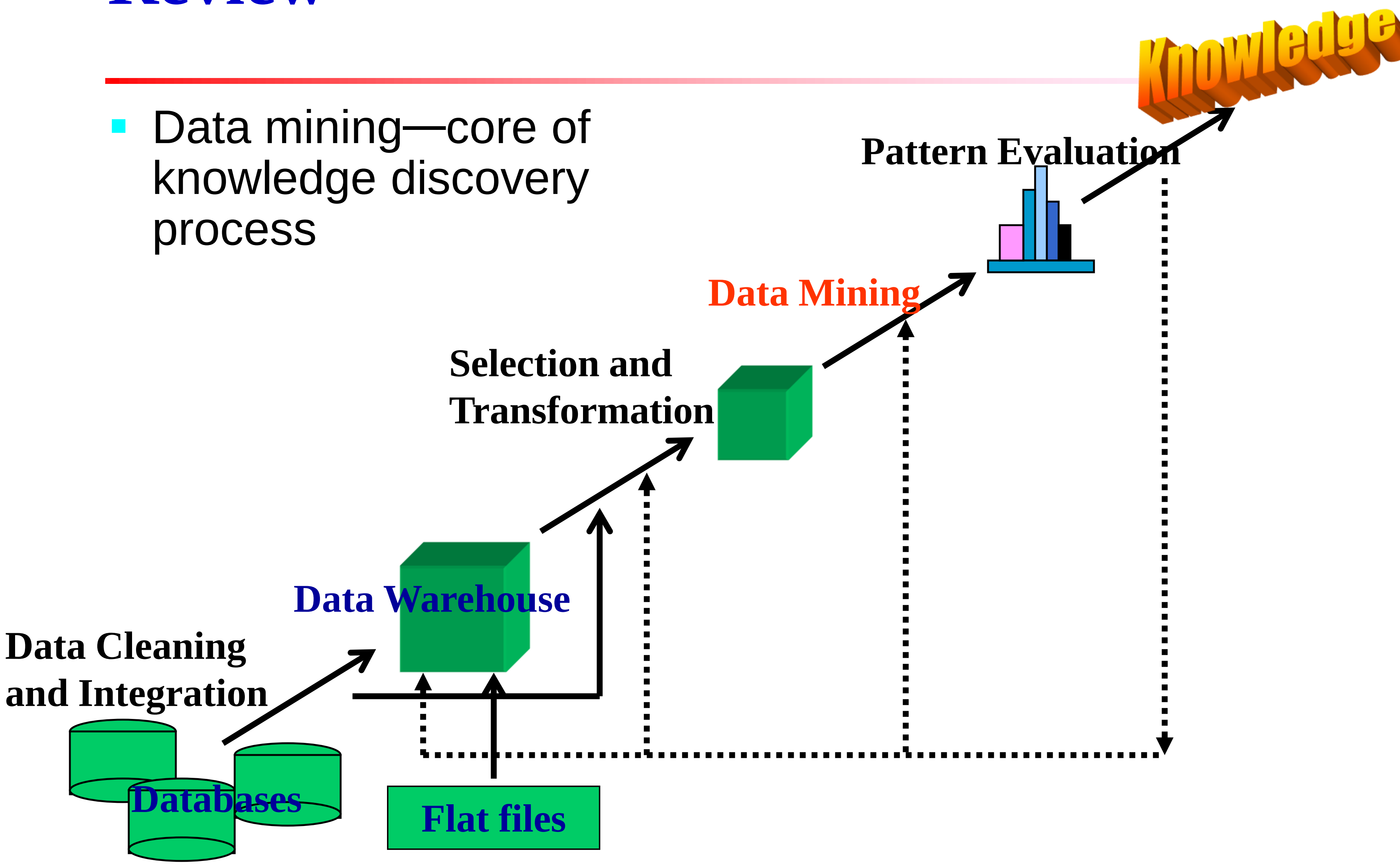
Data Mining

Ying Liu, Prof., Ph.D

University of Chinese Academy of Sciences

Review

- Data mining—core of knowledge discovery process



Review

- Learning the application domain
  - relevant prior knowledge and goals of application
- Creating a target data resource
- Data cleaning and preprocessing: (may take 60% of effort!)
- Data reduction and transformation
  - Find useful features, dimensionality/variable reduction, invariant representation
- Choosing the mining algorithm(s) to search for patterns of interest
- Pattern evaluation and knowledge presentation
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

Data Preprocessing Overview

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data transformation
- Data integration
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Why Data Preprocessing?

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregated 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”

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

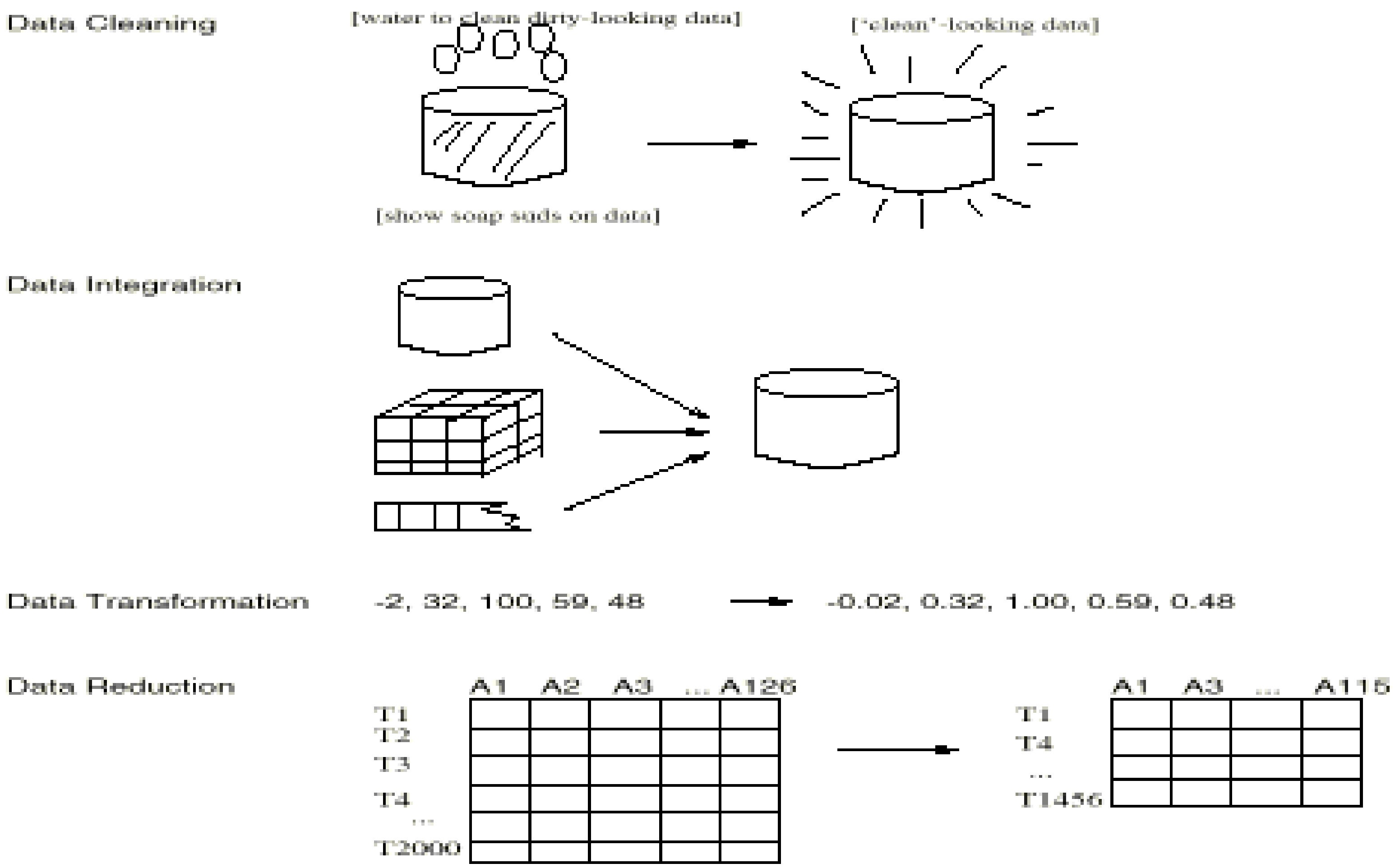
Why Is Data Preprocessing Important?

- 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 extraction, cleaning, and transformation comprises the majority of the work of building a 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, 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
  - Data reduction for numerical data

Forms of Data Preprocessing



Data Preprocessing Overview

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Data Descriptive Characteristics

- Central tendency
  - Mean, weighted mean, etc.
- 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

Measuring the Central Tendency

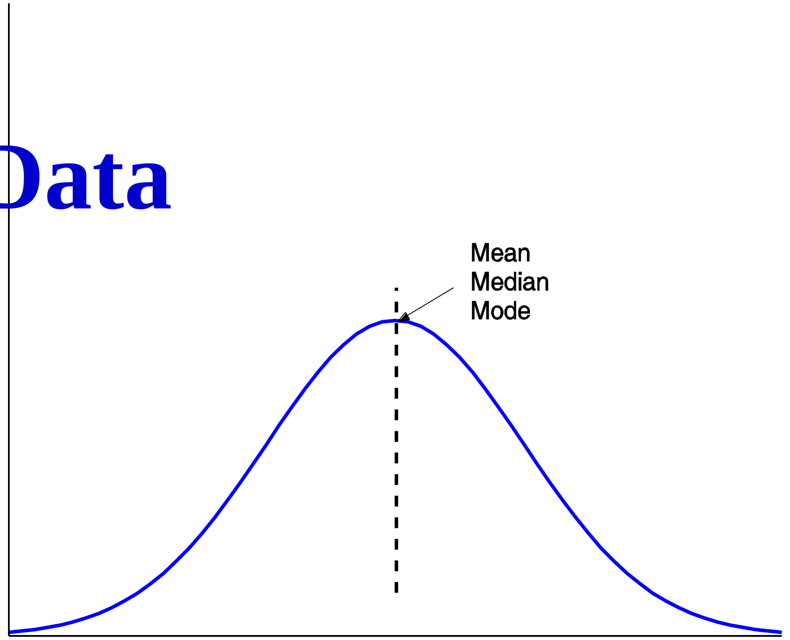
- Mean (algebraic measure):
  - Mean $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
  - Weighted arithmetic mean: $$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$
  - Trimmed mean: chopping extreme values

