Mining
- 13 100/05/10 Text Mining and Web Mining
- 14 100/05/17 Project Presentation
- 15 100/05/24 Final Exam (畢業班考試)

# Typical framework of a data warehouse



# ETL

- Extraction
- Transformation
- Loading

# Relational Database

*customer*

<u>cust_ID</u>	<i>name</i>	<i>address</i>	<i>age</i>	<i>income</i>	<i>credit_info</i>	<i>category</i>	...
C1	Smith, Sandy	1223 Lake Ave., Chicago, IL	31	\$78000	1	3	...
...	...	...	...	...	...	...	...

*item*

<u>item_ID</u>	<i>name</i>	<i>brand</i>	<i>category</i>	<i>type</i>	<i>price</i>	<i>place_made</i>	<i>supplier</i>	<i>cost</i>
I3	hi-res-TV	Toshiba	high resolution	TV	\$988.00	Japan	NikoX	\$600.00
I8	Laptop	Dell	laptop	computer	\$1369.00	USA	Dell	\$983.00
...	...	...	...	...	...	...	...	...

*employee*

<u>empl_ID</u>	<i>name</i>	<i>category</i>	<i>group</i>	<i>salary</i>	<i>commission</i>
E55	Jones, Jane	home entertainment	manager	\$118,000	2%
...	...	...	...	...	...

*branch*

<u>branch_ID</u>	<i>name</i>	<i>address</i>
B1	City Square	396 Michigan Ave., Chicago, IL
...	...	...

*purchases*

<u>trans_ID</u>	<i>cust_ID</i>	<i>empl_ID</i>	<i>date</i>	<i>time</i>	<i>method_paid</i>	<i>amount</i>
T100	C1	E55	03/21/2005	15:45	Visa	\$1357.00
...	...	...	...	...	...	...

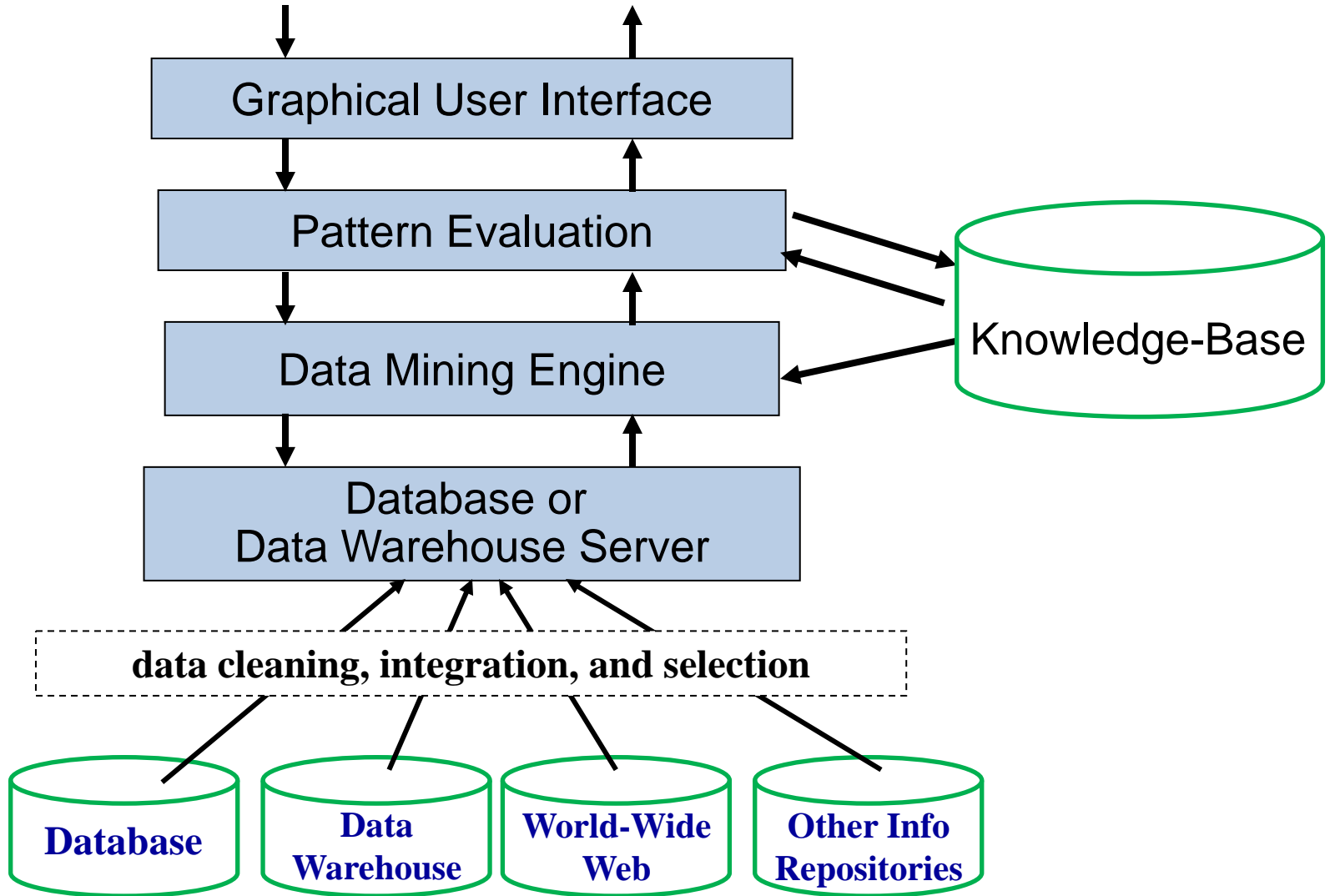
*items\_sold*

<u>trans_ID</u>	<u>item_ID</u>	<i>qty</i>
T100	I3	1
T100	I8	2
...	...	...

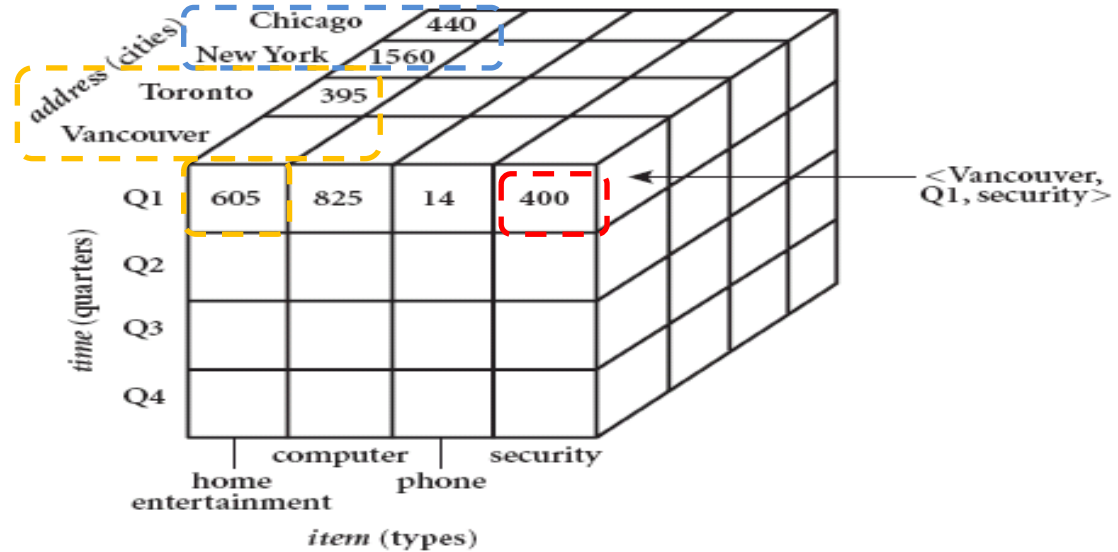
*works\_at*

<u>empl_ID</u>	<u>branch_ID</u>
E55	B1
...	...

