

# Data Mining: Concepts and Techniques

— Slides for Textbook —  
— Chapter 5 —

Jiawei Han and Micheline Kamber

Intelligent Database Systems Research Lab Simon Fraser  
University,

Ari Visa, , Institute of Signal Processing

Tampere University of Technology

# Chapter 5: Concept Description: Characterization and Comparison

- What is concept description?
- Data generalization and summarization-based characterization
- Analytical characterization: Analysis of attribute relevance
- Mining class comparisons: Discriminating between different classes
- Mining descriptive statistical measures in large databases
- Discussion
- Summary

# What is Concept Description?

- Descriptive vs. predictive data mining
  - **Descriptive mining**: describes concepts or task-relevant data sets in concise, summarily, informative, discriminative forms
  - **Predictive mining**: Based on data and analysis, constructs models for the database, and predicts the trend and properties of unknown data
- Concept description:
  - **Characterization**: provides a concise and succinct summarization of the given collection of data
  - **Comparison**: provides descriptions comparing two or more collections of data

# Concept Description vs. OLAP

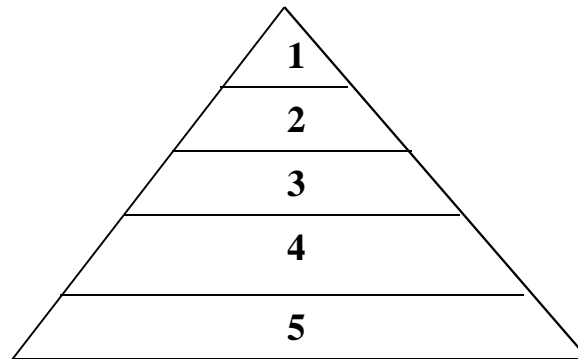
- Concept description:
  - can handle complex data types of the attributes and their aggregations
  - a more automated process
- OLAP:
  - restricted to a small number of dimension and measure types
  - user-controlled process

# Chapter 5: Concept Description: Characterization and Comparison

- What is concept description?
- [Data generalization and summarization-based characterization](#)
- Analytical characterization: Analysis of attribute relevance
- Mining class comparisons: Discriminating between different classes
- Mining descriptive statistical measures in large databases
- Discussion
- Summary

# Data Generalization and Summarization-based Characterization

- Data generalization
  - A process which abstracts a large set of task-relevant data in a database from a low conceptual levels to higher ones.



Conceptual levels

- Approaches:
  - Data cube approach(OLAP approach)
  - Attribute-oriented induction approach

# Characterization: Data Cube Approach (without using AO-Induction)

- Perform computations and store results in data cubes
- Strength
  - An efficient implementation of data generalization
  - Computation of various kinds of measures
    - e.g., `count()`, `sum()`, `average()`, `max()`
  - Generalization and specialization can be performed on a data cube by *roll-up* and *drill-down*
- Limitations
  - handle only dimensions of *simple nonnumeric data* and measures of *simple aggregated numeric values*.
  - Lack of intelligent analysis, can't tell which dimensions should be used and what levels should the generalization reach

# Attribute-Oriented Induction

- Proposed in 1989 (KDD '89 workshop)
- Not confined to categorical data nor particular measures.
- How it is done?
  - Collect the task-relevant data( *initial relation*) using a relational database query
  - Perform generalization by attribute removal or attribute generalization.
  - Apply aggregation by merging identical, generalized tuples and accumulating their respective counts.
  - Interactive presentation with users.



# Basic Principles of Attribute-Oriented Induction

- [Data focusing](#): task-relevant data, including dimensions, and the result is the *initial relation*.
- [Attribute-removal](#): remove attribute  $A$  if there is a large set of distinct values for  $A$  but (1) there is no generalization operator on  $A$ , or (2)  $A$ 's higher level concepts are expressed in terms of other attributes.
- [Attribute-generalization](#): If there is a large set of distinct values for  $A$ , and there exists a set of generalization operators on  $A$ , then select an operator and generalize  $A$ .
- [Attribute-threshold control](#): typical 2-8, specified/default.
- [Generalized relation threshold control](#): control the final relation/rule size. [see example](#)

# Basic Algorithm for Attribute-Oriented Induction

- [InitialRel](#): Query processing of task-relevant data, deriving the *initial relation*.
- [PreGen](#): Based on the analysis of the number of distinct values in each attribute, determine generalization plan for each attribute: removal? or how high to generalize?
- [PrimeGen](#): Based on the PreGen plan, perform generalization to the right level to derive a “prime generalized relation”, accumulating the counts.
- [Presentation](#): User interaction: (1) adjust levels by drilling, (2) pivoting, (3) mapping into rules, cross tabs, visualization presentations.

[See Implementation](#)   [See example](#)   [See complexity](#)

# Example

- **DMQL:** Describe general characteristics of graduate students in the Big-University database

**use** Big\_University\_DB

**mine characteristics as** "Science\_Students"

**in relevance to** name, gender, major, birth\_place,  
birth\_date, residence, phone#, gpa

**from** student

**where** status in "graduate"

- **Corresponding SQL statement:**

**Select** name, gender, major, birth\_place, birth\_date,  
residence, phone#, gpa

**from** student

**where** status in {"Msc", "MBA", "PhD" }

# Class Characterization: An Example

**Initial Relation**

Name	Gender	Major	Birth-Place	Birth_date	Residence	Phone #	GPA
Jim Woodman	M	CS	Vancouver,BC, Canada	8-12-76	3511 Main St., Richmond	687-4598	3.67
Scott Lachance	M	CS	Montreal, Que, Canada	28-7-75	345 1st Ave., Richmond	253-9106	3.70
Laura Lee	F	Physics	Seattle, WA, USA	25-8-70	125 Austin Ave., Burnaby	420-5232	3.83
...	...	...	...	...	...	...	...
<b>Removed</b>	<b>Retained</b>	<b>Sci,Eng, Bus</b>	<b>Country</b>	<b>Age range</b>	<b>City</b>	<b>Removed</b>	<b>Excl, VG,..</b>

**Prime Generalized Relation**

Gender	Major	Birth_region	Age_range	Residence	GPA	Count
M	Science	Canada	20-25	Richmond	Very-good	16
F	Science	Foreign	25-30	Burnaby	Excellent	22
...	...	...	...	...	...	...