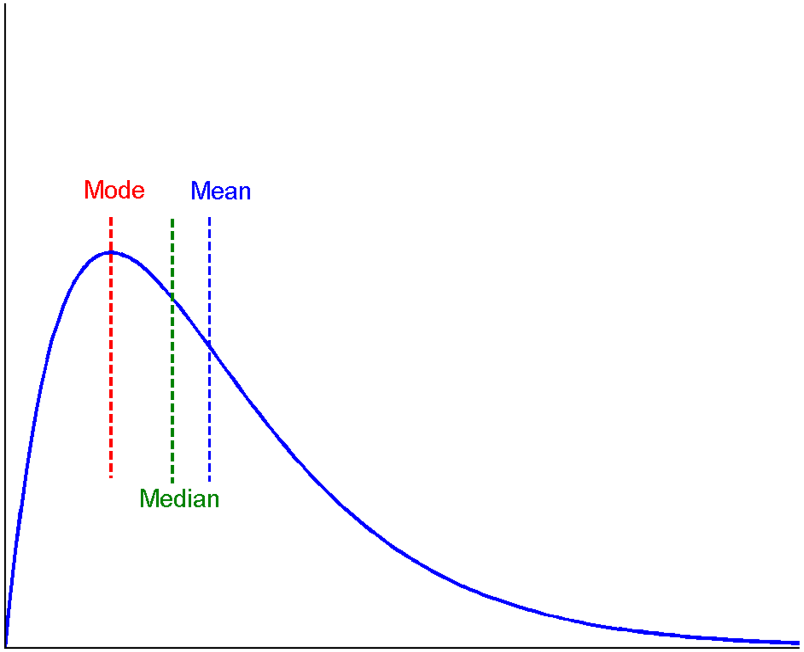
    去掉特别异常的值，目的：找出普遍性的特征
- Median: A holistic measure
  - Middle value if odd number of values, or average of the middle two values otherwise
- Mode 众数
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula: $$mean - mode = 3 \times (mean - median)$$

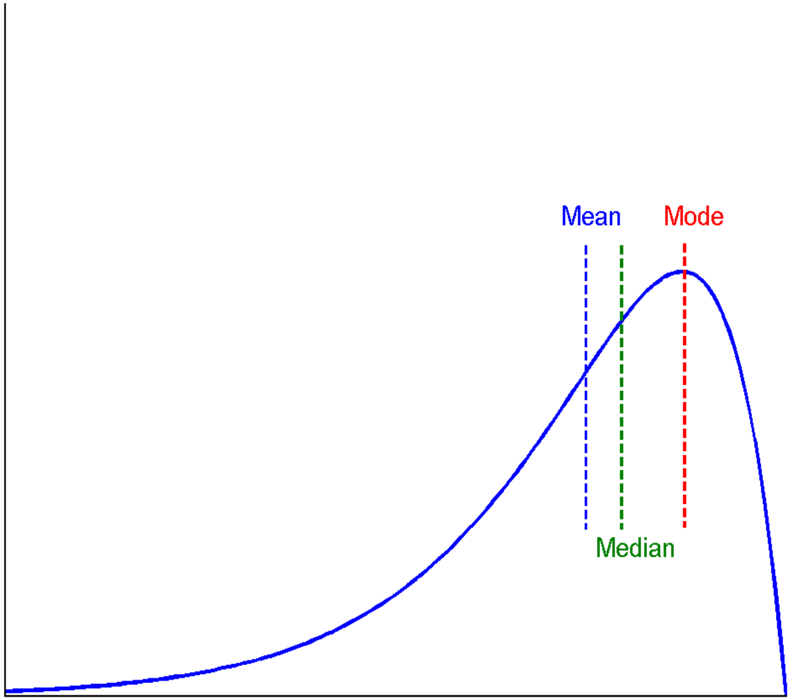
    经验公式

Symmetric vs. Skewed Data

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







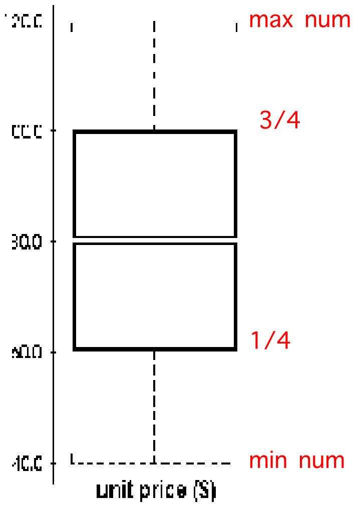
Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  4分位数
  - Quartiles: Q_1 (25th percentile), Q_3 (75th percentile)
  - Inter-quartile range: IQR = $Q_3 - Q_1$
  - Five number summary: min, Q_1, M, Q_3, max 表示数据分布
  - 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 x IQR
- Variance and standard deviation (*sample: s, population: σ*)
  - Variance: (algebraic, scalable computation)

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

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

Boxplot Analysis

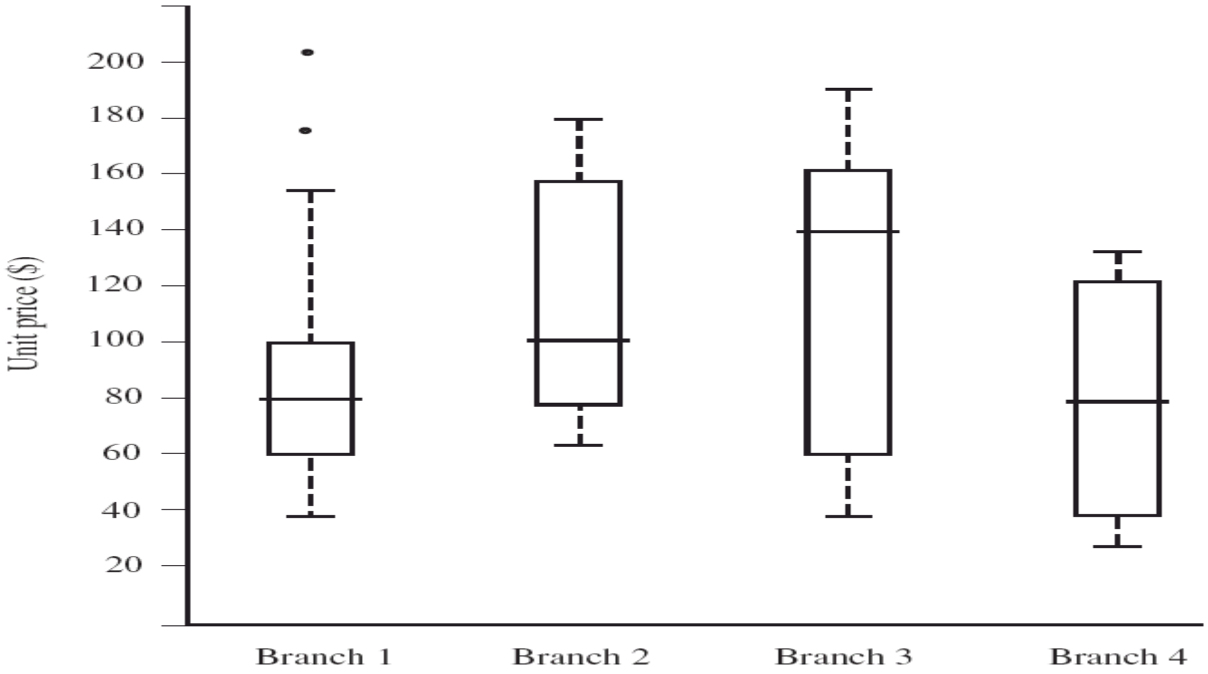


- Five-number summary of a distribution:

  Minimum, Q1, M, Q3, Maximum

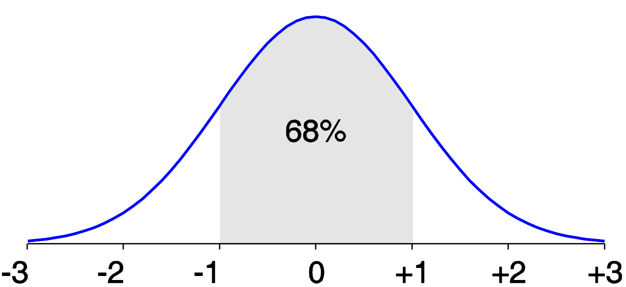
- Boxplot
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extend to Minimum and Maximum, terminate at 1.5xIQR

