

Data Mining: Concepts and Techniques

— Chapter 1 —
— Introduction —
— Information —
— Chapter 1 —

Lingma Acheson
Department of Computer and Information Science, IUPUI
linglu@iupui.edu

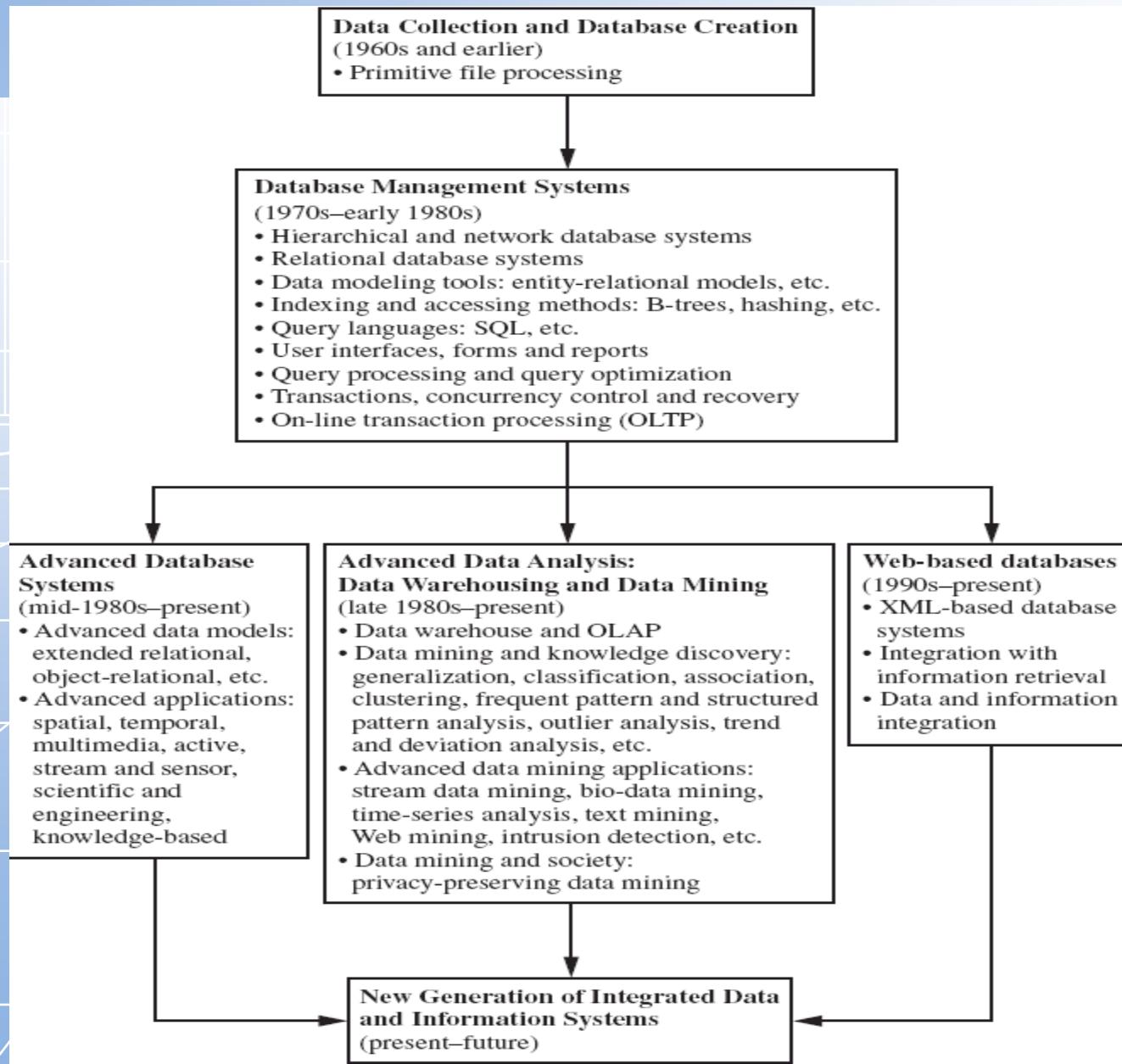
Outline

- 1.1 Motivation: Why data mining?
- 1.2 What is data mining?
- 1.3 Data Mining: On what kind of data?
- 1.4 Data mining functionality: What kinds of Patterns Can Be Mined?
- 1.5 Are all the patterns interesting?
- 1.6 Classification of data mining systems
- 1.7 Data Mining Task Primitives
- 1.8 Integration of data mining system with a DB and DW System
- 1.9 Major issues in data mining