# Architecture of a typical data mining system



# Multidimensional data cube for data warehousing



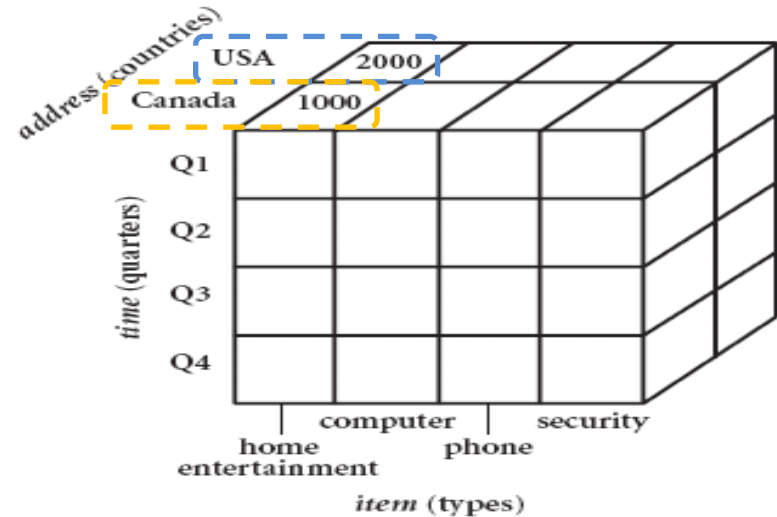
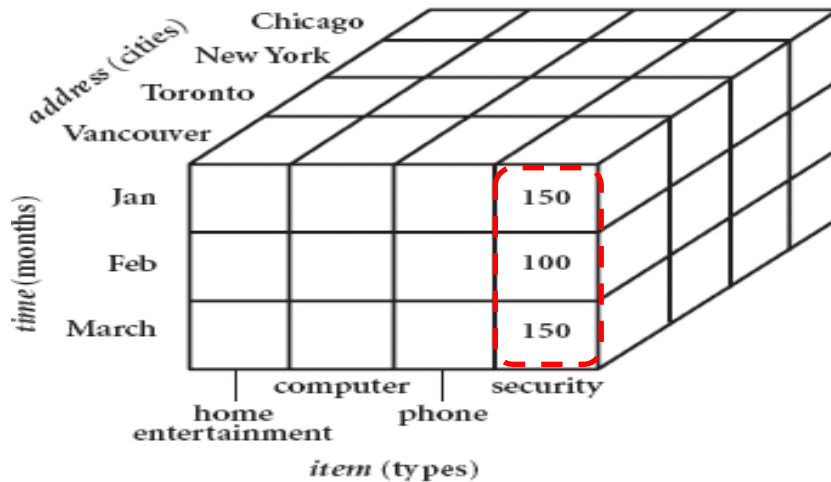
(b)

**Drill-down**

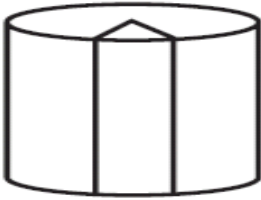
Drill-down  
on time data for Q1

Roll-up  
on address

**Roll-up**



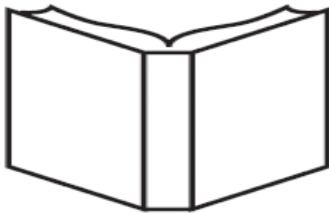
# Primitives for specifying a data mining task



Task-relevant data  
Database or data warehouse name  
Database tables or data warehouse cubes  
Conditions for data selection  
Relevant attributes or dimensions  
Data grouping criteria



Knowledge type to be mined  
Characterization  
Discrimination  
Association/correlation  
Classification/prediction  
Clustering



Background knowledge  
Concept hierarchies  
User beliefs about relationships in the data



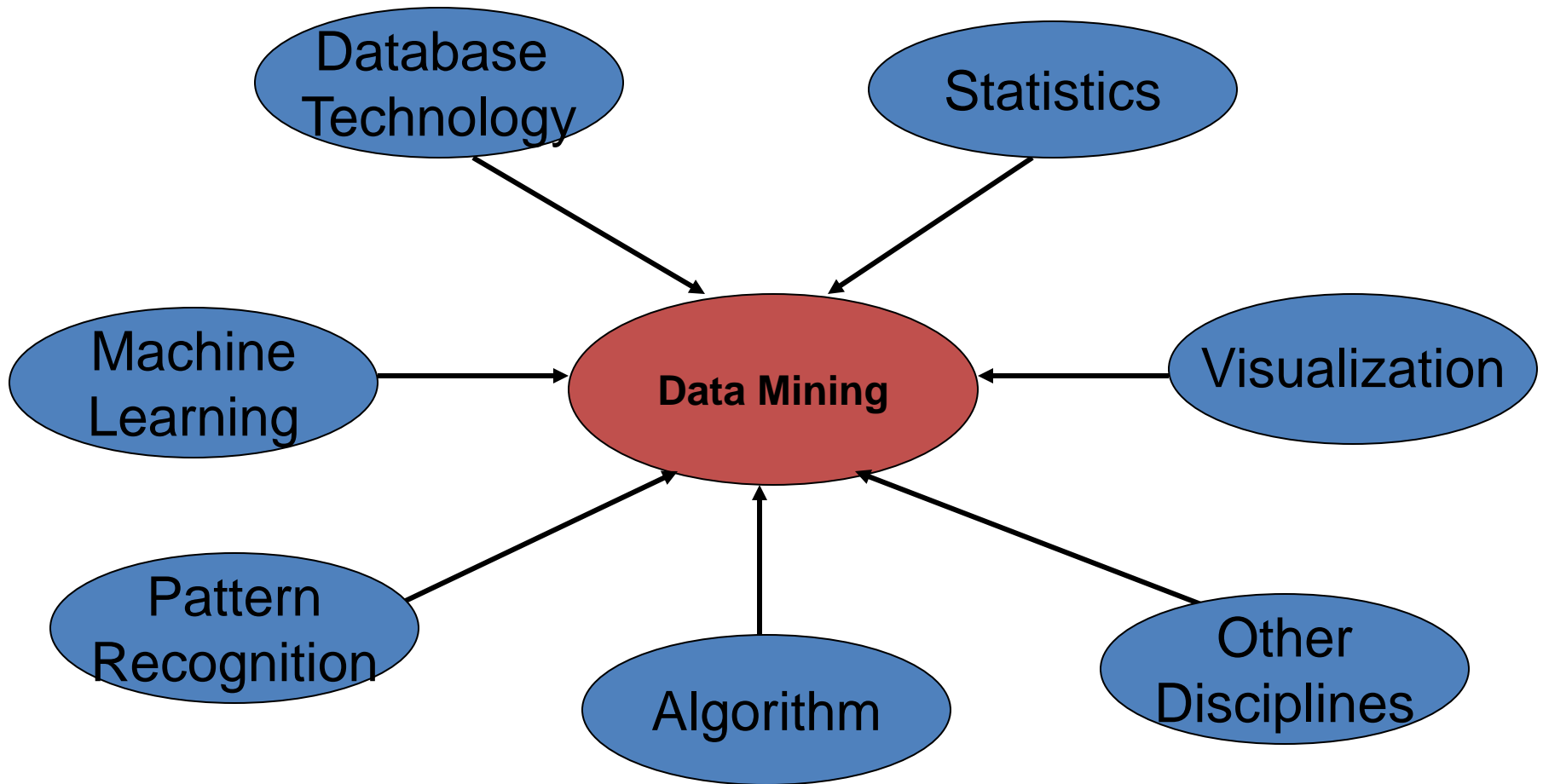
Pattern interestingness measures  
Simplicity  
Certainty (e.g., confidence)  
Utility (e.g., support)  
Novelty



Visualization of discovered patterns  
Rules, tables, reports, charts, graphs, decision trees,  
and cubes  
Drill-down and roll-up



# Data Mining: Confluence of Multiple Disciplines



# Differences between a data warehouse and a database

- Data warehouse:
  - A data warehouse is a repository of information collected from **multiple sources** over a **history of time** stored under a **unified schema** and used for **data analysis** and **decision support**
  - There could be multiple heterogeneous databases where the schema of one database may not agree with the schema of another.
- Database:
  - A database is a collection of **interrelated data** that represents the current status of the stored data.
  - A database system supports ad-hoc query and on-line transaction processing.

# Similarities between a data warehouse and a database

- Both are repositories of information storing huge amounts of persistent data.