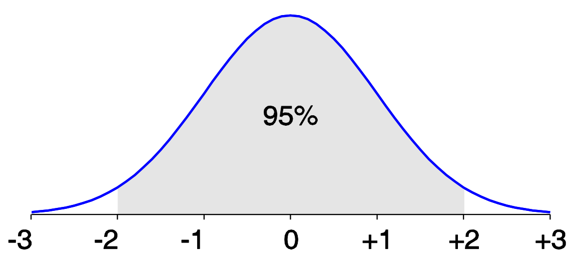
Boxplot Analysis

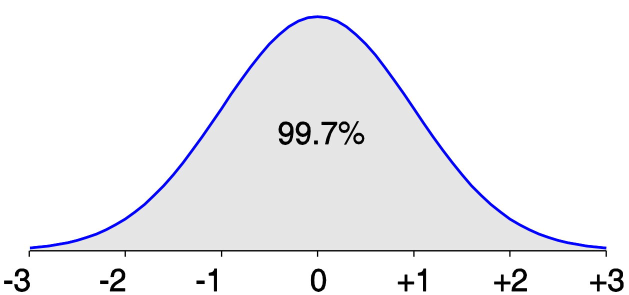


Properties of Normal Distribution Curve

- 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





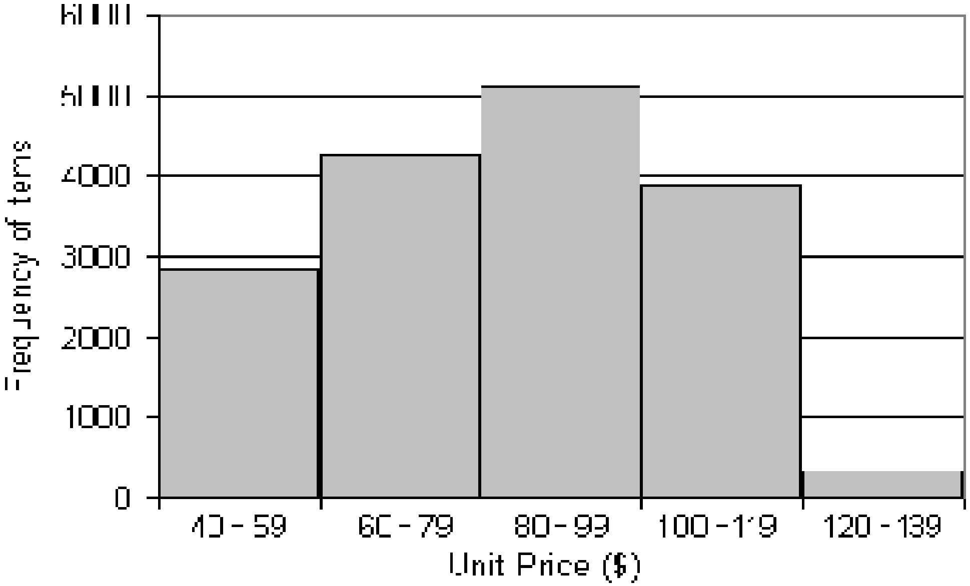


Histogram Analysis

- 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

Example

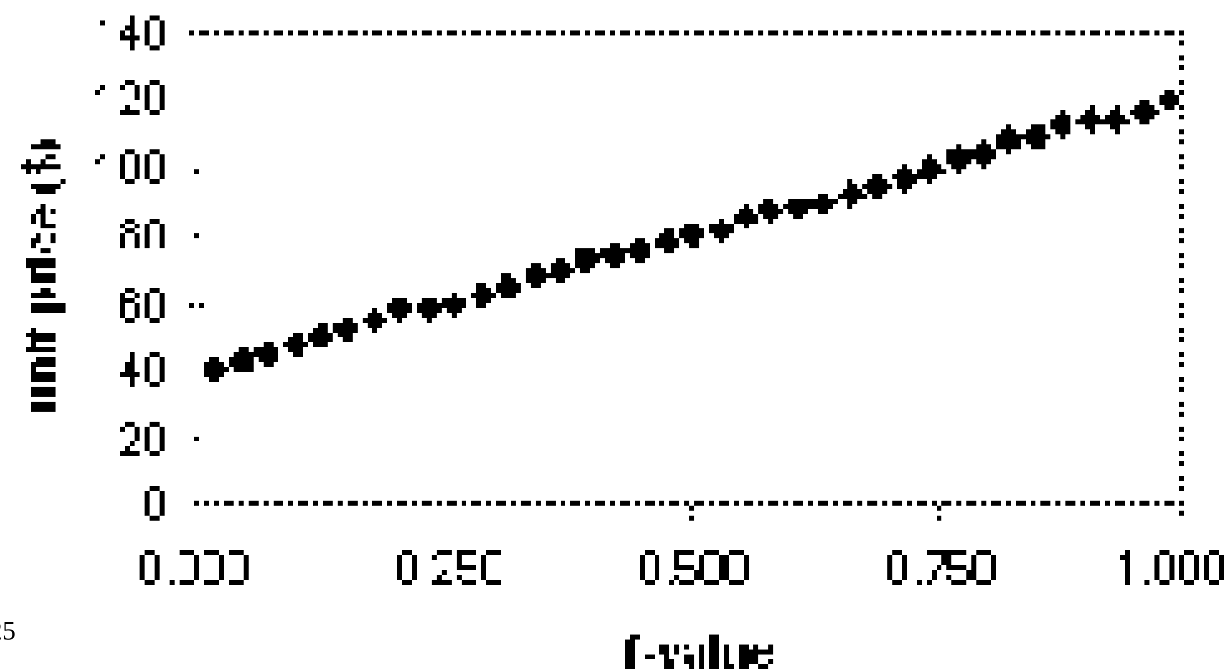
Unit price ($)	Count of items sold
40	275
43	300
47	250
..	..
74	360
75	515
78	540
..	..
115	320
117	270
120	350



Quantile Plot

百分位

- Display all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plot quantile information
  - For a data x_i data sorted in increasing order, f_i indicates that approximately 100 f_i% of the data are below or equal to the value x_i

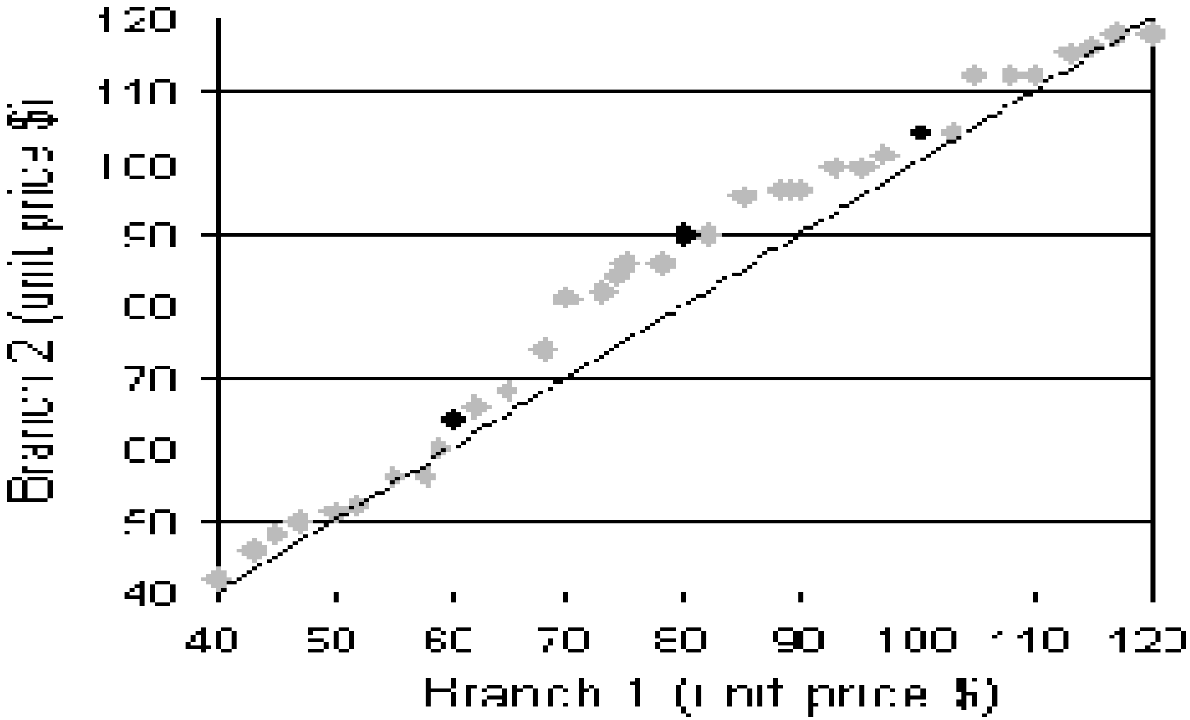


为了使点不画在边界上

$$f_i = \frac{i - 0.5}{n}$$

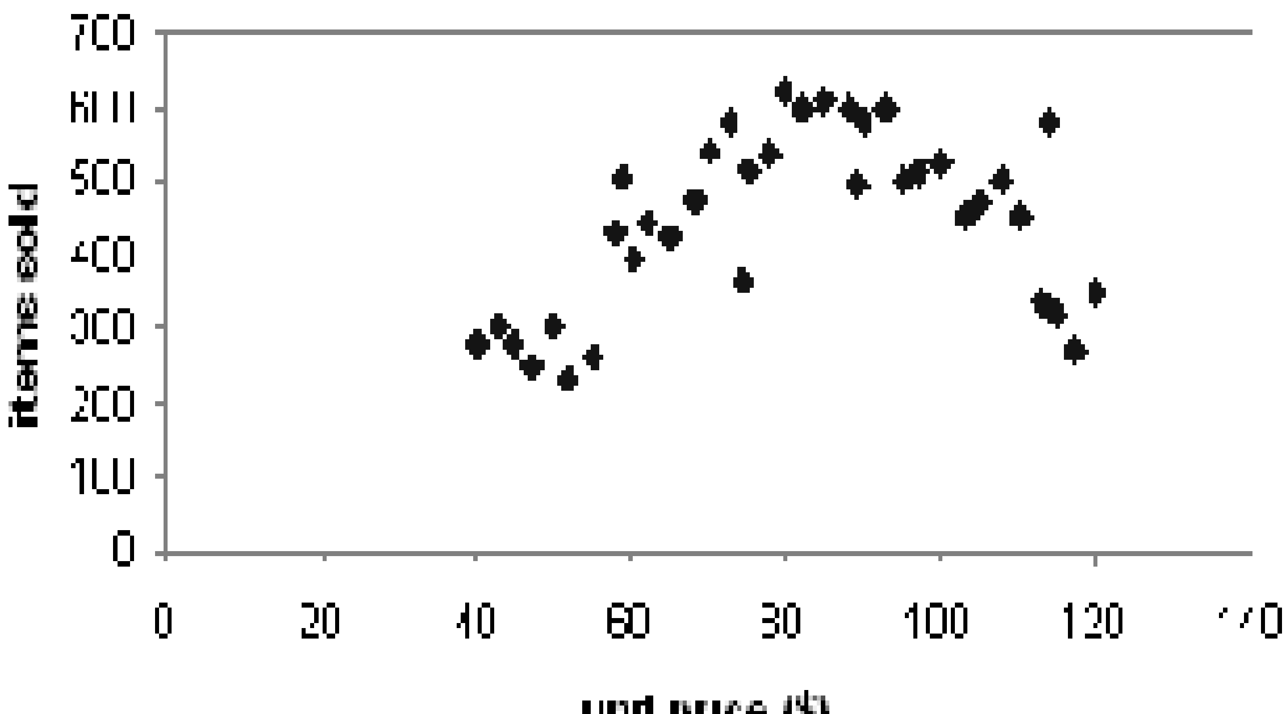
Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- Allows the user to view whether there is a shift in going from one distribution to another



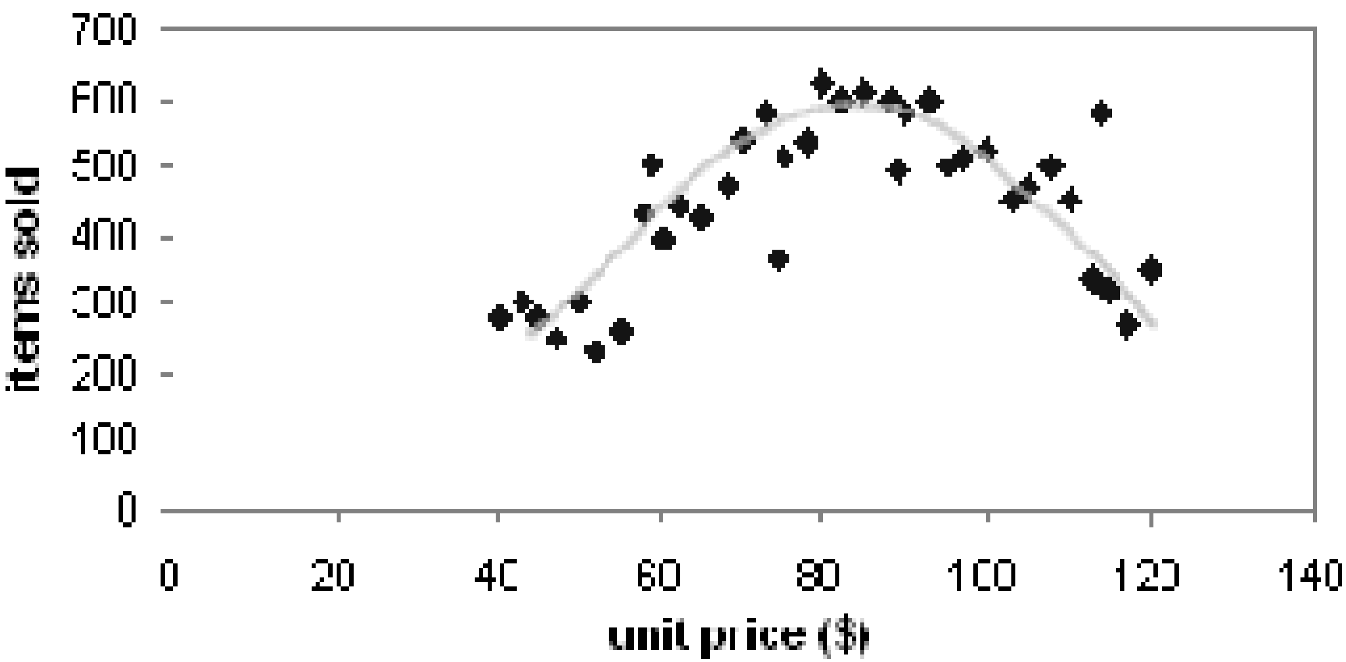
Scatter Plot

- Provides a first look at bivariate data to see clusters of points, outliers, etc.
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Loess Curve

- Adds a smooth curve to a scatter plot in order to provide better perception of the pattern
- Loess curve is fitted by setting two parameters: a smoothing parameter, and the degree of the polynomials that are fitted by the regression



Graphic Displays of Basic Statistical Descriptions

- Boxplot
- Histogram
- 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 univariant 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

Exercise

1. The values of data tuples are 13, 15, 16, 16, 19, 20, 20, 21.

(a) What is the mean of the data? What is the median?