Gender \ Birth_Region	Canada	Foreign	Total
	M	16	14
F	10	22	32
Total	26	36	62

[See Principles](#)

[See Algorithm](#)

[See Implementation](#)

[See Analytical Characterization](#)

# Presentation of Generalized Results

- Generalized relation:
  - Relations where some or all attributes are generalized, with counts or other aggregation values accumulated.
- Cross tabulation:
  - Mapping results into cross tabulation form (similar to contingency tables).
  - Visualization techniques:
    - Pie charts, bar charts, curves, cubes, and other visual forms.
- Quantitative characteristic rules:
  - Mapping generalized result into characteristic rules with quantitative information associated with it, e.g.,

$grad(x) \wedge male(x) \Rightarrow$   
 $birth\_region(x) = "Canada"[t:53\%] \vee birth\_region(x) = "foreign"[t:47\%].$

# Presentation—Generalized Relation

location	item	sales (in million dollars)	count (in thousands)
Asia	TV	15	300
Europe	TV	12	250
North_America	TV	28	450
Asia	computer	120	1000
Europe	computer	150	1200
North_America	computer	200	1800

Table 5.3: A generalized relation for the sales in 1997.

# Presentation—Crosstab

location \ item	TV		computer		<i>both_items</i>	
	sales	count	sales	count	sales	count
Asia	15	300	120	1000	135	1300
Europe	12	250	150	1200	162	1450
North_America	28	450	200	1800	228	2250
<i>all_regions</i>	45	1000	470	4000	525	5000

Table 5.4: A crosstab for the sales in 1997.

# Implementation by Cube Technology

- Construct a data cube on-the-fly for the given data mining query
  - Facilitate efficient drill-down analysis
  - May increase the response time
  - A balanced solution: precomputation of “subprime” relation
- Use a predefined & precomputed data cube
  - Construct a data cube beforehand
  - Facilitate not only the attribute-oriented induction, but also attribute relevance analysis, dicing, slicing, roll-up and drill-down
  - Cost of cube computation and the nontrivial storage overhead



# Chapter 5: Concept Description: Characterization and Comparison

- What is concept description?
- Data generalization and summarization-based characterization
- [Analytical characterization: Analysis of attribute relevance](#)
- Mining class comparisons: Discriminating between different classes
- Mining descriptive statistical measures in large databases
- Discussion
- Summary

# Characterization vs. OLAP

- Similarity:
  - Presentation of data summarization at multiple levels of abstraction.
  - Interactive drilling, pivoting, slicing and dicing.
- Differences:
  - Automated desired level allocation.
  - Dimension relevance analysis and ranking when there are many relevant dimensions.
  - Sophisticated typing on dimensions and measures.
  - Analytical characterization: data dispersion analysis.

# Attribute Relevance Analysis

- Why?
  - Which dimensions should be included?
  - How high level of generalization?
  - Automatic vs. interactive
  - Reduce # attributes; easy to understand patterns
- What?
  - statistical method for preprocessing data
    - filter out irrelevant or weakly relevant attributes
    - retain or rank the relevant attributes
  - relevance related to dimensions and levels
  - analytical characterization, analytical comparison

# Attribute relevance analysis (cont'd)

- How?
  - Data Collection
  - Analytical Generalization
    - Use information gain analysis (e.g., entropy or other measures) to identify highly relevant dimensions and levels.
  - Relevance Analysis
    - Sort and select the most relevant dimensions and levels.
  - Attribute-oriented Induction for class description
    - On selected dimension/level
  - OLAP operations (e.g. drilling, slicing) on relevance rules

# Relevance Measures

- Quantitative relevance measure determines the classifying power of an attribute within a set of data.
- Methods
  - information gain (ID3)
  - gain ratio (C4.5)
  - gini index
  - $\chi^2$  contingency table statistics
  - uncertainty coefficient