1.1 Why Data Mining?

- The Explosive Growth of Data: from terabytes(1000^4) to yottabytes(1000^8)
 - Data collection and data availability
 - Automated data collection tools, database systems, web
 - Major sources of abundant data
 - Business: Web, e-commerce, transactions, stocks, ...
 - Science: bioinformatics, scientific simulation, medical research ...
 - Society and everyone: news, digital cameras, ...
- Data rich but information poor!
 - What does those data mean?
 - How to analyze data?
- Data mining — Automated analysis of massive data sets

Evolution of Database Technology



1.2 What Is Data Mining?



- Data mining (knowledge discovery from data)
 - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
 - Data mining: a misnomer?
- Alternative names
 - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.



Potential Applications

- Data analysis and decision support
 - Market analysis and management
 - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
 - Risk analysis and management
 - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
 - Fraud detection and detection of unusual patterns (outliers)
- Other Applications
 - Text mining (news group, email, documents) and Web mining
 - Stream data mining
 - Bioinformatics and bio-data analysis

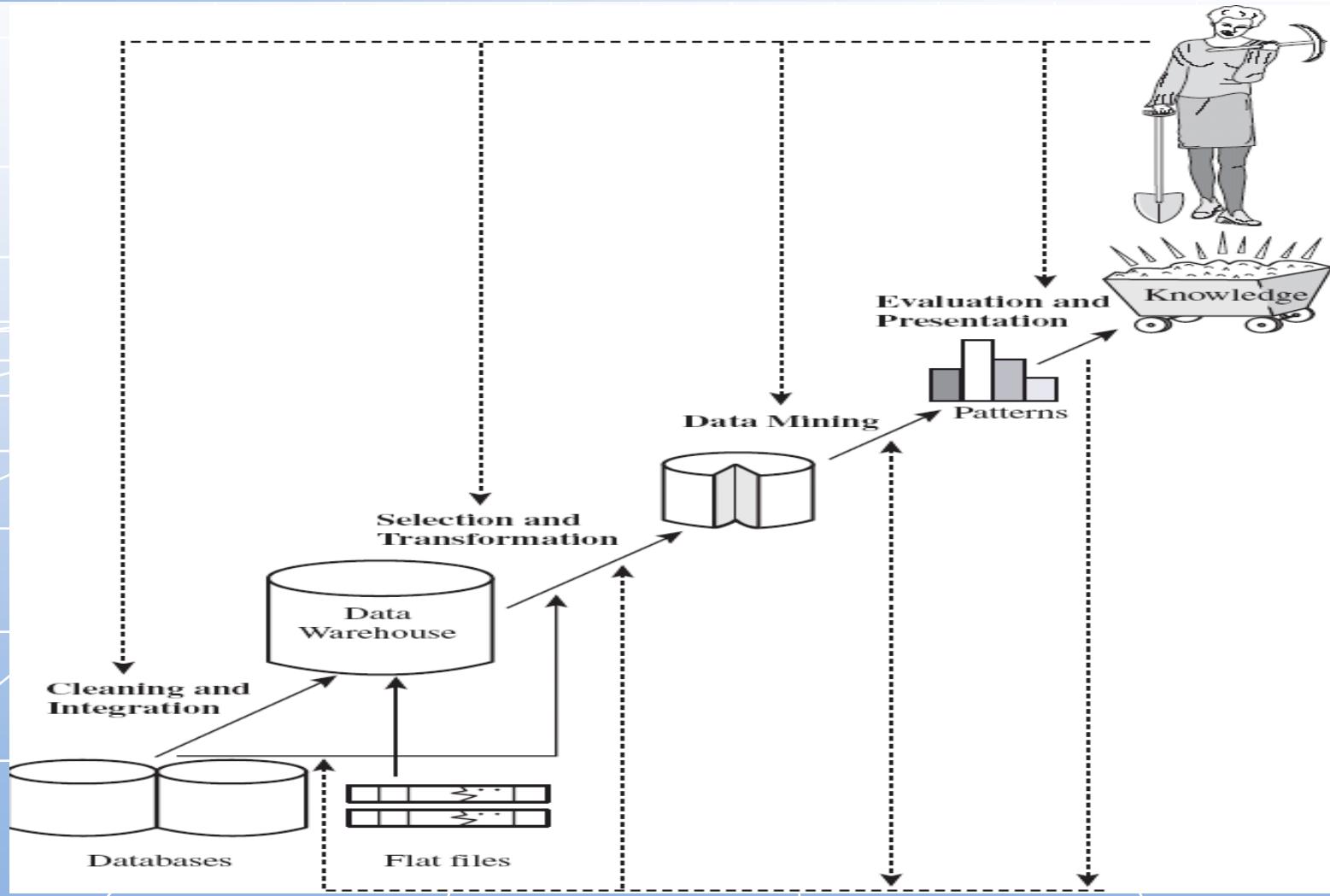
Ex.: Market Analysis and Management

- Where does the data come from?—Credit card transactions, loyalty cards, discount coupons, customer complaint calls, surveys ...
- Target marketing
 - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.,
 - E.g. Most customers with income level 60k – 80k with food expenses \$600 - \$800 a month live in that area
 - Determine customer purchasing patterns over time
 - E.g. Customers who are between 20 and 29 years old, with income of 20k – 29k usually buy this type of CD player
- Cross-market analysis—Find associations/co-relations between product sales, & predict based on such association
 - E.g. Customers who buy computer A usually buy software B

Ex.: Market Analysis and Management (2)

- Customer requirement analysis
 - Identify the best products for different customers
 - Predict what factors will attract new customers
- Provision of summary information
 - Multidimensional summary reports
 - E.g. Summarize all transactions of the first quarter from three different branches
 - Summarize all transactions of last year from a particular branch
 - Summarize all transactions of a particular product
 - Statistical summary information
 - E.g. What is the average age for customers who buy product A?
- Fraud detection
 - Find outliers of unusual transactions
- Financial planning
 - Summarize and compare the resources and spending