# Why Data Preprocessing?

- Data in the real world is dirty
  - **incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=""
  - **noisy**: containing errors or outliers
    - e.g., Salary="-10"
  - **inconsistent**: containing discrepancies in codes or names
    - e.g., Age="42" Birthday="03/07/1997"
    - e.g., Was rating "1,2,3", now rating "A, B, C"
    - e.g., discrepancy between duplicate records

# Why Is Data Dirty?

- Incomplete data may come from
  - “Not applicable” data value when collected
  - Different considerations between the time when the data was collected and when it is analyzed.
  - Human/hardware/software problems
- Noisy data (incorrect values) may come from
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- Inconsistent data may come from
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)
- Duplicate records also need data cleaning

# Why Is Data Preprocessing Important?

- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
  - Data warehouse needs consistent integration of quality data
- Data **extraction, cleaning, and transformation** comprises the majority of the work of building a data warehouse

# Multi-Dimensional Measure of Data Quality

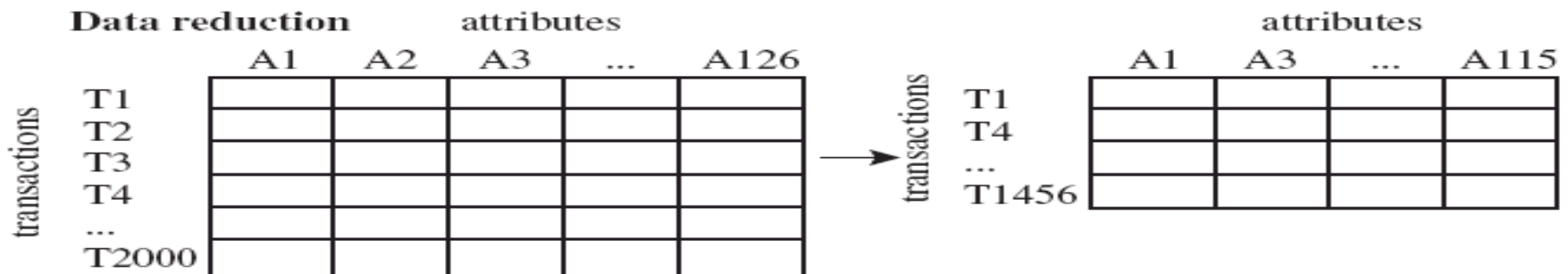
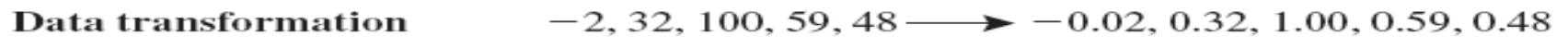
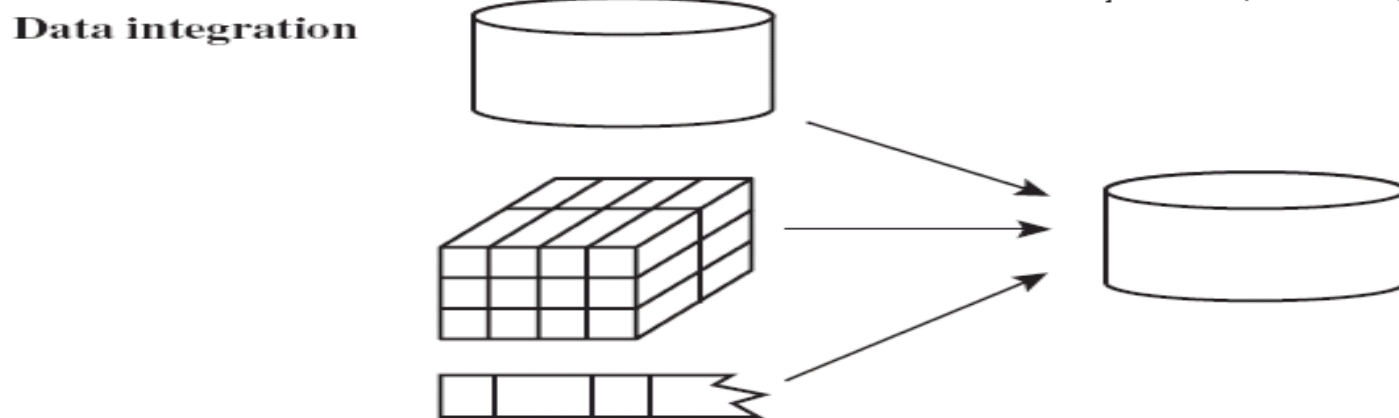
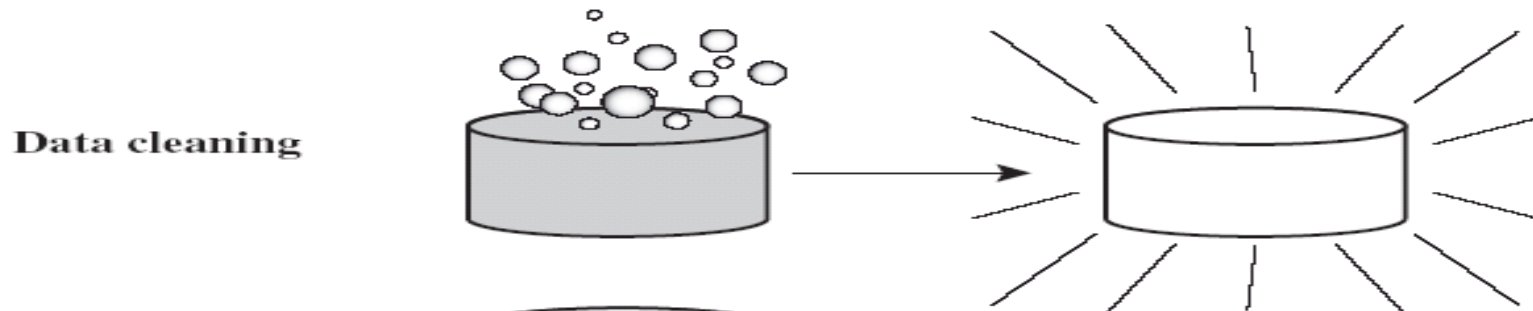
- A well-accepted multidimensional view:
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Value added
  - Interpretability
  - Accessibility
- Broad categories:
  - Intrinsic, contextual, representational, and accessibility

# 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 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



# Forms of Data Preprocessing



# Mining Data Descriptive Characteristics

- Motivation
  - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
  - Data dispersion: analyzed with multiple granularities of precision
  - Boxplot or quantile analysis on sorted intervals
- Dispersion analysis on computed measures
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube

# Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):

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

- Weighted arithmetic mean:
- Trimmed mean: chopping extreme values

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

- Median: A holistic measure

- Middle value if odd number of values, or average of the middle two values otherwise
- Estimated by interpolation (for *grouped data*):

$$median = L_1 + \left( \frac{n/2 - (\sum f)l}{f_{median}} \right) c$$

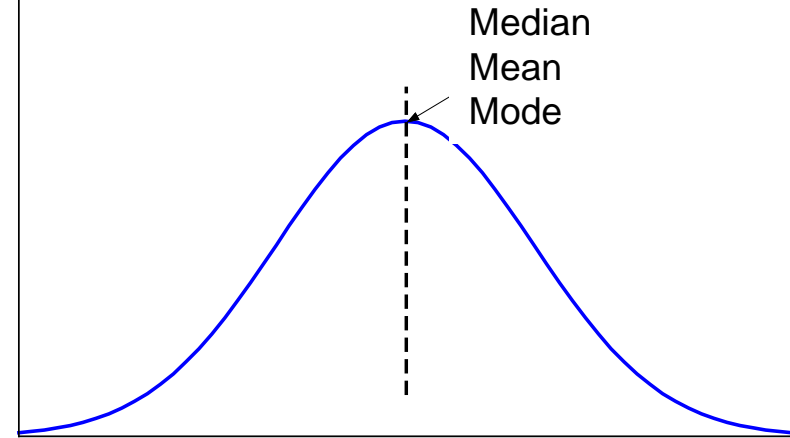
- Mode

- Value that occurs most frequently in the data
- Unimodal, bimodal, trimodal
- Empirical formula:

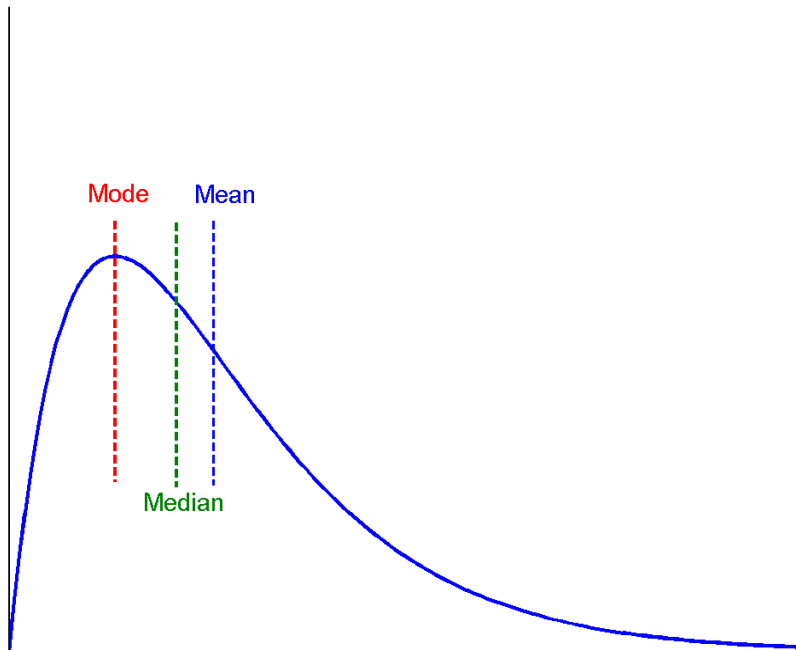
$$mean - mode = 3 \times (mean - median)$$

# Symmetric vs. Skewed Data

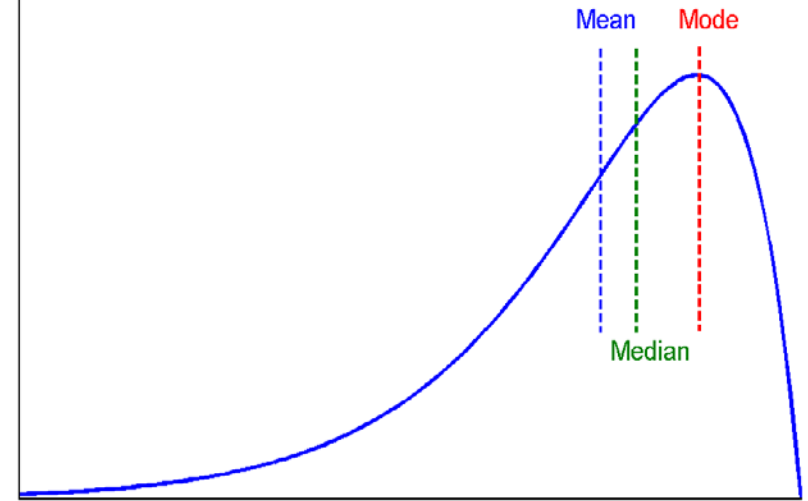
- Median, mean and mode of symmetric, positively and negatively skewed data



Symmetric data



Positively skewed data



Negatively skewed data

# Measuring the Dispersion of Data

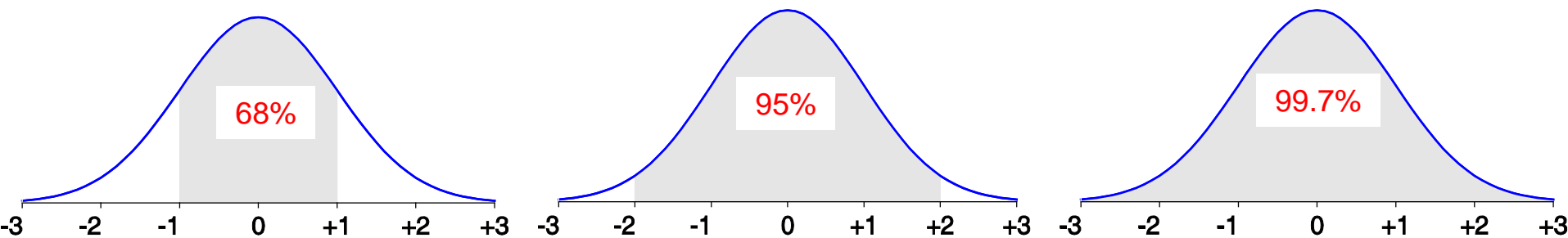
- Quartiles, outliers and boxplots
  - **Quartiles**:  $Q_1$  (25<sup>th</sup> percentile),  $Q_3$  (75<sup>th</sup> percentile)
  - **Inter-quartile range**:  $IQR = Q_3 - Q_1$
  - **Five number summary**: min,  $Q_1$ , M,  $Q_3$ , max
  - **Boxplot**: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually
  - **Outlier**: usually, a value higher/lower than  $1.5 \times IQR$
- Variance and standard deviation (*sample:  $s$ , population:  $\sigma$* )
  - **Variance**: (algebraic, scalable computation)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[ \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \right] \quad \sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^n x_i^2 - \mu^2$$

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

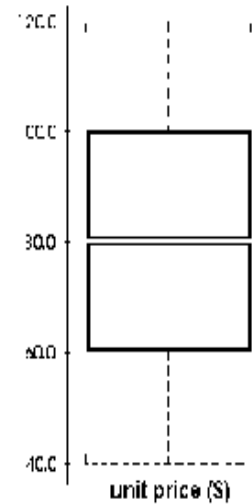
# Properties of Normal Distribution Curve

- The normal (distribution) curve
  - From  $\mu - \sigma$  to  $\mu + \sigma$ : contains about 68% of the measurements ( $\mu$ : mean,  $\sigma$ : standard deviation)
  - From  $\mu - 2\sigma$  to  $\mu + 2\sigma$ : contains about 95% of it
  - From  $\mu - 3\sigma$  to  $\mu + 3\sigma$ : contains about 99.7% of it

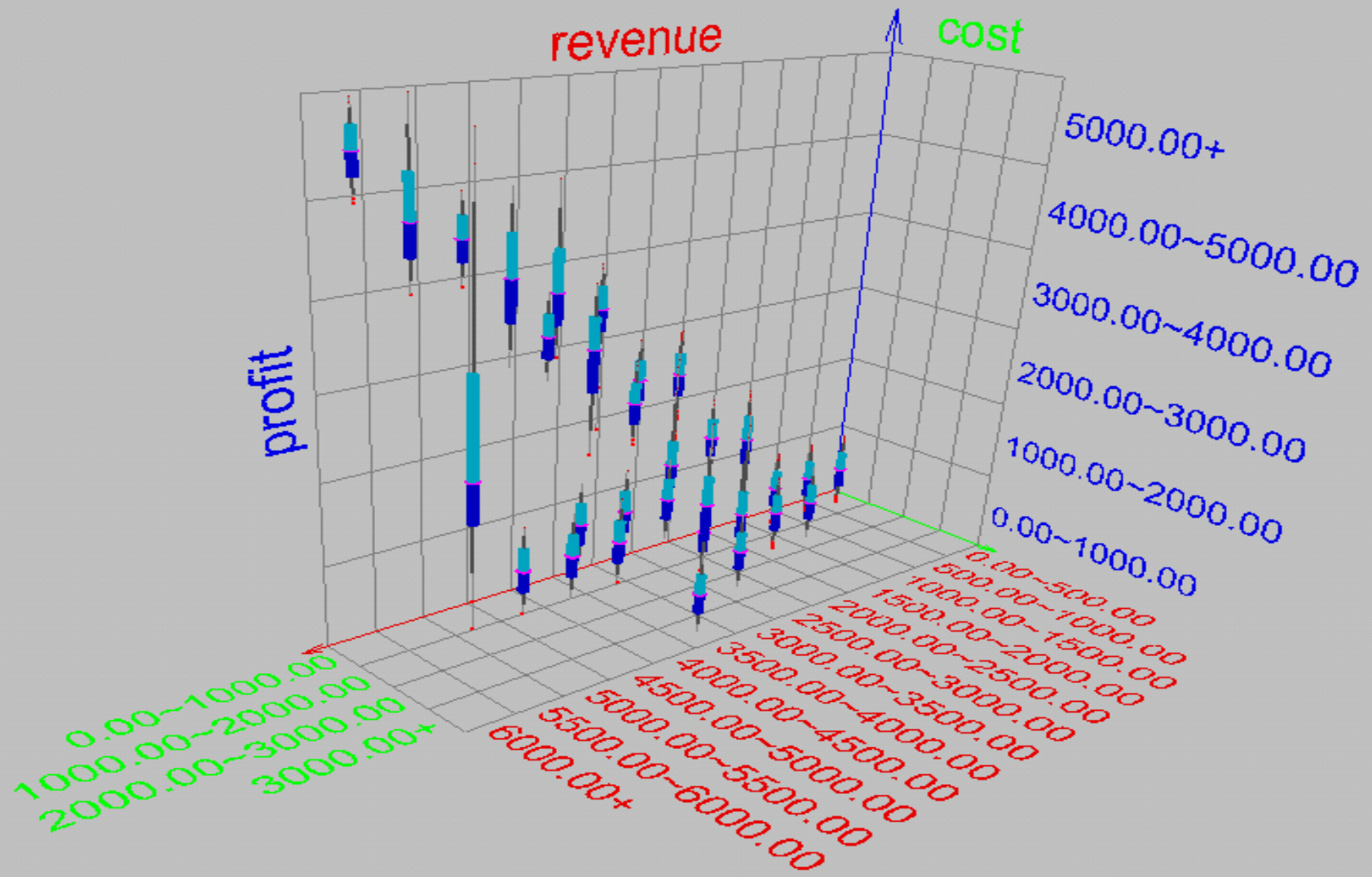


# Boxplot Analysis

- **Five-number summary** of a distribution:  
Minimum, Q1, M, Q3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IRQ
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extend to Minimum and Maximum