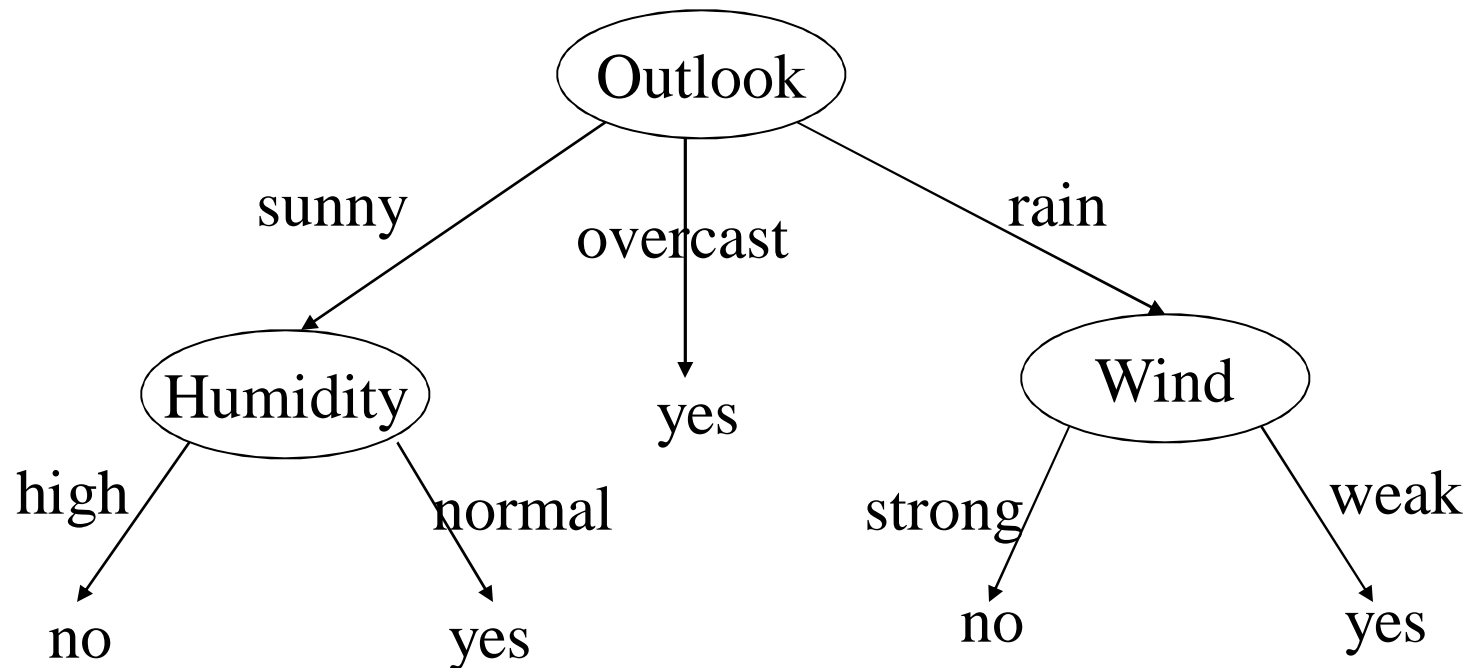
# Information-Theoretic Approach

- Decision tree
  - each internal node tests an attribute
  - each branch corresponds to attribute value
  - each leaf node assigns a classification
- ID3 algorithm
  - build decision tree based on training objects with known class labels to classify testing objects
  - rank attributes with information gain measure
  - minimal height
    - the least number of tests to classify an object

# Top-Down Induction of Decision Tree

Attributes = {Outlook, Temperature, Humidity, Wind}

PlayTennis = {yes, no}



# Entropy and Information Gain

- S contains  $s_i$  tuples of class  $C_i$  for  $i = \{1, \dots, m\}$
- Information measures info required to classify any arbitrary tuple

$$I(s_1, s_2, \dots, s_m) = - \sum_{i=1}^m \frac{s_i}{S} \log_2 \frac{s_i}{S}$$

- Entropy of attribute A with values  $\{a_1, a_2, \dots, a_v\}$

$$E(A) = \sum_{j=1}^v \frac{s_{1j} + \dots + s_{mj}}{S} I(s_{1j}, \dots, s_{mj})$$

- Information gained by branching on attribute A

$$Gain(A) = I(s_1, s_2, \dots, s_m) - E(A)$$



# Example: Analytical Characterization

- Task
  - Mine general characteristics describing graduate students using analytical characterization
- Given
  - attributes *name*, *gender*, *major*, *birth\_place*, *birth\_date*, *phone#*, and *gpa*
  - $Gen(a_i)$  = concept hierarchies on  $a_i$
  - $U_i$  = attribute analytical thresholds for  $a_i$
  - $T_i$  = attribute generalization thresholds for  $a_i$
  - $R$  = attribute relevance threshold

# Example: Analytical Characterization (cont'd)

- 1. Data collection
  - target class: graduate student
  - contrasting class: undergraduate student
- 2. Analytical generalization using  $U_i$ 
  - attribute removal
    - remove *name* and *phone#*
  - attribute generalization
    - generalize *major*, *birth\_place*, *birth\_date* and *grade\_point\_average*
    - accumulate counts
  - **candidate relation**: *gender*, *major*, *birth\_country*, *age\_range* and *gpa*

## Example: Analytical characterization (2)

gender	major	birth_country	age_range	gpa	count
M	Science	Canada	20-25	Very_good	16
F	Science	Foreign	25-30	Excellent	22
M	Engineering	Foreign	25-30	Excellent	18
F	Science	Foreign	25-30	Excellent	25
M	Science	Canada	20-25	Excellent	21
F	Engineering	Canada	20-25	Excellent	18

*Candidate relation for Target class: Graduate students ( $\Sigma=120$ )*

gender	major	birth_country	age_range	gpa	count
M	Science	Foreign	<20	Very_good	18
F	Business	Canada	<20	Fair	20
M	Business	Canada	<20	Fair	22
F	Science	Canada	20-25	Fair	24
M	Engineering	Foreign	20-25	Very_good	22
F	Engineering	Canada	<20	Excellent	24

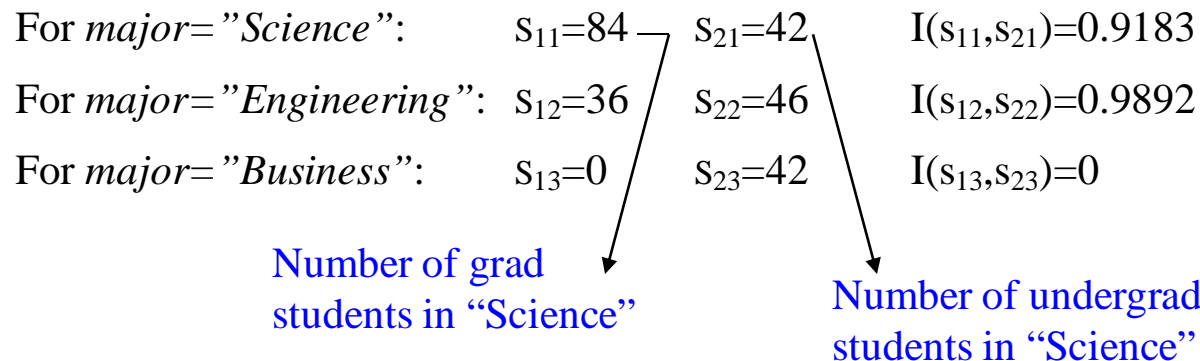
*Candidate relation for Contrasting class: Undergraduate students ( $\Sigma=130$ )*

# Example: Analytical characterization (3)

- 3. Relevance analysis
  - Calculate expected info required to classify an arbitrary tuple

$$I(s_1, s_2) = I(120, 130) = -\frac{120}{250} \log_2 \frac{120}{250} - \frac{130}{250} \log_2 \frac{130}{250} = 0.9988$$

- Calculate entropy of each attribute: e.g. *major*



# Example: Analytical Characterization (4)

- Calculate expected info required to classify a given sample if S is partitioned according to the attribute

$$E(\text{major}) = \frac{126}{250} I(s_{11}, s_{21}) + \frac{82}{250} I(s_{12}, s_{22}) + \frac{42}{250} I(s_{13}, s_{23}) = 0.7873$$