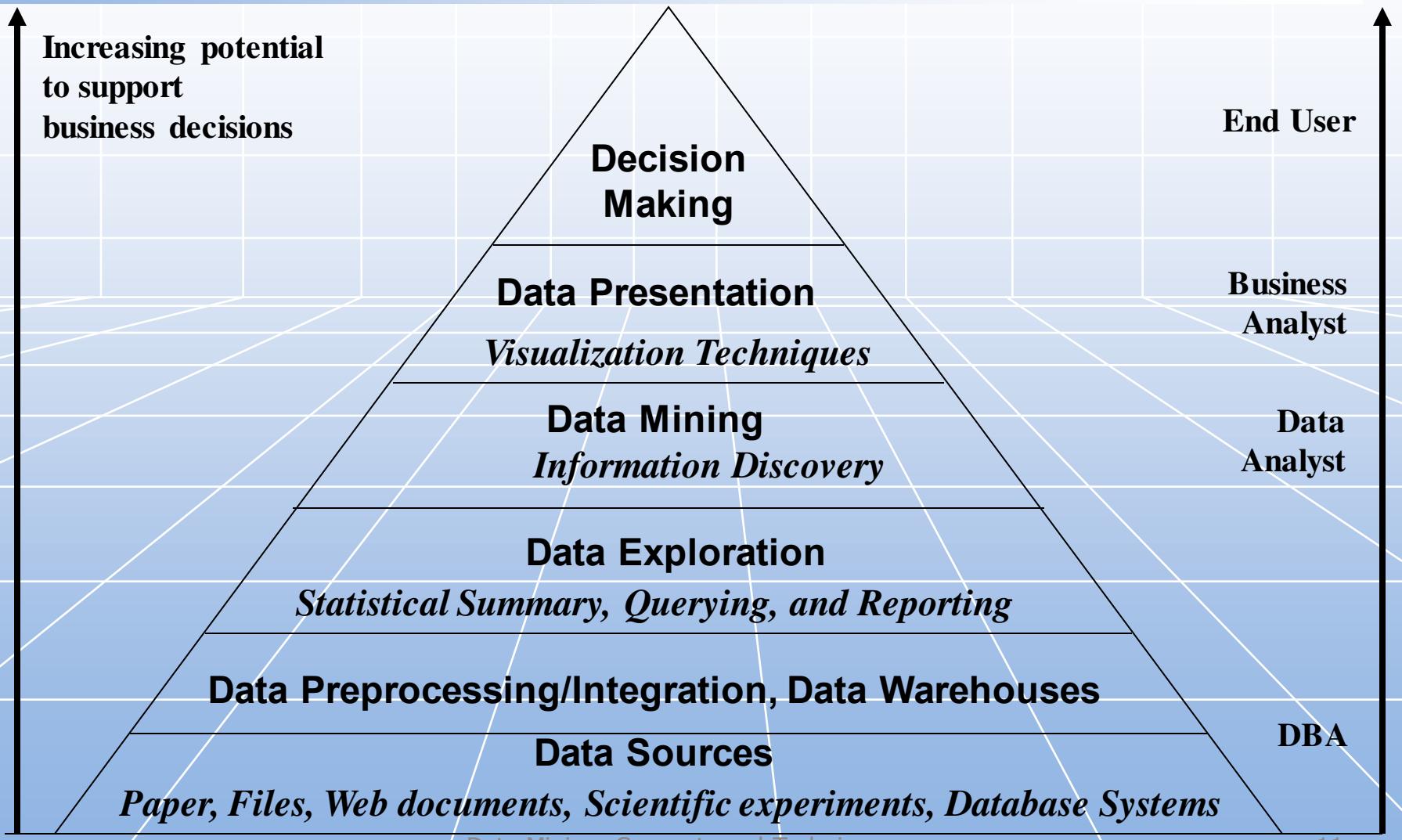
Knowledge Discovery (KDD) Process



KDD Process: Several Key Steps

- Learning the application domain
 - relevant prior knowledge and goals of application
- Identifying a target data set: data selection
- Data processing
 - **Data cleaning** (remove noise and inconsistent data)
 - **Data integration** (multiple data sources maybe combined)
 - **Data selection** (data relevant to the analysis task are retrieved from database)
 - **Data transformation** (data transformed or consolidated into forms appropriate for mining)
(Done with data preprocessing)
 - **Data mining** (an essential process where intelligent methods are applied to extract data patterns)
 - **Pattern evaluation** (indentify the truly interesting patterns)
 - **Knowledge presentation** (mined knowledge is presented to the user with visualization or representation techniques)
- Use of discovered knowledge