(b) What is the mode of the data?

(c) What is Q1 and Q3?

(d) What is the IQR of the data?

(e) Give the five-number-summary of the data.

(f) Show a boxplot for the data.

Overview: Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data transformation
- Data integration
- Data reduction
- Discretization and concept hierarchy generation
- Summary

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

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: 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
  - inconsistent data

How to Handle Noisy Data?

- Binning
  - first sort data and partition into bins
  - Smooth noise by consulting its neighbors, local smooth
  - 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
  - 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

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

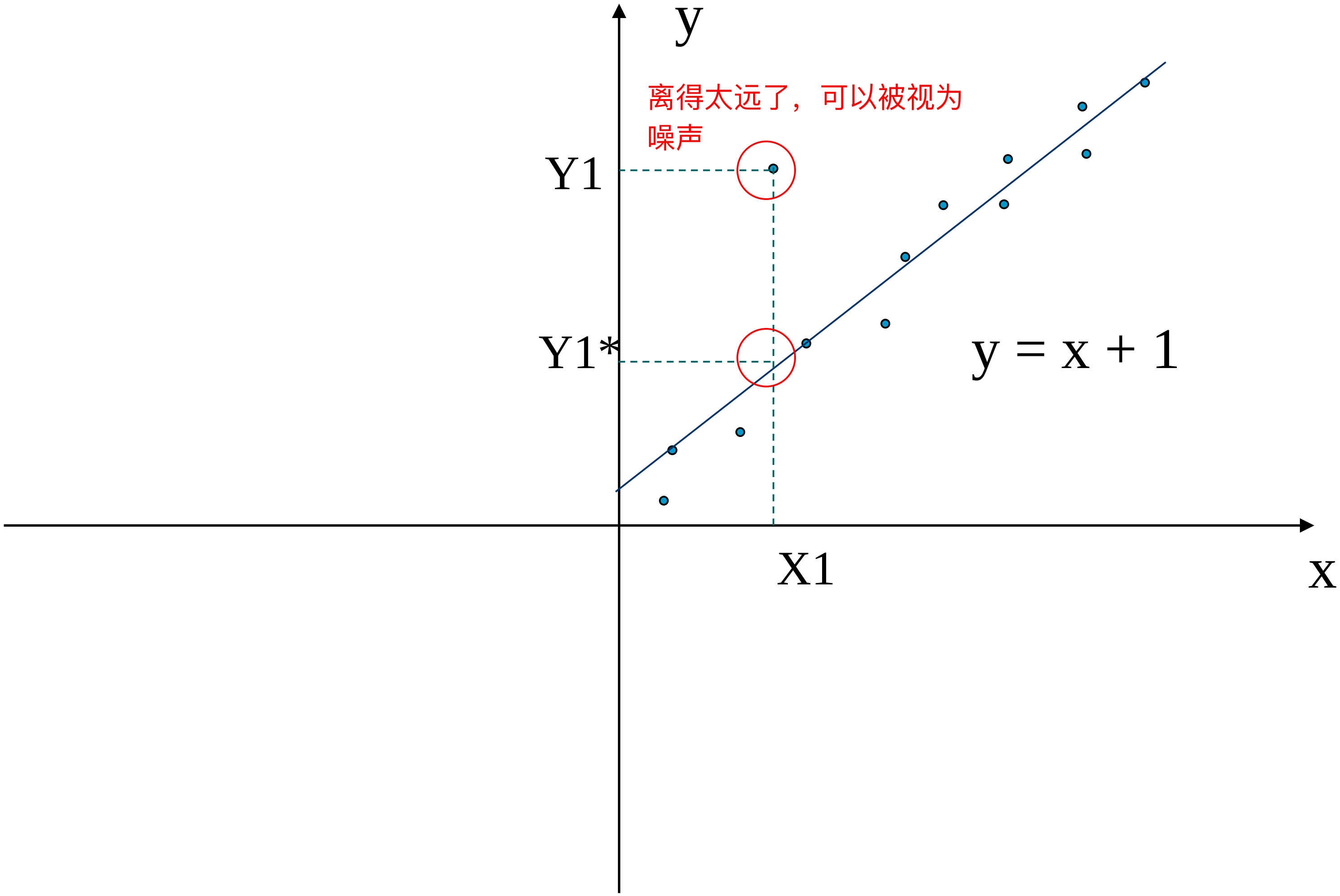
* Smoothing by bin means:

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

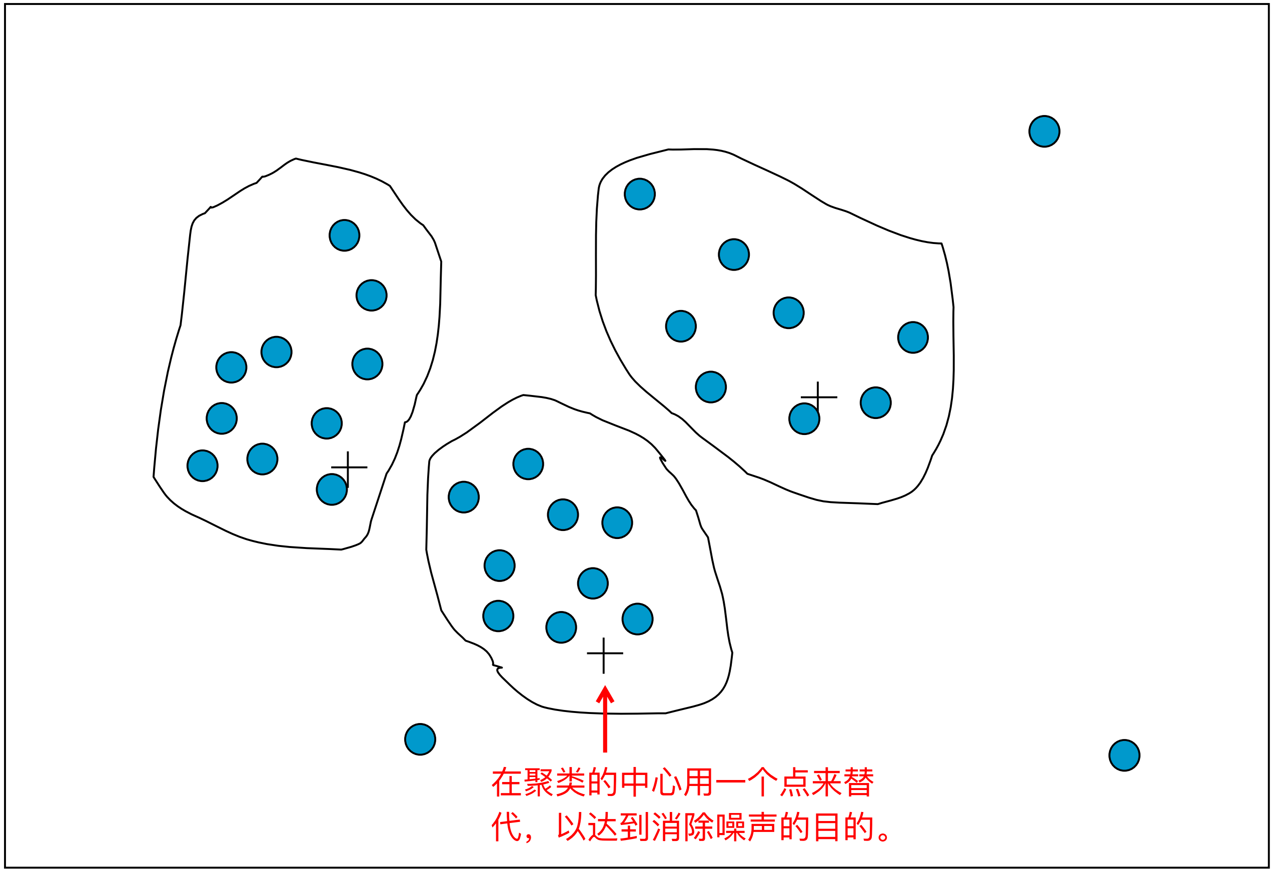
* Smoothing by bin boundaries:

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

Regression



Cluster Analysis



Exercise

1. Suppose a group of 12 sales price records has been sorted as follows: 5, 10, 11, 13, 15, 15, 15, 55, 60, 60, 65, 65.

(a) Smooth the data by bin means, using a bin depth of 4.

(b) Smooth the data by bin boundaries, using a bin depth of 4.

(c) Smooth the data by bin means, using 3 bins of equal-width partitioning.

Overview: Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data transformation
- Data integration
- Data reduction
- Discretization and concept hierarchy generation
  离散化
- Summary

Data Transformation

- Smoothing: remove noise from data
  - Binning, regression, clustering
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Aggregation: summarization, data cube construction

OLAP

- Attribute/feature construction
  - New attributes constructed from the given ones

Data Transformation: Normalization

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

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

一旦min max发生变化，模型将不再适用

Min、max会变化的，因此对于这个方法，无法保留住模型

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,600 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(|v'|) < 1

  - Ex. (-986, 917) => (-0.986, 0.917), j=3

Exercise

1. Use two methods to normalize the following group of data: 200, 300, 400, 600, 1000.

(a) min-max normalization by setting min=0 and max = 1.

(b) Z-score normalization (mean = 500, standard deviation = 316.2).