- Calculate information gain for each attribute

$$\text{Gain}(\text{major}) = I(s_1, s_2) - E(\text{major}) = 0.2115$$

– Information gain for all attributes

$$\text{Gain}(\text{gender}) = 0.0003$$

$$\text{Gain}(\text{birth\_country}) = 0.0407$$

$$\text{Gain}(\text{major}) = 0.2115$$

$$\text{Gain}(\text{gpa}) = 0.4490$$

$$\text{Gain}(\text{age\_range}) = 0.5971$$

# Example: Analytical characterization (5)

- 4. Initial working relation ( $W_0$ ) derivation
  - $R = 0.1$
  - remove irrelevant/weakly relevant attributes from candidate relation  
=> drop *gender*, *birth\_country*
  - remove contrasting class candidate relation

major	age_range	gpa	count
Science	20-25	Very_good	16
Science	25-30	Excellent	47
Science	20-25	Excellent	21
Engineering	20-25	Excellent	18
Engineering	25-30	Excellent	18

**Initial target class working relation  $W_0$ : Graduate students**

- 5. Perform attribute-oriented induction on  $W_0$  using  $T_i$

# Chapter 5: Concept Description: Characterization and Comparison

- What is concept description?
- Data generalization and summarization-based characterization
- Analytical characterization: Analysis of attribute relevance
- Mining class comparisons: Discriminating between different classes
- Mining descriptive statistical measures in large databases
- Discussion
- Summary

# Mining Class Comparisons

- Comparison: Comparing two or more classes.
- Method:
  - Partition the set of relevant data into the target class and the contrasting class(es)
  - Generalize both classes to the same high level concepts
  - Compare tuples with the same high level descriptions
  - Present for every tuple its description and two measures:
    - support - distribution within single class
    - comparison - distribution between classes
  - Highlight the tuples with strong discriminant features
- Relevance Analysis:
  - Find attributes (features) which best distinguish different classes.



# Example: Analytical comparison

- Task
  - Compare graduate and undergraduate students using discriminant rule.
  - DMQL query

```
use Big_University_DB
mine comparison as "grad_vs_undergrad_students"
in relevance to name, gender, major, birth_place, birth_date, residence, phone#, gpa
for "graduate_students"
where status in "graduate"
versus "undergraduate_students"
where status in "undergraduate"
analyze count%
from student
```

# Example: Analytical comparison (2)

- Given
  - attributes *name, gender, major, birth\_place, birth\_date, residence, phone#* and *gpa*
  - $Gen(a_i)$  = concept hierarchies on attributes  $a_i$
  - $U_i$  = attribute analytical thresholds for attributes  $a_i$
  - $T_i$  = attribute generalization thresholds for attributes  $a_i$
  - $R$  = attribute relevance threshold

# Example: Analytical comparison (3)

- 1. Data collection
  - target and contrasting classes
- 2. Attribute relevance analysis
  - remove attributes *name, gender, major, phone#*
- 3. Synchronous generalization
  - controlled by user-specified dimension thresholds
  - prime target and contrasting class(es) relations/cuboids

# Example: Analytical comparison (4)

Birth_country	Age_range	Gpa	Count%
Canada	20-25	Good	5.53%
Canada	25-30	Good	2.32%
Canada	Over_30	Very_good	5.86%
...	...	...	...
Other	Over_30	Excellent	4.68%

**Prime generalized relation for the target class: Graduate students**

Birth_country	Age_range	Gpa	Count%
Canada	15-20	Fair	5.53%
Canada	15-20	Good	4.53%
...	...	...	...
Canada	25-30	Good	5.02%
...	...	...	...
Other	Over_30	Excellent	0.68%

**Prime generalized relation for the contrasting class: Undergraduate students**

# Example: Analytical comparison (5)

- 4. Drill down, roll up and other OLAP operations on target and contrasting classes to adjust levels of abstractions of resulting description
- 5. Presentation
  - as generalized relations, crosstabs, bar charts, pie charts, or rules
  - contrasting measures to reflect comparison between target and contrasting classes
    - e.g. count%

# Quantitative Discriminant Rules

- $C_j$  = target class
- $q_a$  = a generalized tuple covers some tuples of class
  - but can also cover some tuples of contrasting class
- d-weight
  - range:  $[0, 1]$

$$d\text{-weight} = \frac{\text{count}(q_a \in C_j)}{\sum_{i=1}^m \text{count}(q_a \in C_i)}$$

- quantitative discriminant rule form

$$\forall X, \text{target\_class}(X) \Leftarrow \text{condition}(X) \quad [d : d\_weight]$$

# Example: Quantitative Discriminant Rule

Status	Birth_country	Age_range	Gpa	Count
Graduate	Canada	25-30	Good	90
Undergraduate	Canada	25-30	Good	210

Count distribution between graduate and undergraduate students for a generalized tuple

- Quantitative discriminant rule

$\forall X, graduate\_student(X) \Leftarrow$

$birth\_country(X) = "Canada" \wedge age\_range(X) = "25-30" \wedge gpa(X) = "good" [d:30\%]$

– where  $90/(90+210) = 30\%$

# Class Description

- Quantitative characteristic rule

$\forall X, \text{target\_class}(X) \Rightarrow \text{condition}(X) \ [t : t\_weight]$   
– necessary

- Quantitative discriminant rule

$\forall X, \text{target\_class}(X) \Leftarrow \text{condition}(X) \ [d : d\_weight]$   
– sufficient

- Quantitative description rule

$\forall X, \text{target\_class}(X) \Leftrightarrow$   
 $\text{condition}_1(X) [t : w_1, d : w'_1] \vee \dots \vee \text{condition}_n(X) [t : w_n, d : w'_n]$   
– necessary and sufficient