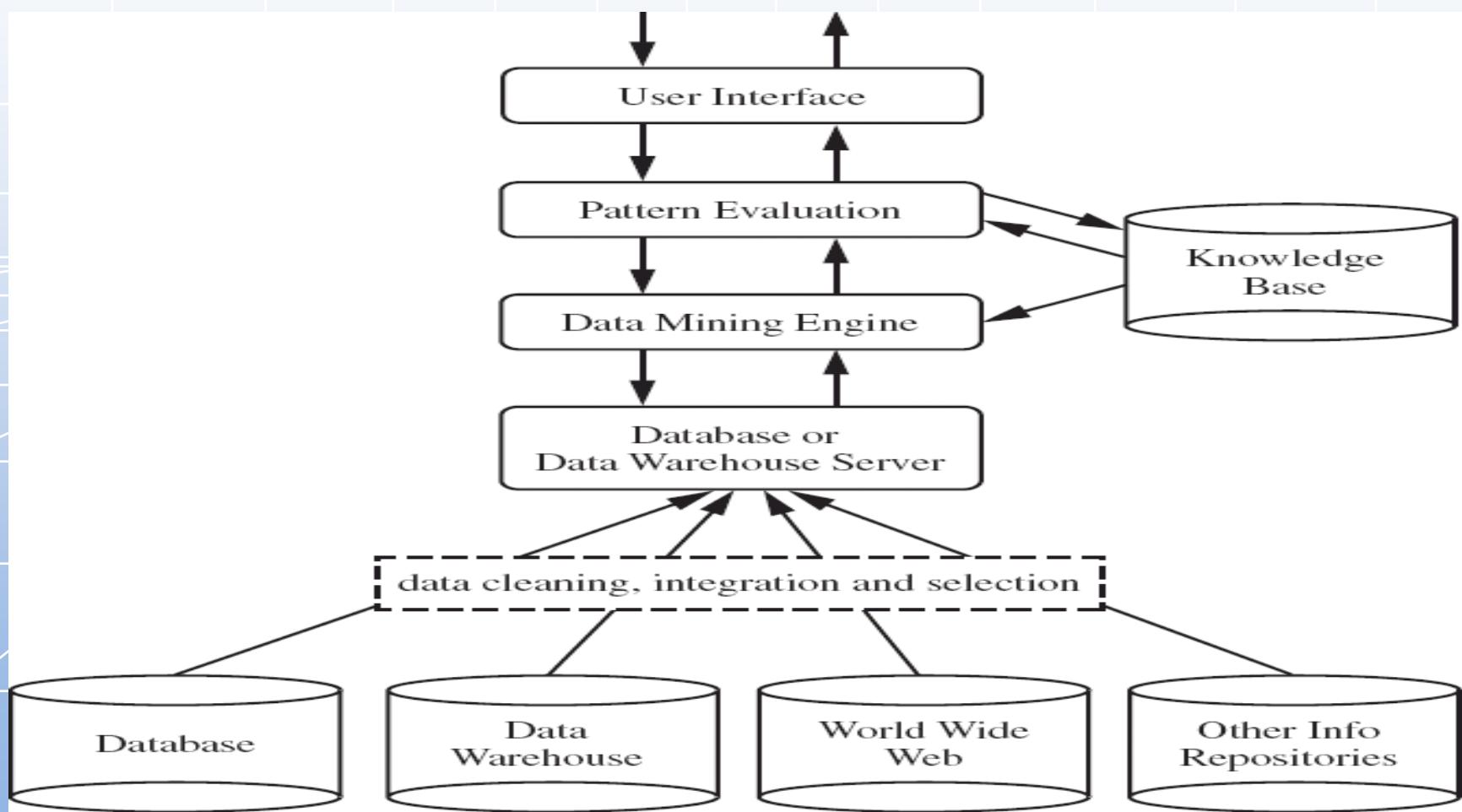
Data Mining and Business Intelligence



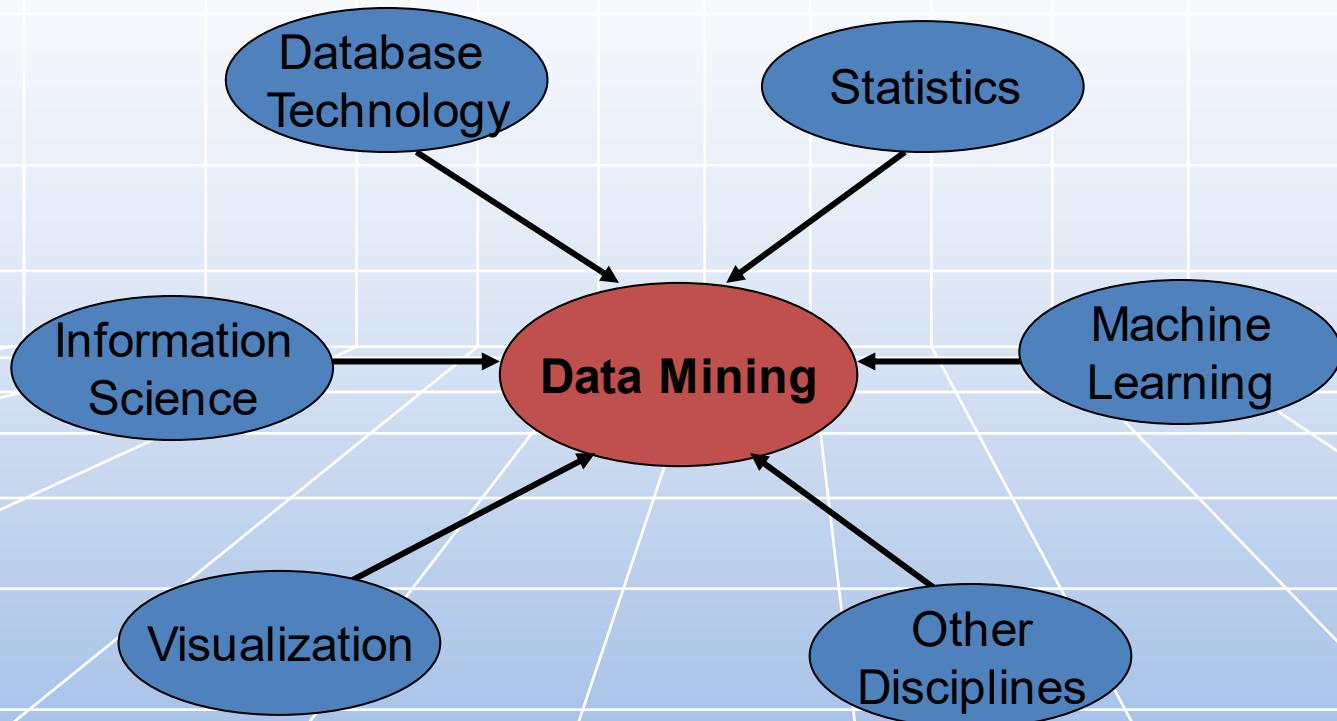
A typical DM System Architecture

- Database, data warehouse, WWW or other information repository (**store data**)
- Database or data warehouse server (**fetch and combine data**)
- Knowledge base (**turn data into meaningful groups according to domain knowledge**)
- Data mining engine (**perform mining tasks**)
- Pattern evaluation module (**find interesting patterns**)
- User interface (**interact with the user**)

A typical DM System Architecture (2)



Confluence of Multiple Disciplines



- Not all “Data Mining System” performs true data mining
 - machine learning system, statistical analysis (small amount of data)
 - Database system (information retrieval, deductive querying...)