Overview: Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data transformation
- Data integration
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Data Integration

- Data integration:
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id ≡ 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 often occur
  - *Derivable data:* One attribute may be a “derived” attribute in another table, e.g., payment, payment rate
- Redundant attributes may be able to be detected by *correlation analysis*

Correlation Analysis (Numerical Data)

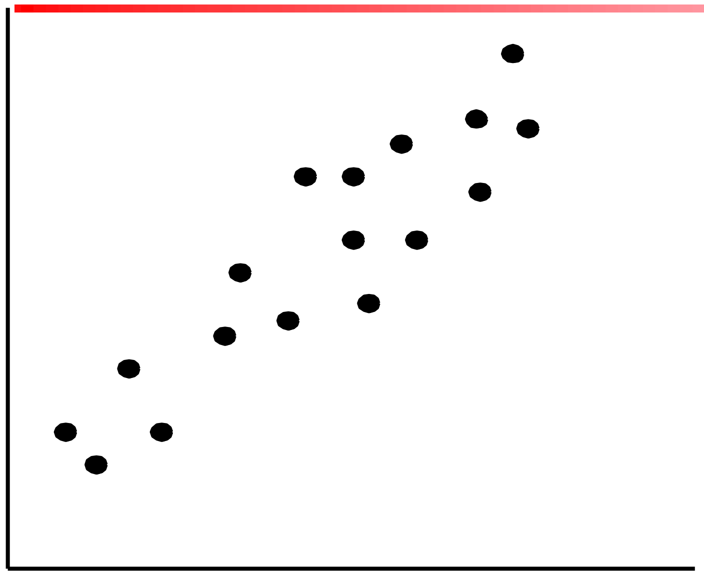
- Correlation coefficient (also called Pearson's product moment coefficient)

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

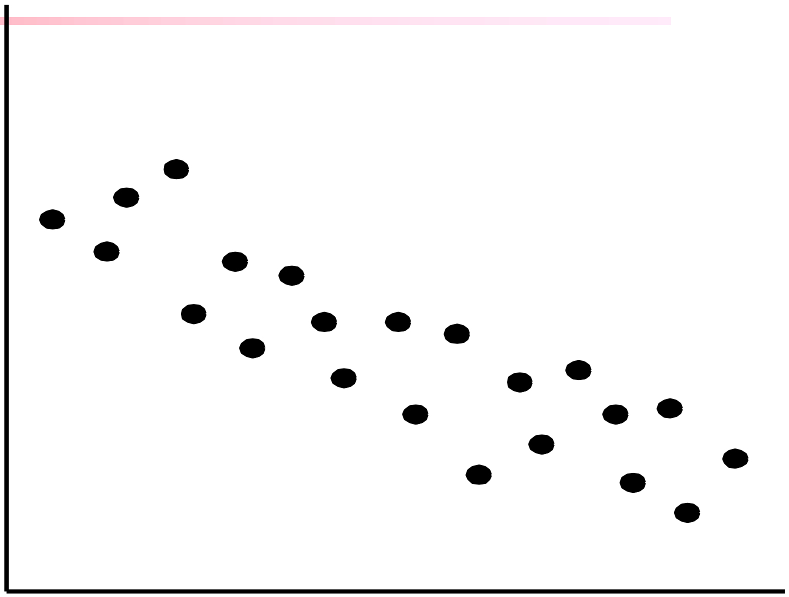
where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\Sigma(a_i b_i)$ is the dot-product of A and B.

- 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

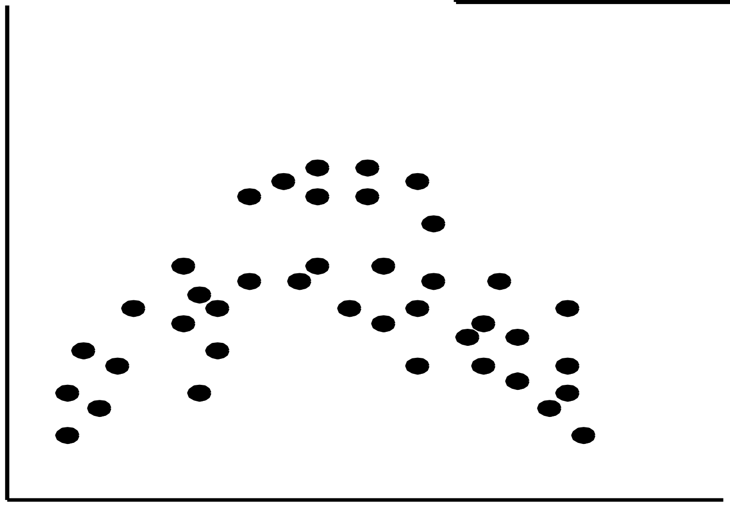
Positively and Negatively Correlated Data



(a)

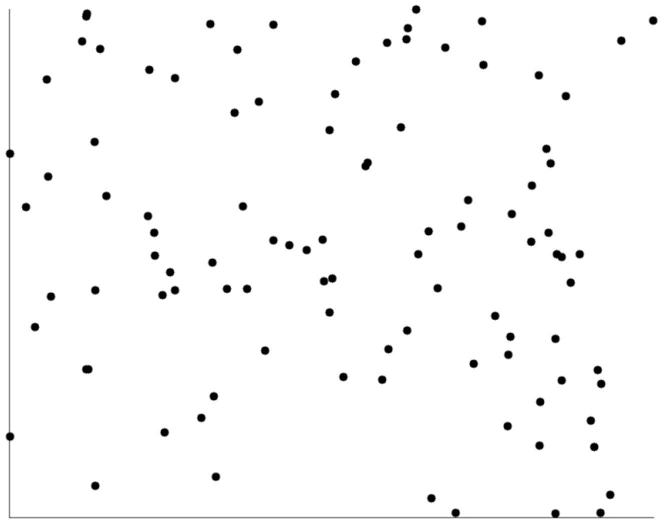


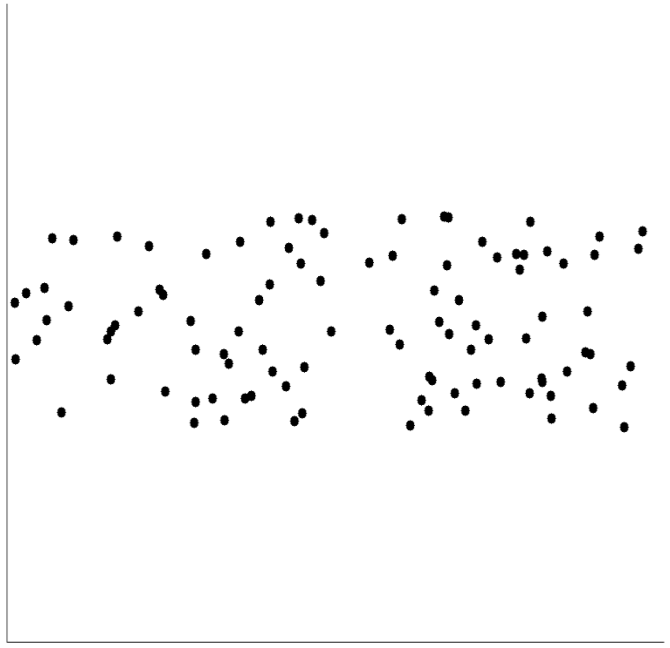
(b)

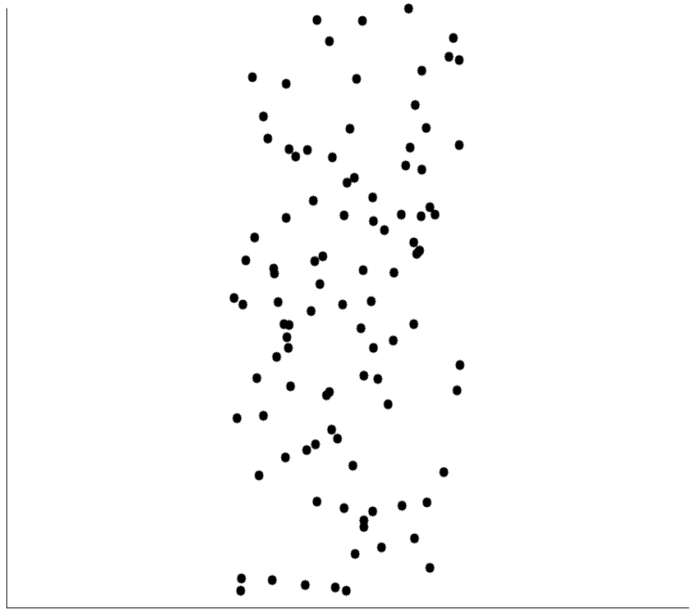


(c)

Not Correlated Data







Correlation Analysis (Categorical Data)

- X^2 (chi-square) test

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

- The larger the X^2 value, the more likely the variables are correlated
- The cells that contribute the most to the X^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

	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

- X^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, dependent

Handling Redundancy in Data Integration

- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Excerise

1. 邻接表
The following contingence table summarizes supermarket transaction data.

(a) Based on the given data, is the purchase of hot dogs independent of the purchase of hamburgers?

(b) If correlated, what kind of correlation relationship exists between the two items?

	hot dogs	not hot dogs	sum
hamburgers	4000	3500	7500
not hamburgers	2000	500	2500
sum	6000	4000	10000

Overview: Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data transformation
- Data integration
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Data Reduction

- 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 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
- Multiple levels of aggregation in data cubes
  - Further reduce the size of data to deal with
- Reference appropriate concept levels
  - Use the smallest representation which is enough to solve the task

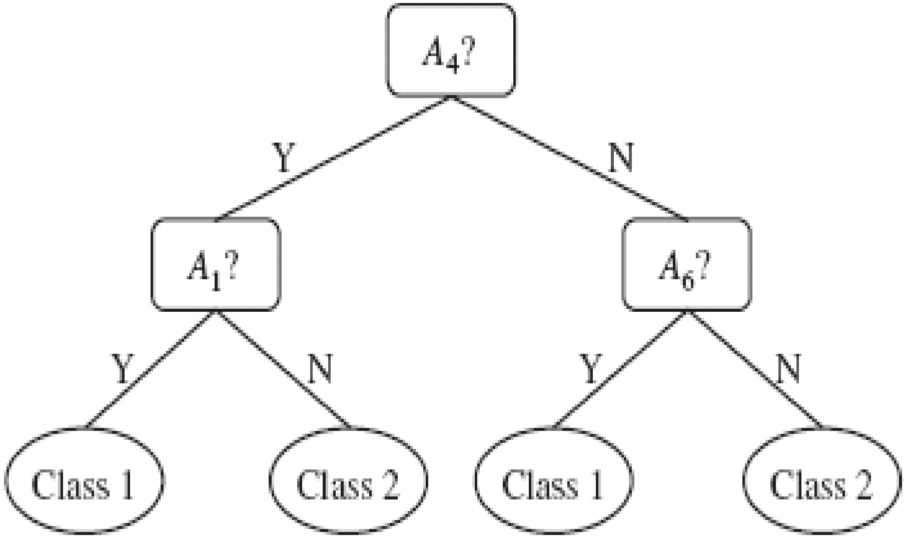
Attribute Subset Selection

- 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 features 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