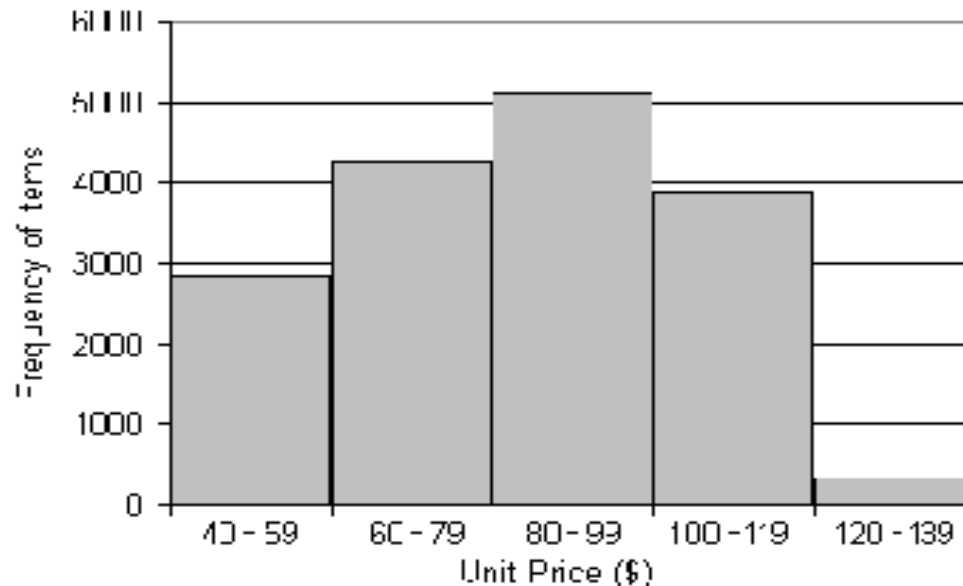
# Visualization of Data Dispersion: Boxplot Analysis





# 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



# Graphic Displays of Basic Statistical Descriptions

- Histogram: (shown before)
- Boxplot: (covered before)
- Quantile plot: each value  $x_i$  is paired with  $f_i$  indicating that approximately  $100 f_i \%$  of data are  $\leq x_i$
- Quantile-quantile (q-q) plot: graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- Scatter plot: each pair of values is a pair of coordinates and plotted as points in the plane
- Loess (local regression) curve: add a smooth curve to a scatter plot to provide better perception of the pattern of dependence

# 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 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
- Fill in the missing value manually
- 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: inference-based such as Bayesian formula or decision tree

# Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may 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?