1.3 On What Kinds of Data?

- Database-oriented data sets and applications
 - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
 - Object-Relational Databases
 - Temporal Databases, Sequence Databases, Time-Series databases
 - Spatial Databases and Spatiotemporal Databases
 - Text databases and Multimedia databases
 - Heterogeneous Databases and Legacy Databases
 - Data Streams
 - The World-Wide Web

Relational Databases

- DBMS – database management system, contains a collection of interrelated databases
 - e.g. Faculty database, student database, publications database
- Each database contains a collection of tables and functions to manage and access the data.
 - e.g. student_bio, student_graduation, student_parking
- Each table contains columns and rows, with columns as attributes of data and rows as records.
- Tables can be used to represent the relationships between or among multiple tables.

Relational Databases (2) - AllElectronics store

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>	<u>cust_ID</u>	<u>empl_ID</u>	<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

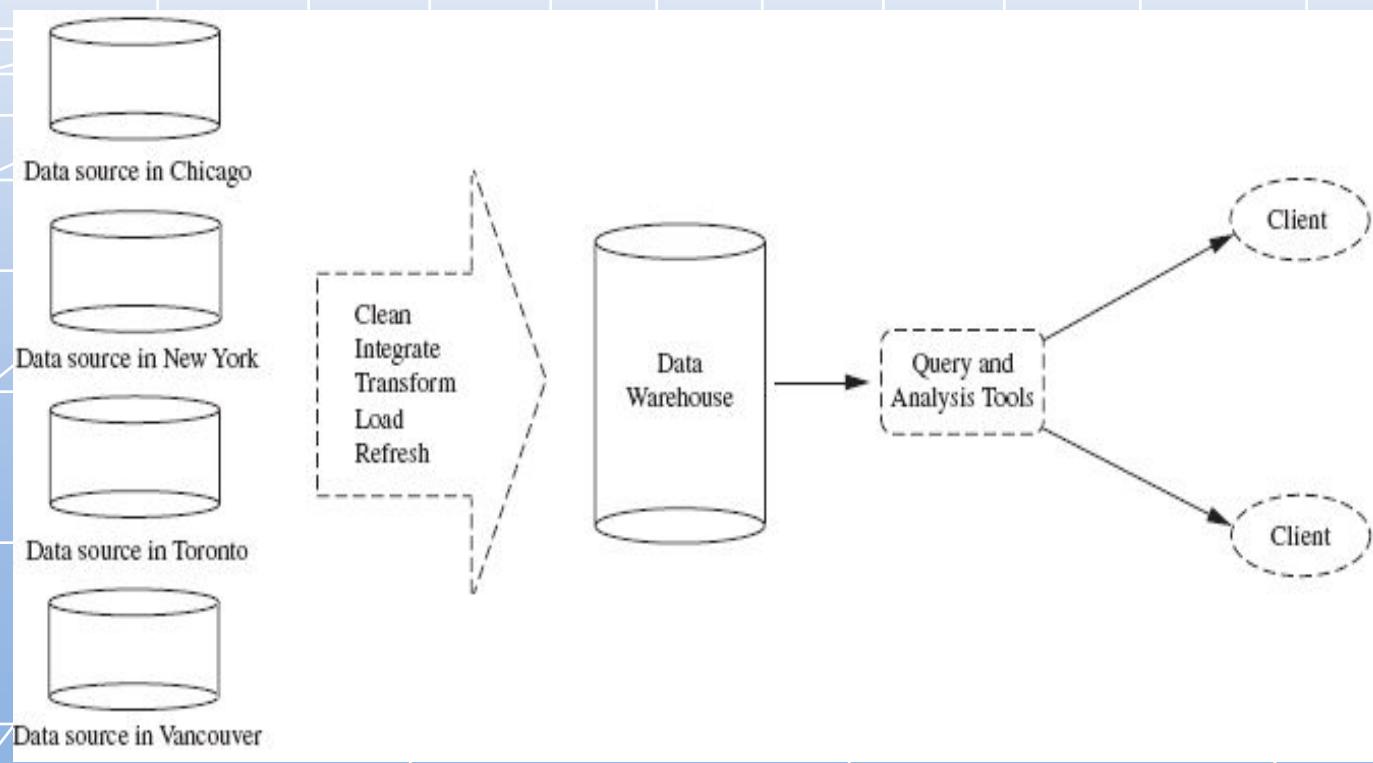
<u>empl_ID</u>	<u>branch_ID</u>
E55	B1
...	...