Feature Selection Methods

- There are 2^d possible sub-features of d features
- Greedy methods: locally optimal
  - Choose by “statistical significance” tests
  - Best step-wise forward selection:
    - The best single-feature is picked first
    - Then next best feature condition to the first, ...
  - Step-wise backward elimination:
    - Repeatedly eliminate the worst feature
  - Best combined feature selection and elimination
- Decision tree induction

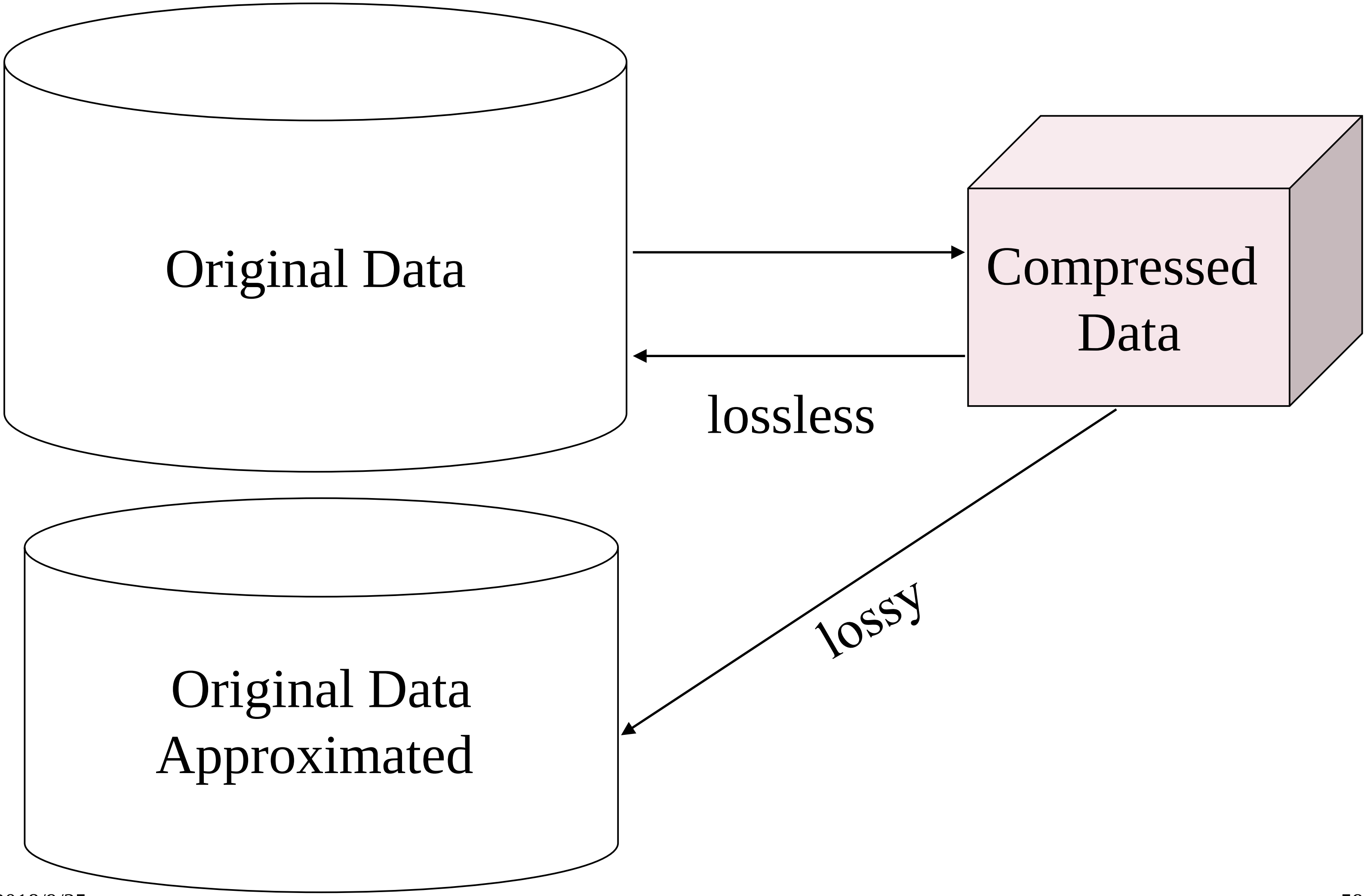
Example

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  => Reduced attribute set: $\{A_1, A_4, A_6\}$

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

Data Compression



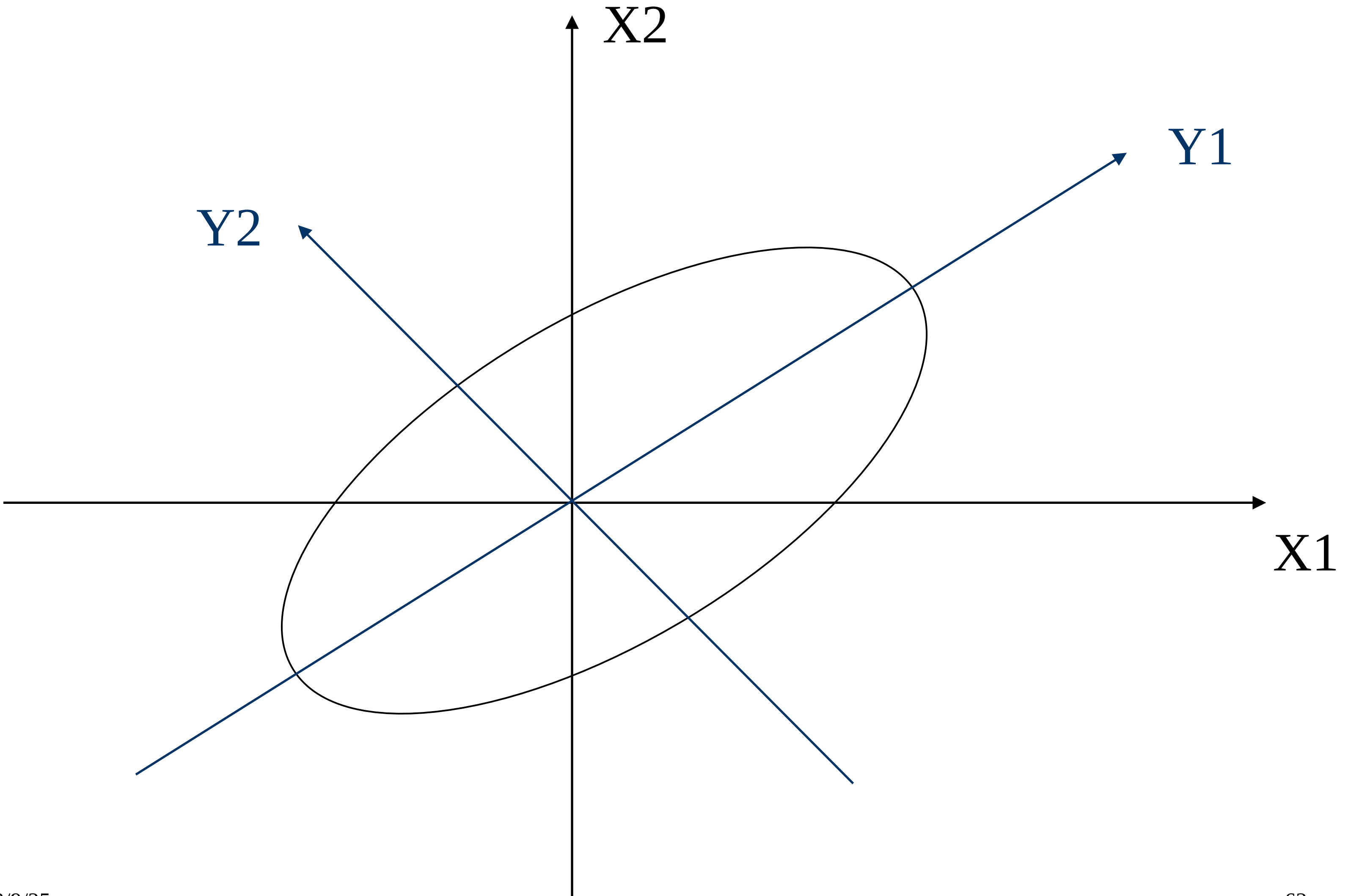
Wavelet Transformation

- Discrete wavelet transform (DWT): linear signal processing, multi-resolutional analysis
- Compressed approximation: store only a small fraction of the strongest wavelet coefficients
- Similar to discrete Fourier transform (DFT), but better lossy compression, localized in space
- Method:
  - Length, L, must be an integer power of 2 (padding with 0's, when necessary)
  - Each transform has 2 functions: smoothing, difference
  - Applies to pairs of data, resulting in two sets of data of length L/2
  - Applies two functions recursively, until reaches the desired length

Dimensionality Reduction: Principal Component Analysis (PCA)

- Given N data vectors from n-dimensions, find $k \le n$ vectors (*principal components*) that can be best used to represent data
- Steps
  - Normalize input data: Each attribute falls within the same range
  - Compute k vectors, i.e., *principal components*
  - Each input data (vector) is a linear combination of the k principal component vectors
  - The principal components are sorted in order of decreasing "significance" or strength
  - Since the components are sorted, the size of the data can be reduced by eliminating the weak components, i.e., those with low variance. (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data)
- Work for numeric data only