# Example: Quantitative Description Rule

Location/item	TV			Computer			Both_items		
	Count	t-wt	d-wt	Count	t-wt	d-wt	Count	t-wt	d-wt
Europe	80	25%	40%	240	75%	30%	320	100%	32%
N_Am	120	17.65%	60%	560	82.35%	70%	680	100%	68%
Both_regions	200	20%	100%	800	80%	100%	1000	100%	100%

Crosstab showing associated t-weight, d-weight values and total number (in thousands) of TVs and computers sold at AllElectronics in 1998

- Quantitative description rule for target class *Europe*

$$\forall X, Europe(X) \Leftrightarrow$$

$$(item(X) = "TV" ) [t : 25\%, d : 40\%] \vee (item(X) = "computer" ) [t : 75\%, d : 30\%]$$

# Chapter 5: Concept Description: Characterization and Comparison

- What is concept description?
- Data generalization and summarization-based characterization
- Analytical characterization: Analysis of attribute relevance
- Mining class comparisons: Discriminating between different classes
- Mining descriptive statistical measures in large databases
- Discussion
- Summary

# Mining Data Dispersion 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  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ 
  - Weighted arithmetic mean

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

- Median: A holistic measure

- Middle value if odd number of values, or average of the middle two values otherwise
- estimated by interpolation

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

- Mode

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

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

# Measuring the Dispersion of Data

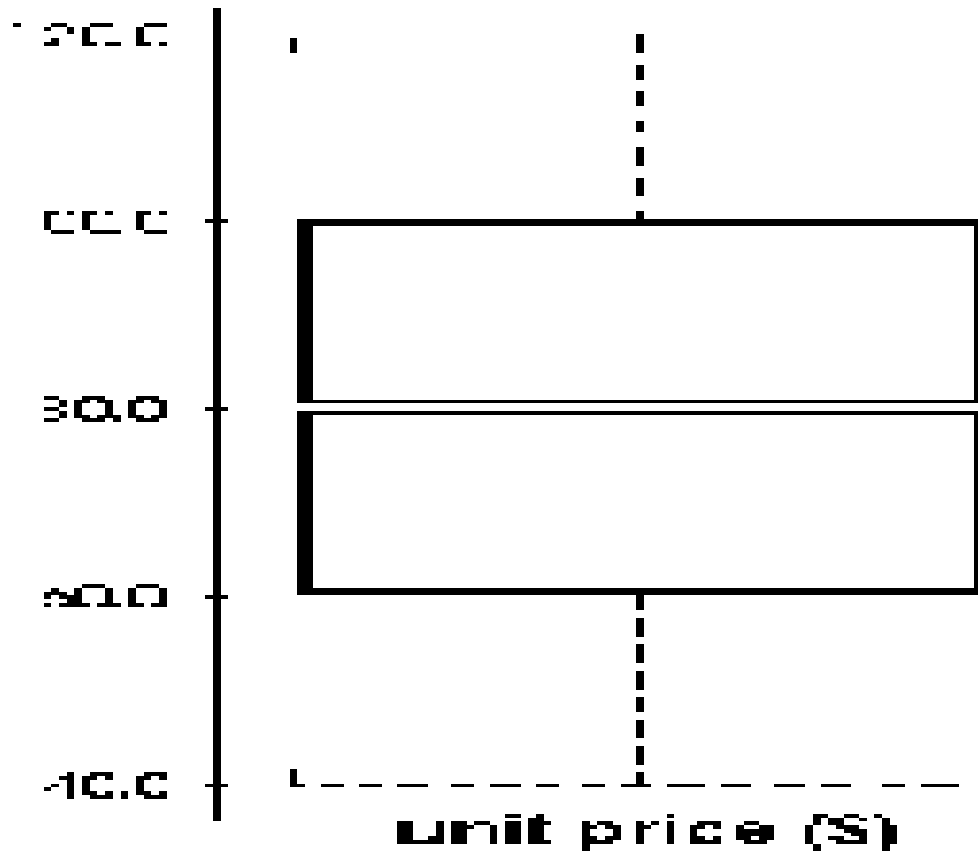
- Quartiles, outliers and boxplots
  - **Quartiles**:  $Q_1$  (25<sup>th</sup> percentile),  $Q_3$  (75<sup>th</sup> percentile)
  - **Inter-quartile range**:  $IQR = Q_3 - Q_1$
  - **Five number summary**: min,  $Q_1$ , M,  $Q_3$ , max
  - **Boxplot**: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually
  - **Outlier**: usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation
  - **Variance**  $s^2$ : (algebraic, scalable computation)
$$s^2 = \frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n - 1} \left[ \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \right]$$
  - **Standard deviation**  $s$  is the square root of variance  $s^2$