Relational Databases (3)

- With a relational query language, e.g. SQL, we will be able to find answers to questions such as:
 - How many items were sold last year?
 - Who has earned commissions higher than 10%?
 - What is the total sales of last month for Dell laptops?
- When data mining is applied to relational databases, we can search for trends or data patterns.
- Relational databases are one of the most commonly available and rich information repositories, and thus are a major data form in our study.

Data Warehouses

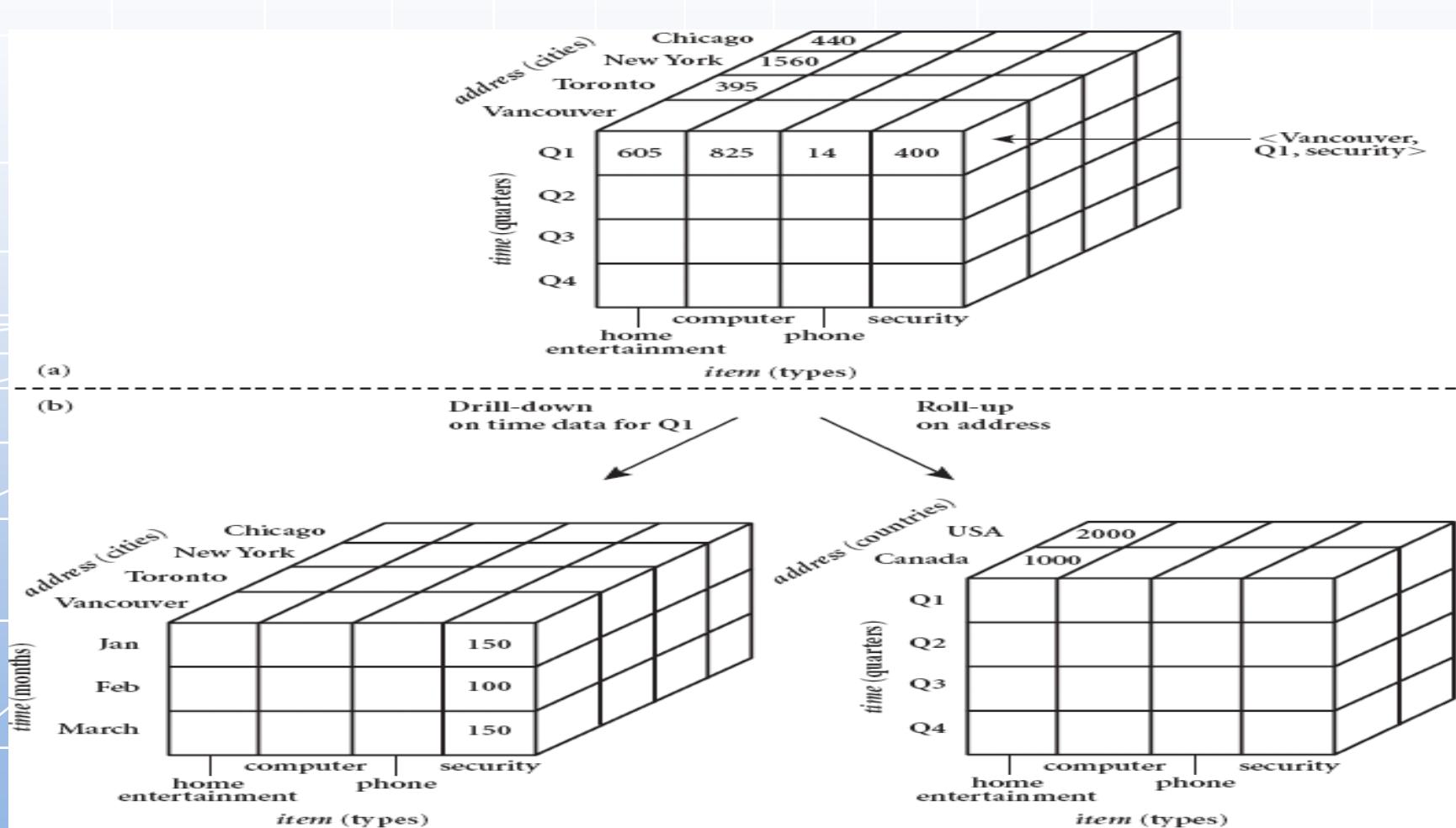
- A repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site.
- Constructed via a process of data cleaning, data integration, data transformation, data loading and periodic data refreshing.