Principal Component Analysis



Numerosity Reduction

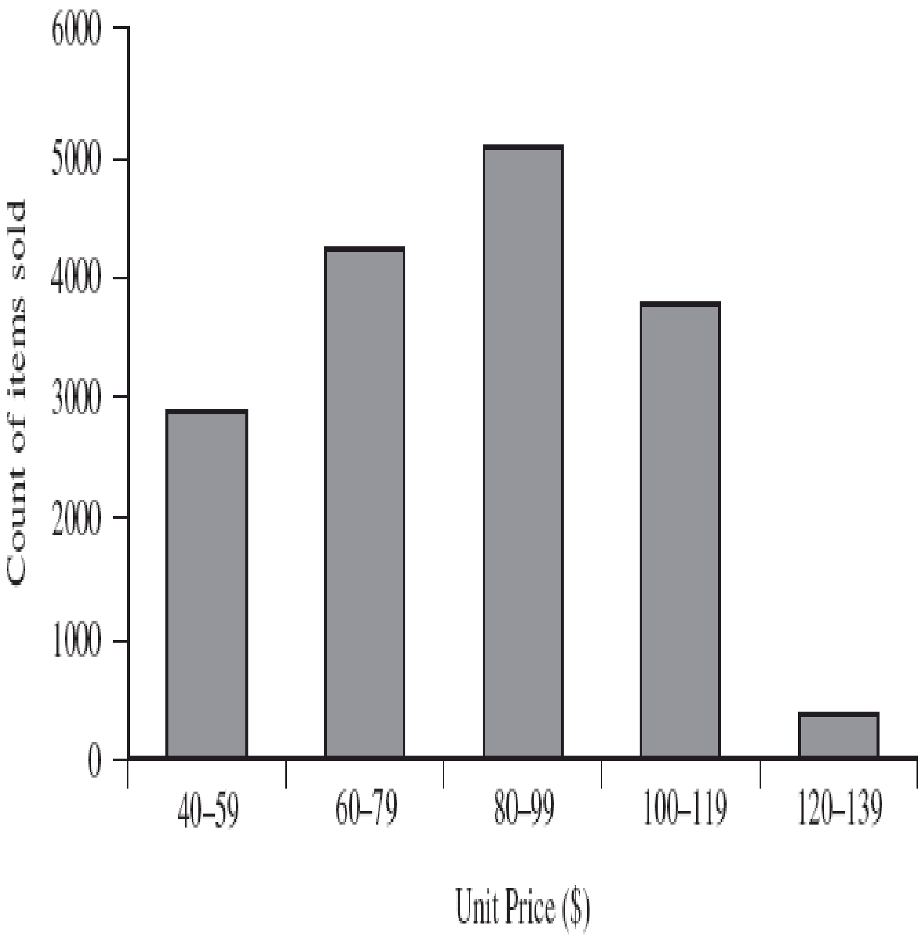
- Reduce data volume by choosing alternative, smaller forms of data representation
- Parametric methods
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
- Non-parametric methods
  - Do not assume models
  - Major families: histograms, clustering, sampling

Data Reduction Method (1): Regression Models

- Linear regression: Data are modeled to fit a straight line
  - Often uses the least-square method to fit the line
  - $Y = \alpha + \beta X$， two regression coefficients
- Multiple regression: allows a response variable Y to be modeled as a linear function of multidimensional feature vector
  - $Y = \alpha + \beta_1 X_1 + \beta_2 X_2$

Data Reduction Method (2): Histograms

- Divide data into buckets and store frequency for each bucket
- Partitioning rules:
  - Equal-width: equal bucket range
  - Equal-frequency (or equal-depth)
  - V-optimal: with the least *histogram variance* (weighted sum of the original values that each bucket represents)



minimize sum of (bar value = counts -sum of the Origen value in the bar)

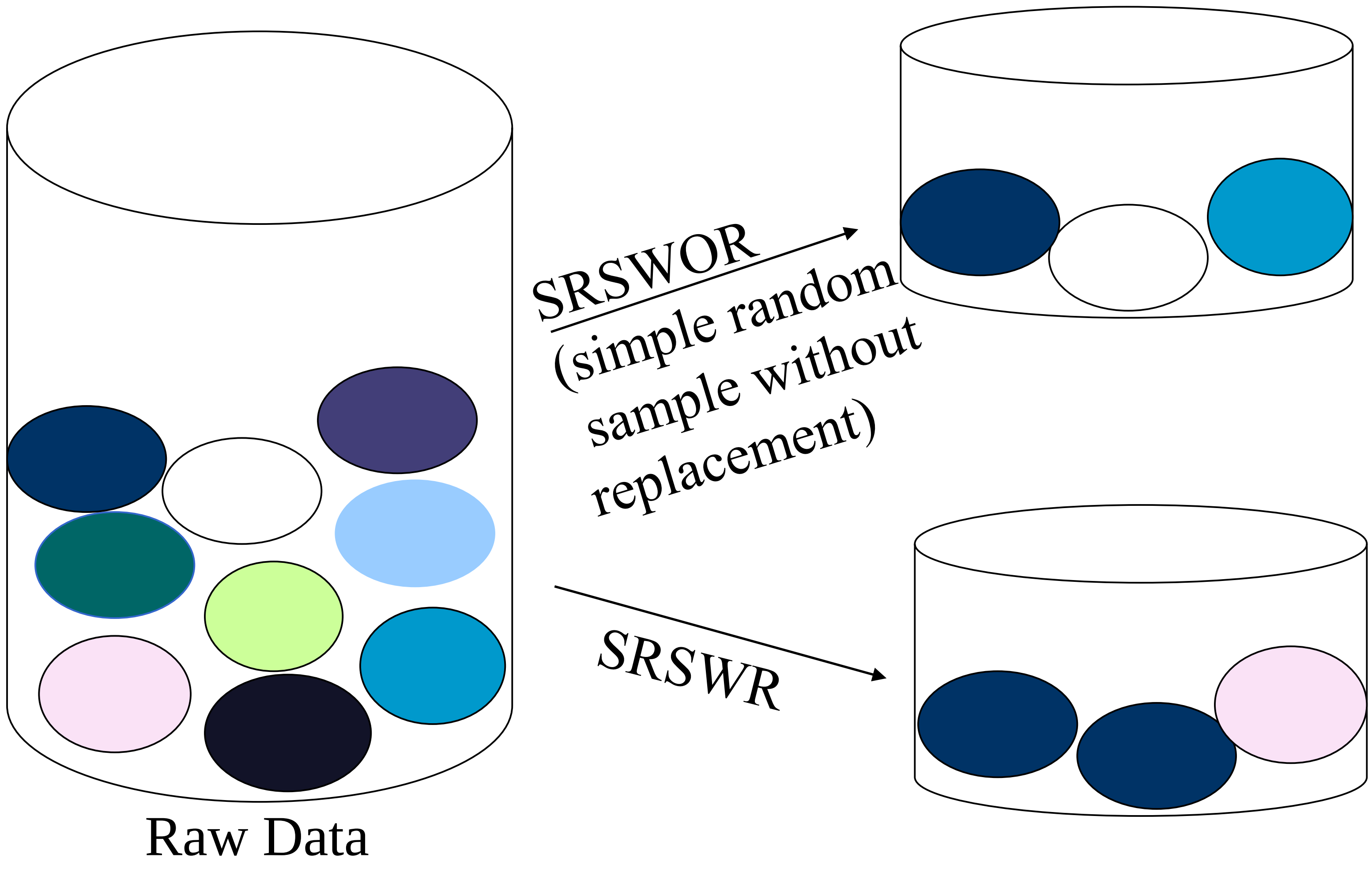
Data Reduction Method (3): Clustering

- Partition data set into clusters based on similarity
- Store cluster representation (e.g., centroid and diameter) only
- Can be very effective for data that can be clustered but not for “smeared” data
- There are many choices of clustering definitions and clustering algorithms

Data Reduction Method (4): Sampling

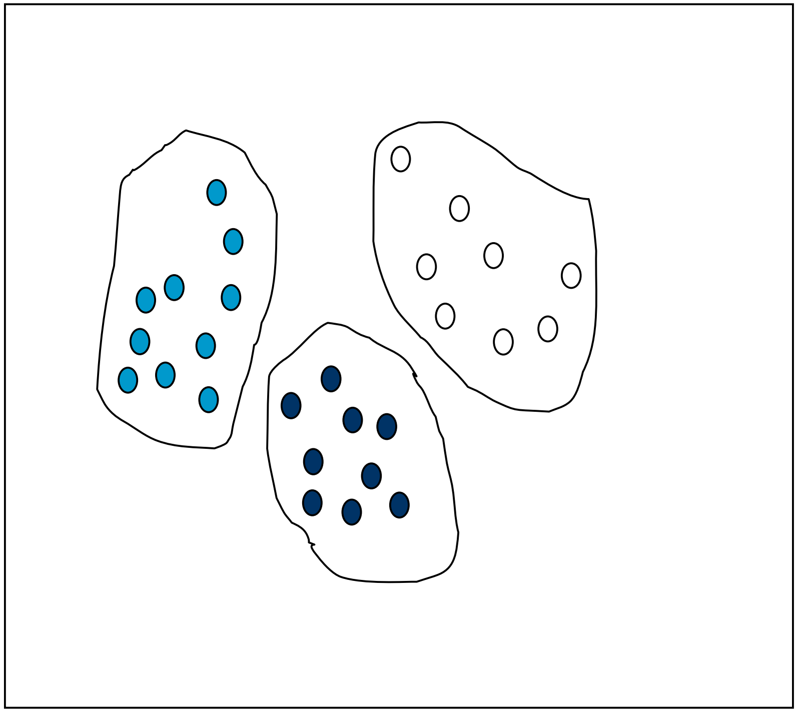
- Sampling: obtaining a small sample *s* to represent the whole data set *N*
- Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew Negative in outlier data
- Adaptive sampling methods
  - Stratified sampling:
    - Approximate the percentage of each class (or subpopulation of interest) in the overall database
    - Used in conjunction with skewed data
- Fast, scan database once