# Boxplot Analysis

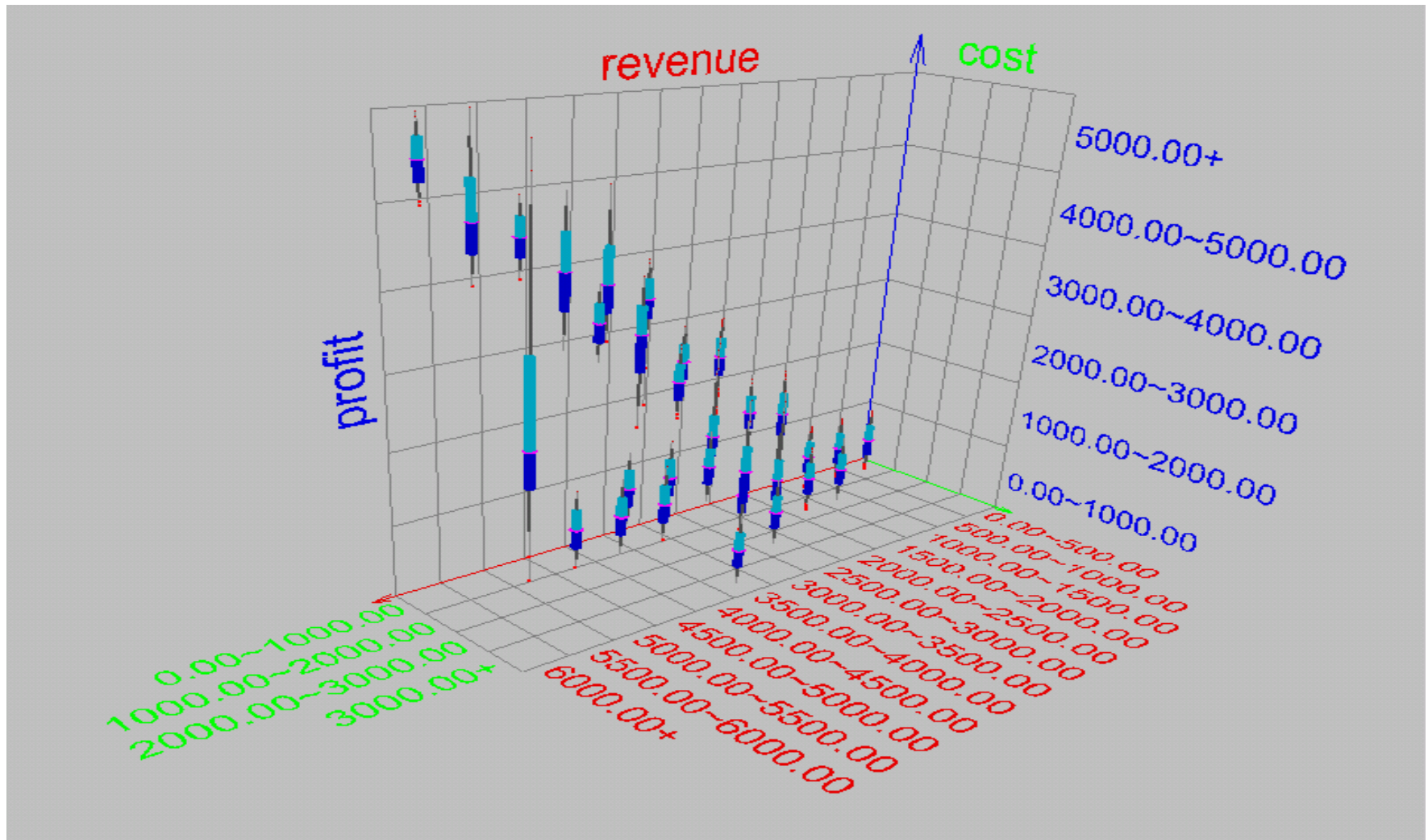
- **Five-number summary** of a distribution:  
Minimum, Q1, M, Q3, Maximum
- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IRQ
  - The median is marked by a line within the box
  - Whiskers: two lines outside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

# A Boxplot

A boxplot



# Visualization of Data Dispersion: Boxplot Analysis





# Mining Descriptive Statistical Measures in Large Databases

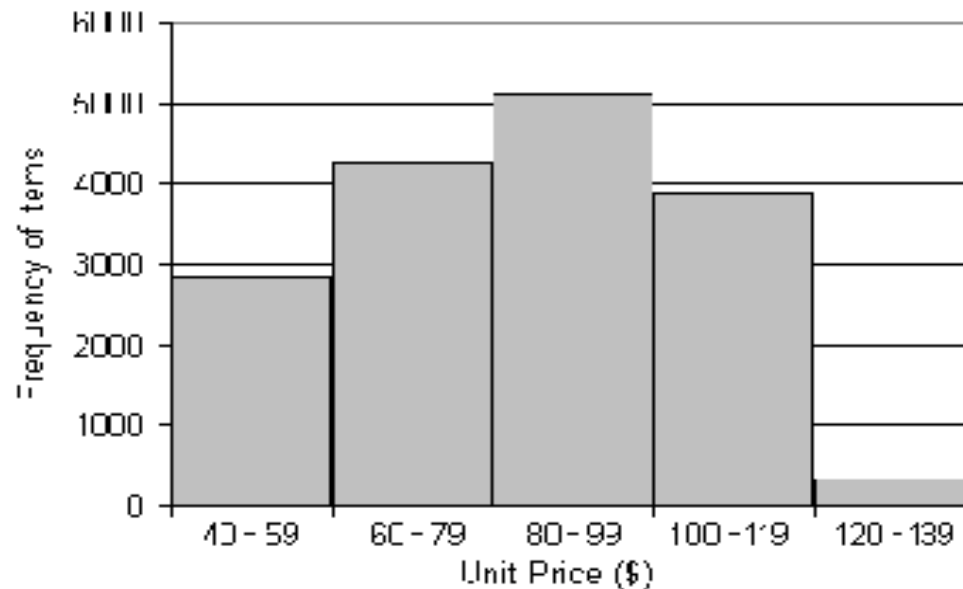
- Variance

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

- **Standard deviation**: the square root of the variance
  - Measures spread about the mean
  - It is zero if and only if all the values are equal
  - Both the deviation and the variance are algebraic