## 1. 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.

## 2. Regression

- smooth by fitting the data into regression functions

## 3. 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:

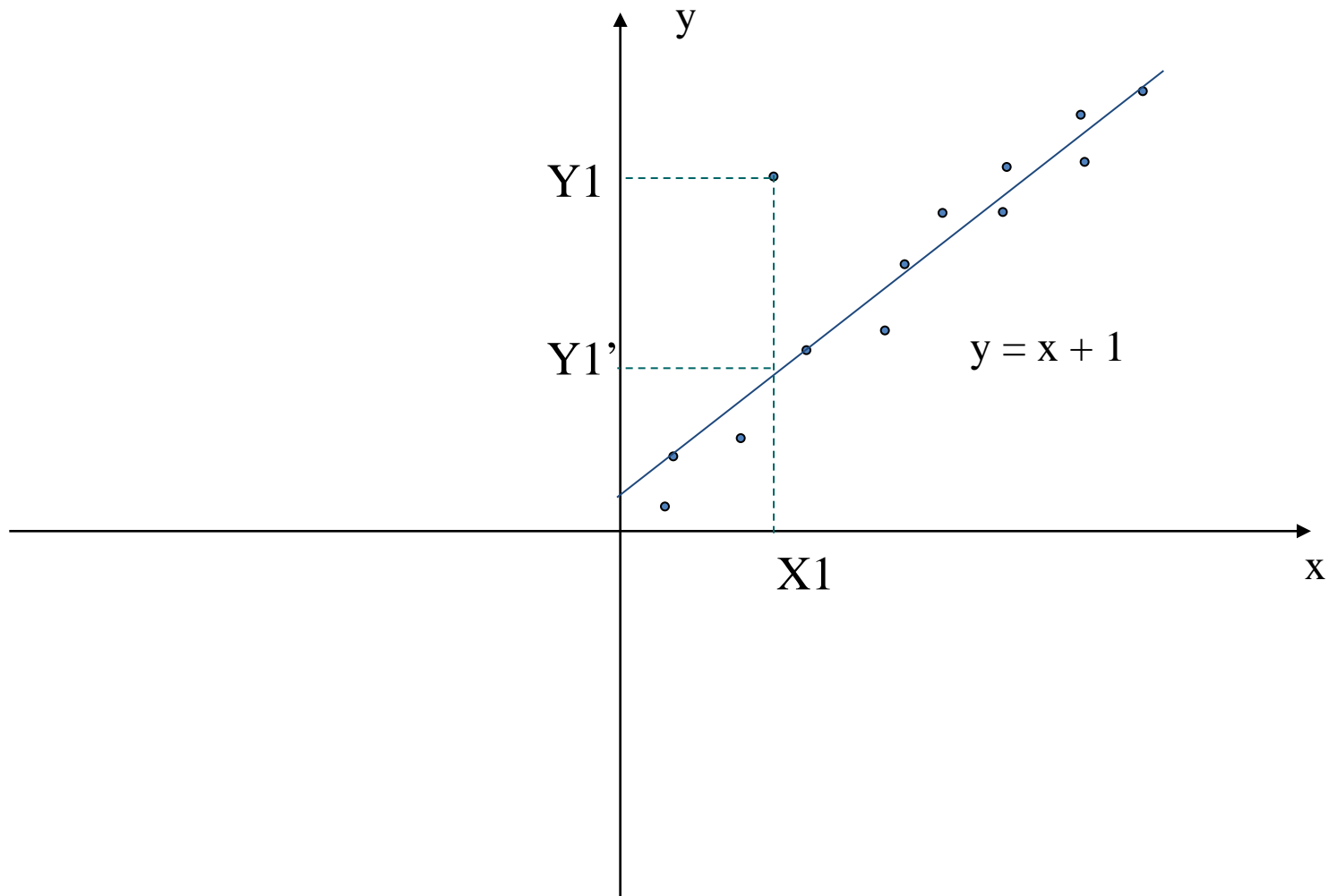
## Binning

- **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

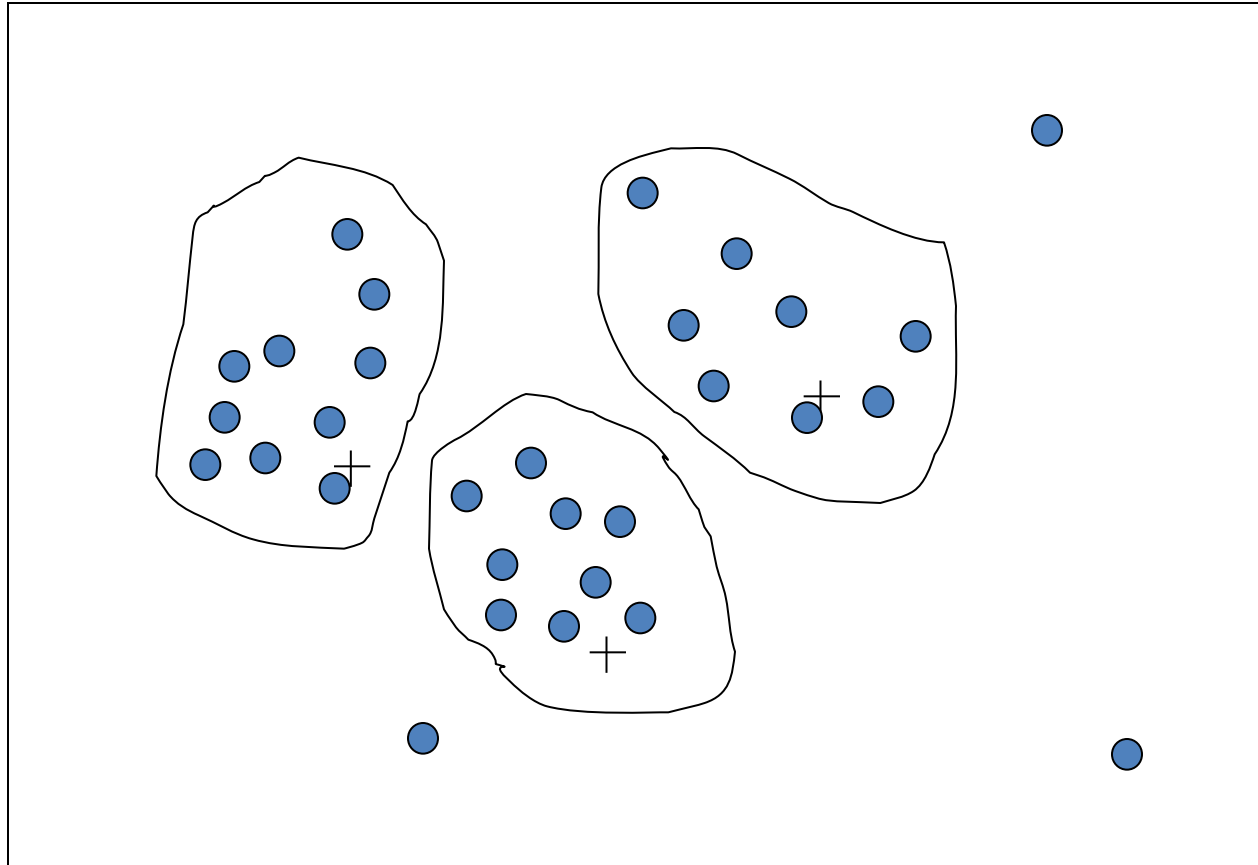
# Binning Methods for Data Smoothing

- 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

# Regression



# Cluster Analysis



# Data Cleaning as a Process

- Data discrepancy detection
  - Use metadata
    - (e.g., domain, range, dependency, distribution)
  - Check field overloading
  - Check uniqueness rule, consecutive rule and null rule
  - Use commercial tools
    - Data scrubbing
      - use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
    - Data auditing
      - by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)

# Data Cleaning as a Process

- Data migration and integration
  - Data migration tools
    - allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools
    - allow users to specify transformations through a graphical user interface
- Integration of the two processes
  - Iterative and interactive (e.g., Potter's Wheels)

# 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

# Handling Redundancy in Data Integration

- 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



# Correlation Analysis (Numerical Data)

- 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  $B$ ,  $\sigma_A$  and  $\sigma_B$  are the respective standard deviation of  $A$  and  $B$ , and  $\sum(AB)$  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

# Correlation Analysis (Categorical Data)

- $\chi^2$  (chi-square) test

$$\chi^2 = \sum \frac{(\textit{Observed} - \textit{Expected})^2}{\textit{Expected}}$$

- The larger the  $\chi^2$  value, the more likely the variables are related
- The cells that contribute the most to the  $\chi^2$  value are those whose actual count is very different from the expected count
- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

# Chi-Square Calculation: An Example

$$e_{11} = \frac{300 \times 450}{1500} = 90$$

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- $\chi^2$  (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{840} = 507.93$$

- It shows that like\_science\_fiction and play\_chess are correlated in the group

# Data Transformation

- 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 ( $\mu$ : mean,  $\sigma$ : standard deviation):

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

- Ex. Let  $\mu = 54,000$ ,  $\sigma = 16,000$ . Then  $\frac{73,600 - 54,000}{16,000} = 1.225$

- Normalization by decimal scaling

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

# Why data reduction?

- A database/data warehouse may store terabytes of data
- Complex data analysis/mining may take a very long time to run on the complete data set

# Data reduction

- Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results

# Data Reduction Strategies

- Data cube aggregation:
- Dimensionality reduction
  - e.g., remove unimportant attributes
- Data Compression
- Numerosity reduction
  - e.g., fit data into models
- Discretization and concept hierarchy generation



# Data Cube Aggregation

- The lowest level of a data cube (base cuboid)
  - The aggregated data for an **individual entity of interest**
  - E.g., a customer in a phone calling data warehouse
- Multiple levels of aggregation in data cubes
  - Further reduce the size of data to deal with
- Reference appropriate levels
  - Use the smallest representation which is enough to solve the task
- Queries regarding aggregated information should be answered using data cube, when possible

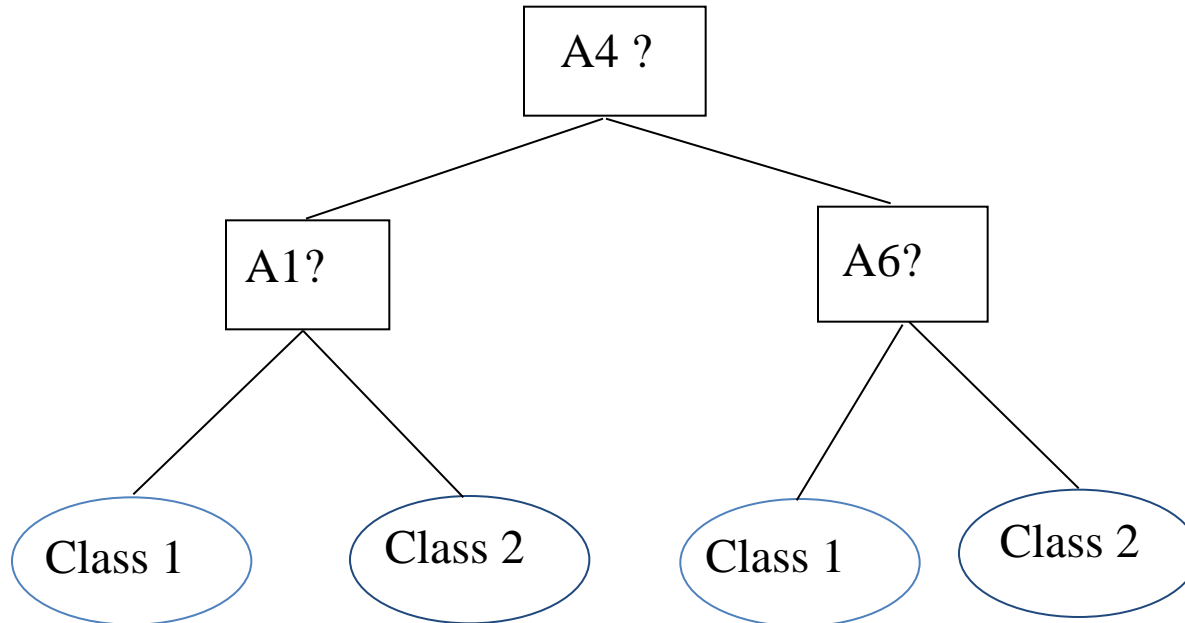
# Attribute Subset Selection

- Feature selection (i.e., attribute subset selection):
  - Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
  - reduce # of patterns in the patterns, easier to understand
- Heuristic methods (due to exponential # of choices):
  - Step-wise forward selection
  - Step-wise backward elimination
  - Combining forward selection and backward elimination
  - Decision-tree induction

# Example of Decision Tree Induction

Initial attribute set:

{A1, A2, A3, A4, A5, A6}



-----> Reduced attribute set: {A1, A4, A6}

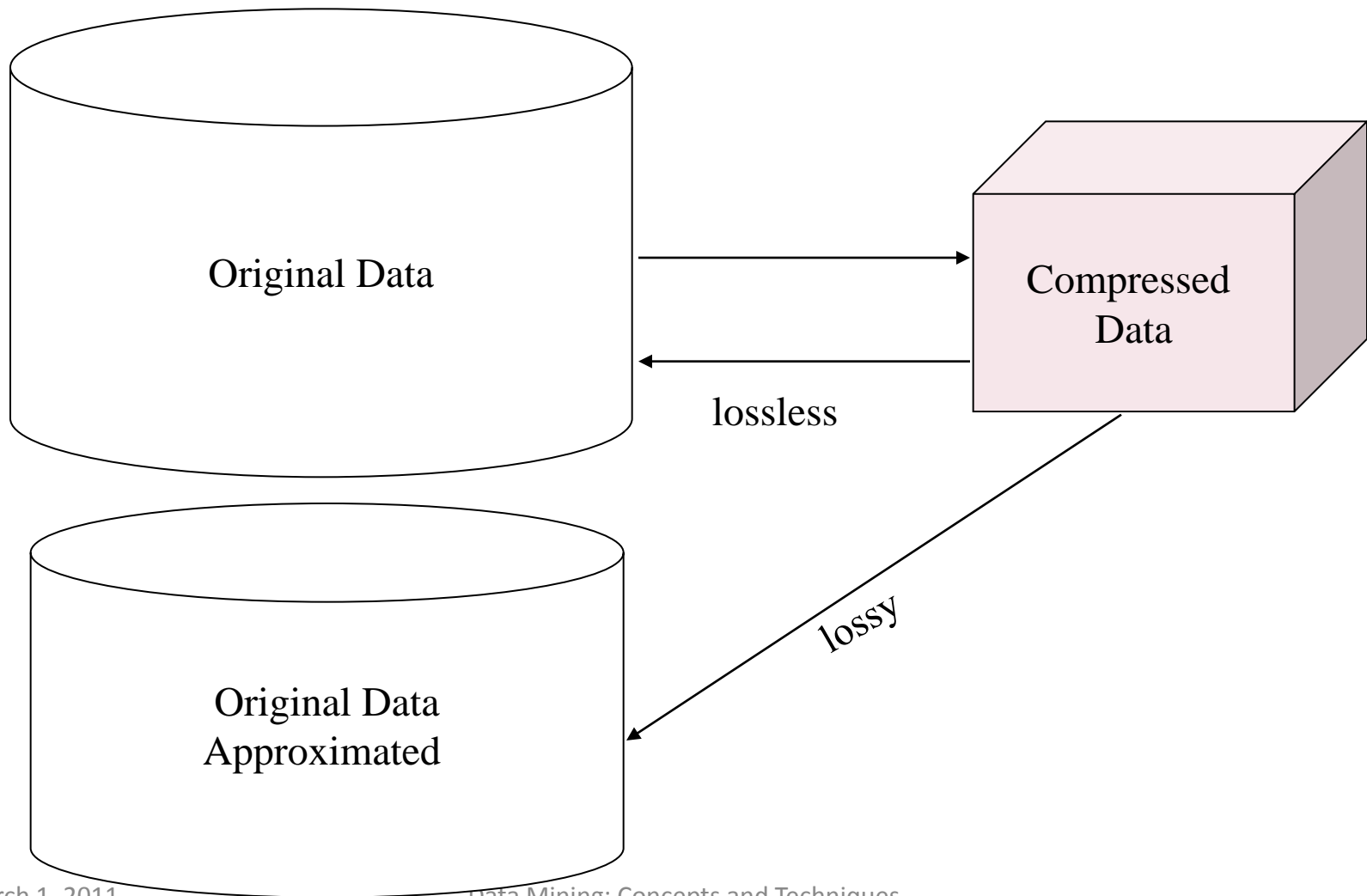
# Heuristic Feature Selection Methods

- There are  $2^d$  possible sub-features of  $d$  features
- Several heuristic feature selection methods:
  - Best single features under the feature independence assumption: choose by significance tests
  - Best step-wise feature selection:
    - The best single-feature is picked first
    - Then next best feature condition to the first, ...
  - Step-wise feature elimination:
    - Repeatedly eliminate the worst feature
  - Best combined feature selection and elimination
  - Optimal branch and bound:
    - Use feature elimination and backtracking

# Data Compression

- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically lossless
  - But only limited manipulation is possible without expansion
- Audio/video compression
  - Typically lossy compression, with progressive refinement
  - Sometimes small fragments of signal can be reconstructed without reconstructing the whole
- Time sequence is not audio
  - Typically short and vary slowly with time

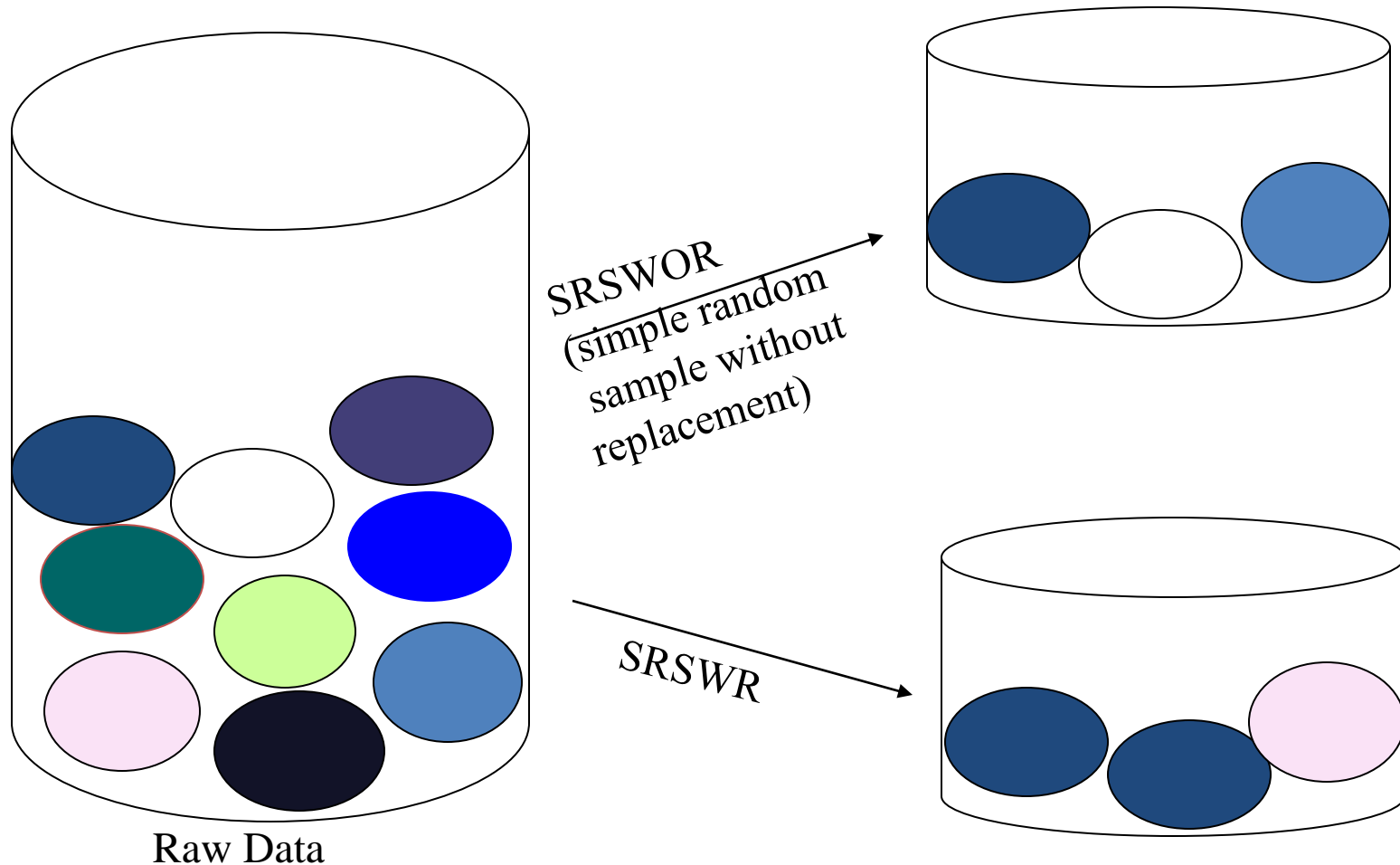
# Data Compression



# Data Reduction Method: Sampling

- Sampling: obtaining a small sample  $s$  to represent the whole data set  $N$
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Choose a **representative** subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
- Develop adaptive sampling methods
  - Stratified sampling:
    - Approximate the percentage of each class (or subpopulation of interest) in the overall database
    - Used in conjunction with skewed data
- Note: Sampling may not reduce database I/Os (page at a time)