Sampling:With or Without Replacement

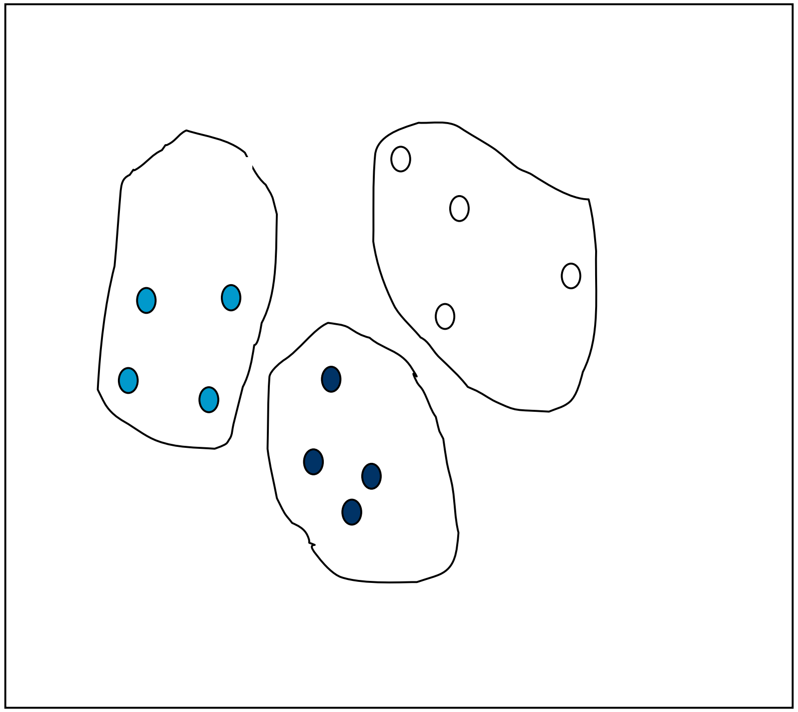


Sampling: Cluster or Stratified Sampling

Raw Data



Cluster/Stratified Sample



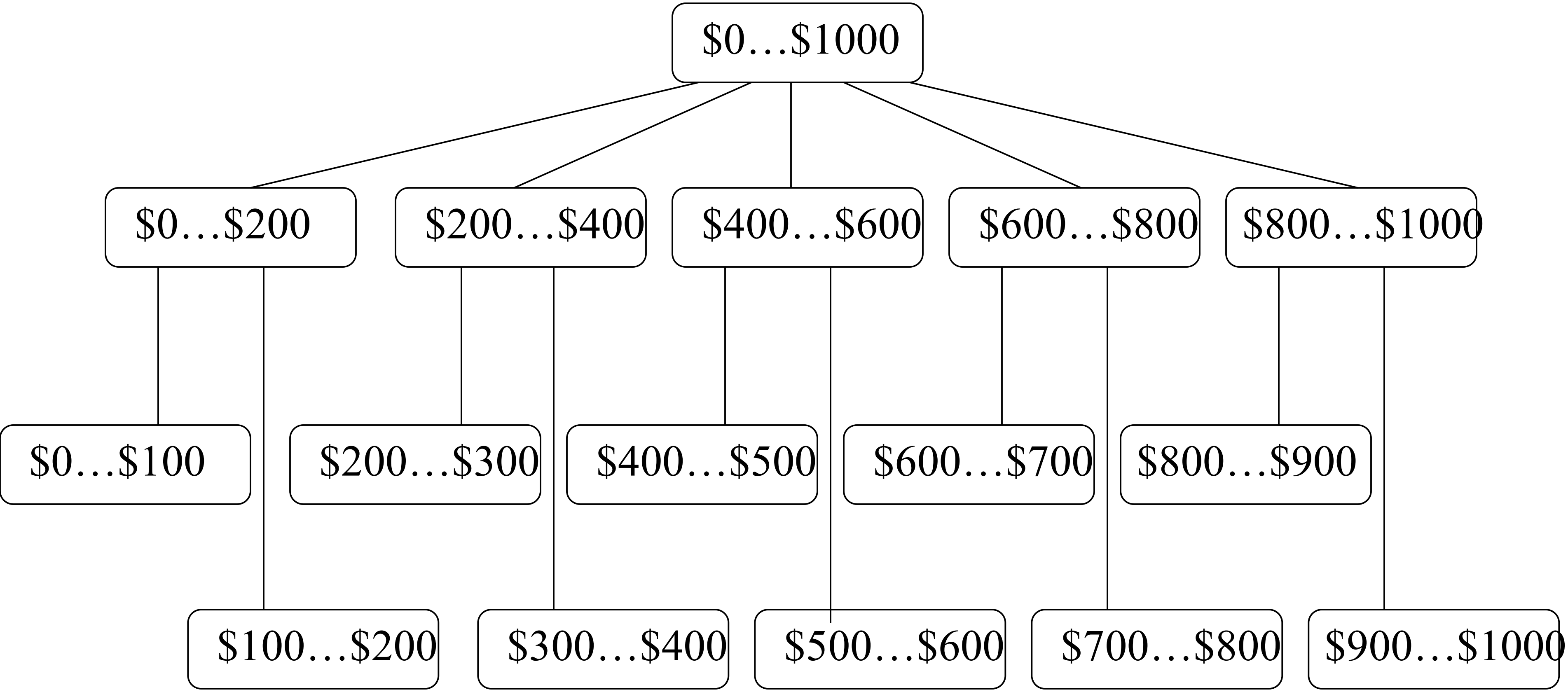
Overview: Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data transformation
- Data integration
- Data reduction
- Discretization and concept hierarchy generation
- Summary

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
  - 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)

Example of Concept Hierarchy



Discretization and Concept Hierarchy Generation for Numeric Data

- Typical methods: All the methods can be applied recursively
  - Binning
    - Top-down split, replace the value by bin mean or median
  - Histogram analysis
    - Top-down split
  - Clustering analysis
    - Top-down split
  - Entropy-based discretization: supervised, top-down split
  - Segmentation by natural partitioning: top-down split

Entropy-Based Discretization

- Entropy is calculated based on class distribution of the samples in the set. Given *m* classes, the entropy of *S* is

$$Entropy(S) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$ where p_i is the probability of class *i* in *S*

- Given a set of samples *S*, if *S* is partitioned into two intervals S_1 and S_2 using boundary *T*, the entropy after partitioning is

?

$$\boxed{Entropy(S,T) = \frac{|S_1|}{|S|} Entropy(S_1) + \frac{|S_2|}{|S|} Entropy(S_2)}$$

- The boundary that maximizes the information gain over all possible boundaries is selected as a binary discretization

$$Gain(S,T) = Entropy(S) - Entropy(S,T)$$

- The process is recursively applied to partitions obtained until some stopping criterion is met
- Such a boundary may reduce data size and improve classification accuracy

An Example of Entropy-based Partitioning

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

An Example of Entropy-based Partitioning

$$Gain(age) = 0.246$$

$$Gain(income) = 0.029$$

$$Gain(student) = 0.151$$

$$Gain(credit\_rating) = 0.048$$

age?

<=30

income	*student*	*credit_rating*	*class*
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

30..40

income	*student*	*credit_rating*	*class*
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

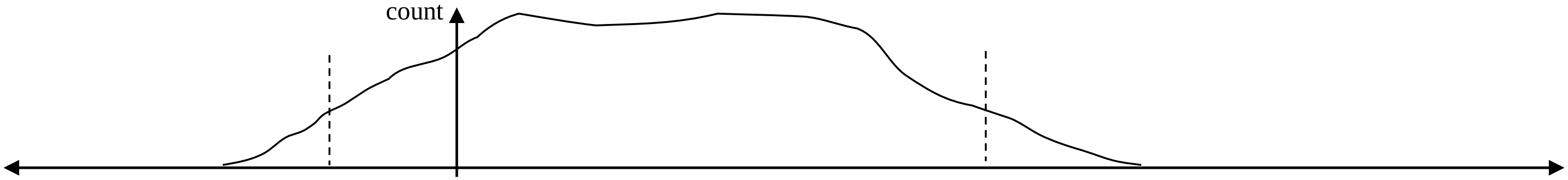
>40

income	*student*	*credit_rating*	*class*
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

Segmentation by Natural Partitioning

- A simply 3-4-5 rule can be used to segment numeric data into relatively uniform, “natural” intervals
  - If an interval covers 3, 6, 7 or 9 distinct values at the “most significant digit”, partition the range into 3 equal-width intervals
  - If it covers 2, 4, or 8 distinct values at the “most significant digit”, partition the range into 4 intervals
  - If it covers 1, 5, or 10 distinct values at the “most significant digit”, partition the range into 5 intervals
  - The top-level segmentation represents the majority

Example of 3-4-5 Rule



Step 1: -$351 -$159 profit $1,838 $4,700

Min Low (i.e, 5%-tile) High(i.e, 95%-0 tile) Max

Cut the extreme value, remove the outlier data

Step 2: msd=1,000 Low=-$1,000 High=$2,000

Step 3:

(-$1,000 - $2,000)

(-$1,000 - 0) (0 -$ 1,000) ($1,000 - $2,000)

Step 4:

(-$400 -$5,000)

(-$400 - 0)
- (-$400 - -$300)
- (-$300 - -$200)
- (-$200 - -$100)
- (-$100 - 0)

(0 - $1,000)
- (0 - $200)
- ($200 - $400)
- ($400 - $600)
- ($600 - $800)
- ($800 - $1,000)

($1,000 - $2, 000)
- ($1,000 - $1,200)
- ($1,200 - $1,400)
- ($1,400 - $1,600)
- ($1,600 - $1,800)
- ($1,800 - $2,000)

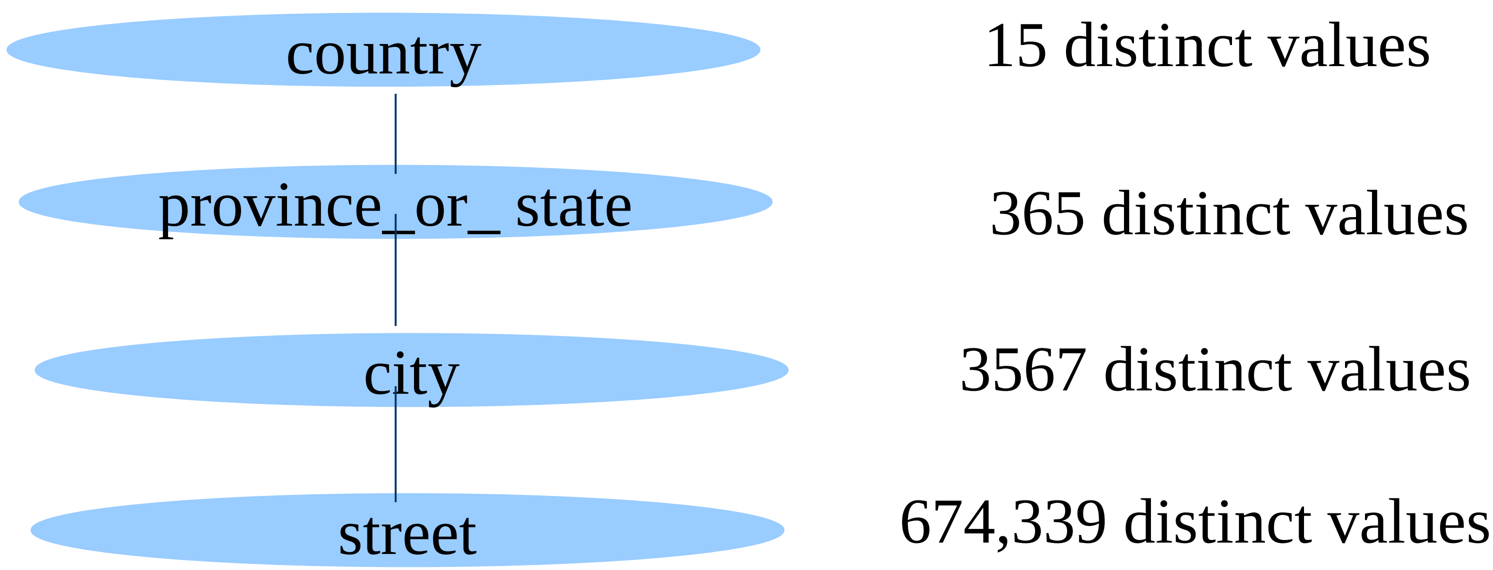
($2,000 - $5, 000)
- ($2,000 - $3,000)
- ($3,000 - $4,000)
- ($4,000 - $5,000)

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} < Ilinois
- 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



The more distinct values, the level lower, the lest distinct values, the higher levels

Overview: Data Preprocessing

- Why preprocess the data?
- Descriptive data summarization
- Data cleaning
- Data transformation
- Data integration
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Summary

- Data preparation or preprocessing is a big issue for both data warehousing and data mining
- Descriptive data summarization is needed 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