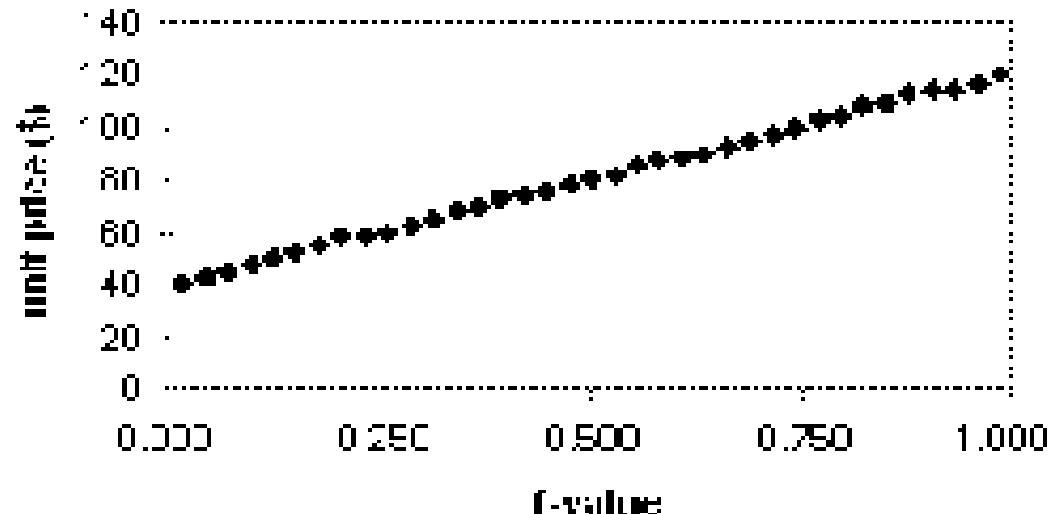
# Histogram Analysis

- 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



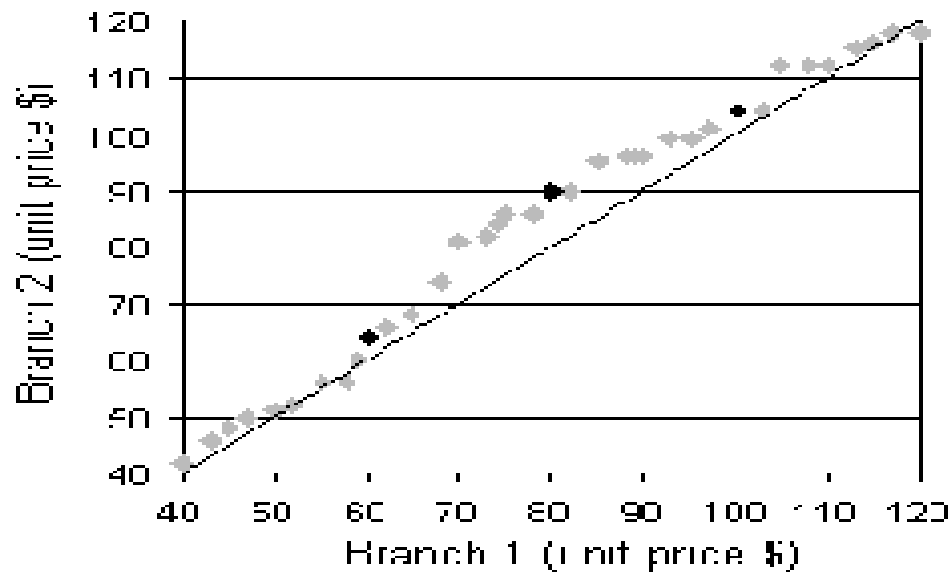
# Quantile Plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots **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  $x_i$



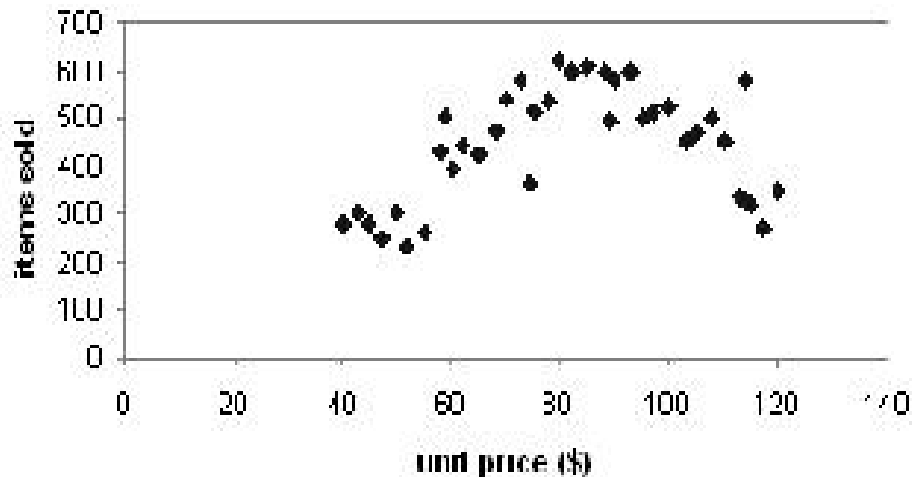
# Quantile-Quantile (Q-Q) Plot

- 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



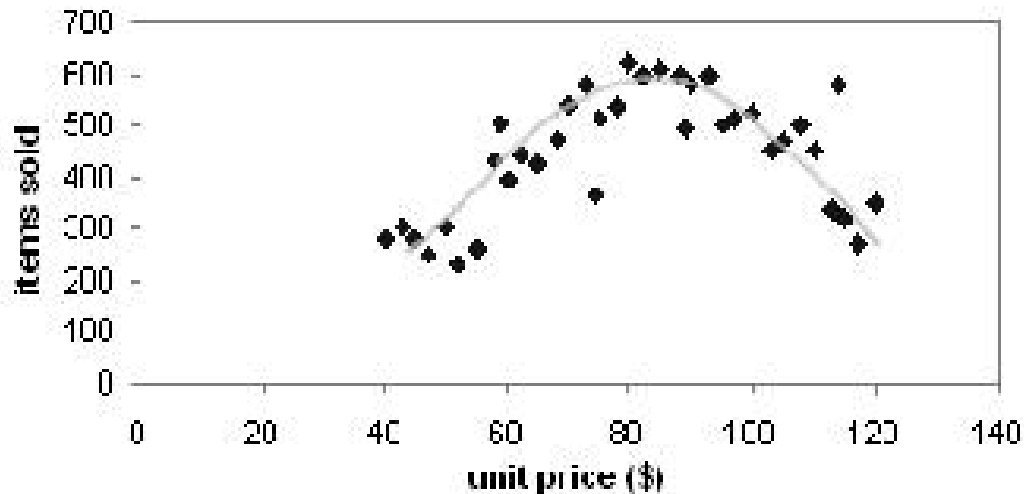
# Scatter plot

- 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 of dependence
- Loess (**local regression**) 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

- 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

# Chapter 5: Concept Description: Characterization and Comparison

- What is concept description?
- Data generalization and summarization-based characterization
- Analytical characterization: Analysis of attribute relevance
- Mining class comparisons: Discriminating between different classes
- Mining descriptive statistical measures in large databases
- Discussion
- Summary