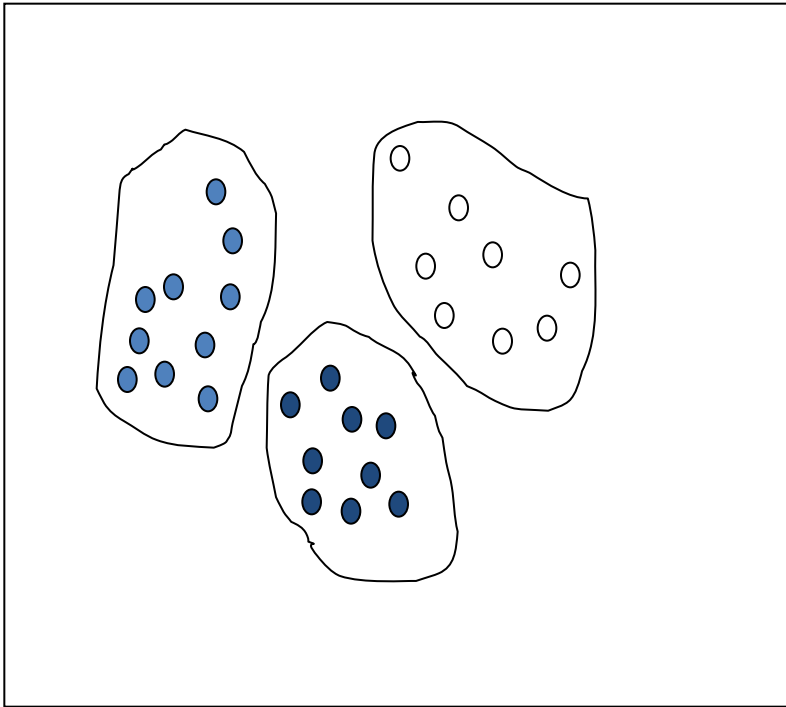
# Sampling: with or without Replacement



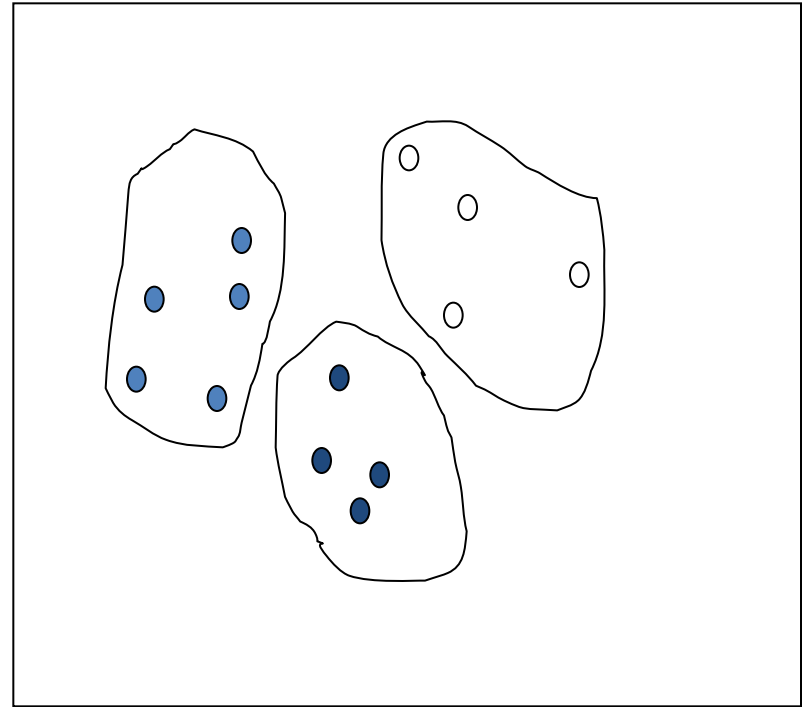


# Sampling: Cluster or Stratified Sampling

Raw Data



Cluster/Stratified Sample



# Discretization

- Three types of attributes:
  - Nominal — values from an unordered set, e.g., color, profession
  - Ordinal — values from an ordered set, e.g., military or academic rank
  - Continuous — real numbers, e.g., integer or real numbers
- Discretization:
  - Divide the range of a continuous attribute into intervals
  - Some classification algorithms only accept categorical attributes.
  - Reduce data size by discretization
  - Prepare for further analysis

# Discretization and Concept Hierarchy

- Discretization
  - Reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute
- Concept hierarchy formation
  - Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for age) by higher level concepts (such as young, middle-aged, or senior)

# Discretization and Concept Hierarchy Generation for Numeric Data

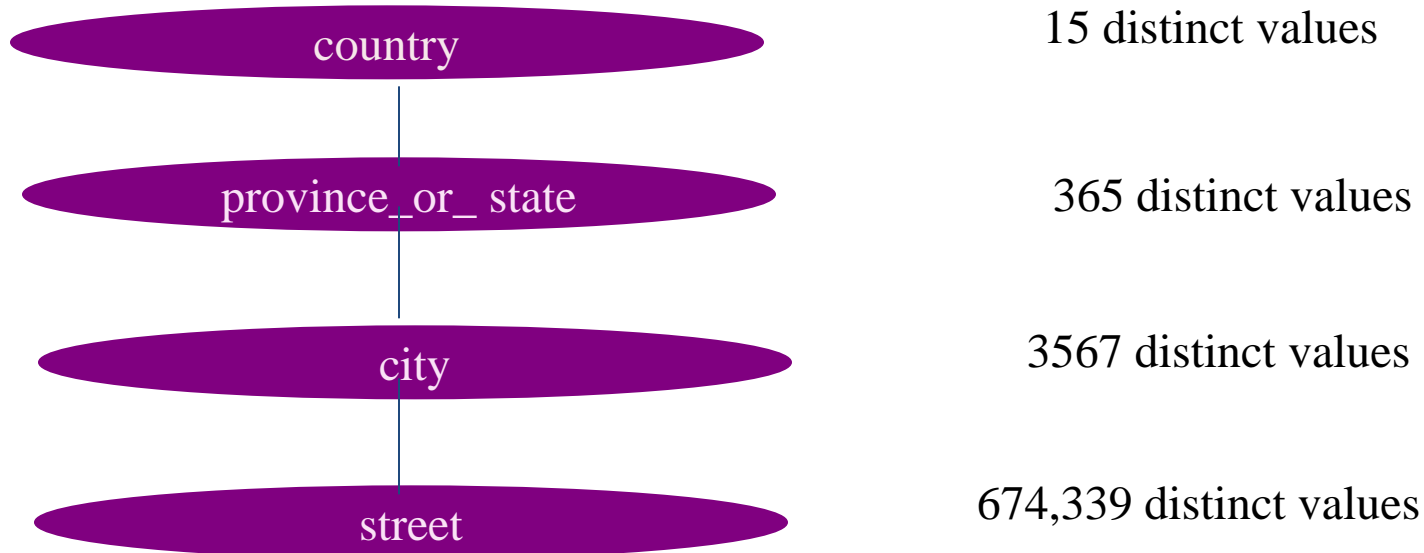
- Typical methods: All the methods can be applied recursively
  - Binning (covered above)
    - Top-down split, unsupervised,
  - Histogram analysis (covered above)
    - Top-down split, unsupervised
  - Clustering analysis (covered above)
    - Either top-down split or bottom-up merge, unsupervised
  - Entropy-based discretization: supervised, top-down split
  - Interval merging by  $\chi^2$  Analysis: unsupervised, bottom-up merge
  - Segmentation by natural partitioning: top-down split, unsupervised

# Concept Hierarchy Generation for Categorical Data

- Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
  - street < city < state < country
- Specification of a hierarchy for a set of values by explicit data grouping
  - {Urbana, Champaign, Chicago} < Illinois
- Specification of only a partial set of attributes
  - E.g., only street < city, not others
- Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values
  - E.g., for a set of attributes: {street, city, state, country}

# Automatic Concept Hierarchy Generation

- Some hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set
  - The attribute with the most distinct values is placed at the lowest level of the hierarchy
  - Exceptions, e.g., weekday, month, quarter, year



# Summary

- Data preparation or preprocessing is a big issue for both data warehousing and data mining
- Descriptive data summarization is need for quality data preprocessing
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- A lot a methods have been developed but data preprocessing still an active area of research

# References

- Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition, 2006, Elsevier