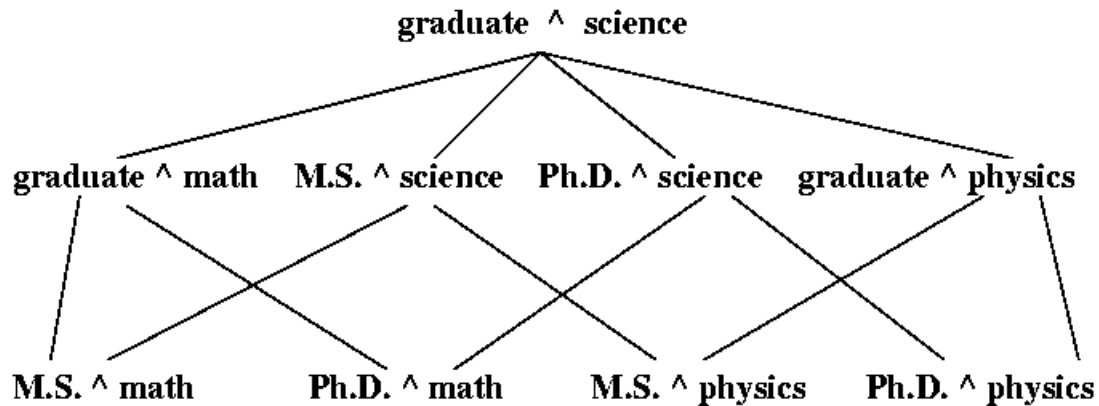


# Attribute Oriented Induction vs. Learning-from-example Paradigm

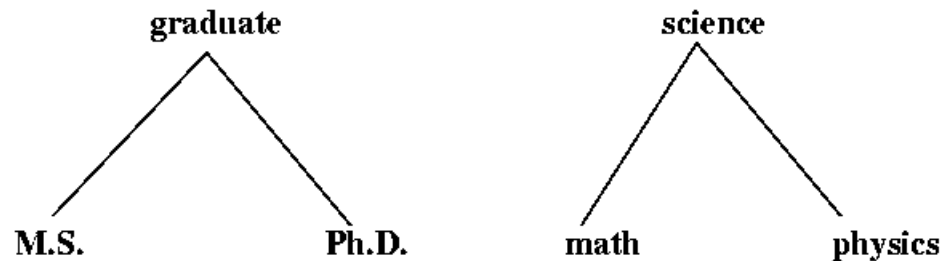
- Difference in philosophies and basic assumptions
  - Positive and negative samples in learning-from-example: positive used for generalization, negative - for specialization (not always)
  - Positive samples only in data mining: hence generalization-based, to drill-down backtrack the generalization to a previous state
- Difference in methods of generalizations
  - Machine learning generalizes on a tuple by tuple basis
  - Data mining generalizes on an attribute by attribute basis

# Comparison of Entire vs. Factored Version Space

The entire version space



The factored version space



# Incremental and Parallel Mining of Concept Description

- Incremental mining: revision based on newly added data  $\Delta DB$ 
  - Generalize  $\Delta DB$  to the same level of abstraction in the generalized relation  $R$  to derive  $\Delta R$
  - Union  $R \cup \Delta R$ , i.e., merge counts and other statistical information to produce a new relation  $R'$
- Similar philosophy can be applied to data sampling, parallel and/or distributed mining, etc.

# Chapter 5: Concept Description: Characterization and Comparison

- What is concept description?
- Data generalization and summarization-based characterization
- Analytical characterization: Analysis of attribute relevance
- Mining class comparisons: Discriminating between different classes
- Mining descriptive statistical measures in large databases
- Discussion
- [Summary](#)

# Summary

- Concept description: characterization and discrimination
- OLAP-based vs. attribute-oriented induction
- Efficient implementation of AOI
- Analytical characterization and comparison
- Mining descriptive statistical measures in large databases
- Discussion
  - Incremental and parallel mining of description
  - Descriptive mining of complex types of data

# References

- Y. Cai, N. Cercone, and J. Han. Attribute-oriented induction in relational databases. In G. Piatetsky-Shapiro and W. J. Frawley, editors, *Knowledge Discovery in Databases*, pages 213-228. AAAI/MIT Press, 1991.
- S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. *ACM SIGMOD Record*, 26:65-74, 1997
- C. Carter and H. Hamilton. Efficient attribute-oriented generalization for knowledge discovery from large databases. *IEEE Trans. Knowledge and Data Engineering*, 10:193-208, 1998.
- W. Cleveland. *Visualizing Data*. Hobart Press, Summit NJ, 1993.
- J. L. Devore. *Probability and Statistics for Engineering and the Science*, 4th ed. Duxbury Press, 1995.
- T. G. Dietterich and R. S. Michalski. A comparative review of selected methods for learning from examples. In Michalski et al., editor, *Machine Learning: An Artificial Intelligence Approach*, Vol. 1, pages 41-82. Morgan Kaufmann, 1983.
- J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. *Data Mining and Knowledge Discovery*, 1:29-54, 1997.
- J. Han, Y. Cai, and N. Cercone. Data-driven discovery of quantitative rules in relational databases. *IEEE Trans. Knowledge and Data Engineering*, 5:29-40, 1993.

# References (cont.)

- J. Han and Y. Fu. Exploration of the power of attribute-oriented induction in data mining. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 399-421. AAAI/MIT Press, 1996.
- R. A. Johnson and D. A. Wichern. *Applied Multivariate Statistical Analysis*, 3rd ed. Prentice Hall, 1992.
- E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. VLDB'98, New York, NY, Aug. 1998.
- H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, 1998.
- R. S. Michalski. A theory and methodology of inductive learning. In Michalski et al., editor, *Machine Learning: An Artificial Intelligence Approach*, Vol. 1, Morgan Kaufmann, 1983.
- T. M. Mitchell. Version spaces: A candidate elimination approach to rule learning. IJCAI'97, Cambridge, MA.
- T. M. Mitchell. Generalization as search. *Artificial Intelligence*, 18:203-226, 1982.
- T. M. Mitchell. *Machine Learning*. McGraw Hill, 1997.
- J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81-106, 1986.
- D. Subramanian and J. Feigenbaum. Factorization in experiment generation. AAAI'86, Philadelphia, PA, Aug. 1986.