Data Warehouses (2)

- Data are organized around major subjects, e.g. customer, item, supplier and activity.
- Provide information from a historical perspective (e.g. from the past 5 – 10 years)
- Typically summarized to a higher level (e.g. a summary of the transactions per item type for each store)
- User can perform drill-down or roll-up operation to view the data at different degrees of summarization

Data Warehouses (3)



Transactional Databases

- Consists of a file where each record represents a transaction
- A transaction typically includes a unique transaction ID and a list of the items making up the transaction.

<i>trans_ID</i>	<i>list of item_IDs</i>
T100	I1, I3, I8, I16
T200	I2, I8
...	...

- Either stored in a flat file or unfolded into relational tables
- Easy to identify items that are frequently sold together

1.4 Data Mining Functionalities

– What kinds of patterns can be mined?

- Concept/Class Description: Characterization and Discrimination
 - Data can be associated with classes or concepts.
 - E.g. classes of items – computers, printers, ...
concepts of customers – bigSpenders, budgetSpenders, ...
 - How to describe these items or concepts?
 - Descriptions can be derived via
 - Data characterization – summarizing the general characteristics of a target class of data.
 - E.g. summarizing the characteristics of customers who spend more than \$1,000 a year at *AllElectronics*. Result can be a general profile of the customers, such as 40 – 50 years old, employed, have excellent credit ratings.

1.4 Data Mining Functionalities

– What kinds of patterns can be mined?

- Data discrimination – comparing the target class with one or a set of comparative classes
 - E.g. Compare the general features of software products whose sales increase by 10% in the last year with those whose sales decrease by 30% during the same period
 - Or both of the above
- Mining Frequent Patterns, Associations and Correlations
 - Frequent itemset: a set of items that frequently appear together in a transactional data set (e.g. milk and bread)
 - Frequent subsequence: a pattern that customers tend to purchase product A, followed by a purchase of product B

1.4 Data Mining Functionalities

- What kinds of patterns can be mined?
- Association Analysis: find frequent patterns
 - E.g. a sample analysis result – an association rule:
 $\text{buys}(X, \text{"computer"}) \Rightarrow \text{buys}(X, \text{"software"})$ [support = 1%, confidence = 50%]
(if a customer buys a computer, there is a 50% chance that she will buy software.
1% of all of the transactions under analysis showed that computer and software
are purchased together.)
 - Associations rules are discarded as uninteresting if they do not satisfy both a minimum support threshold and a minimum confidence threshold.
- Correlation Analysis: additional analysis to find statistical correlations between associated pairs

1.4 Data Mining Functionalities

– What kinds of patterns can be mined?

- Classification and Prediction
 - Classification
 - The process of finding a model that describes and distinguishes the data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown.
 - The derived model is based on the analysis of a set of training data (data objects whose class label is known).
 - The model can be represented in *classification (IF-THEN) rules*, decision trees, *neural networks*, etc.
 - Prediction
 - Predict missing or unavailable numerical data values

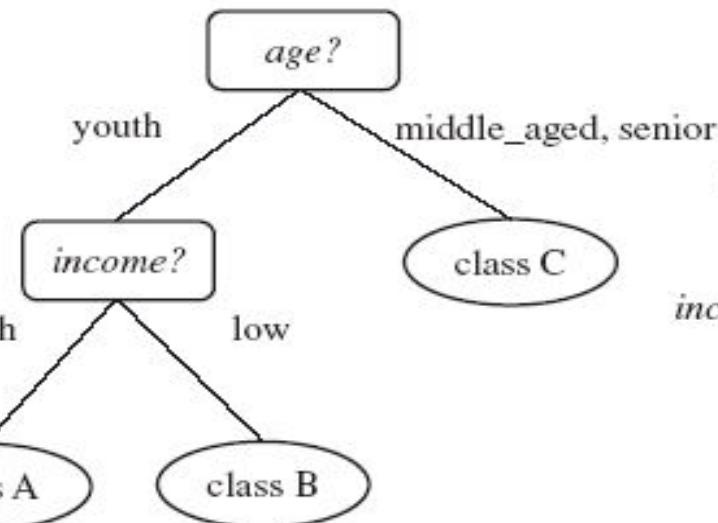
1.4 Data Mining Functionalities

- What kinds of patterns can be mined?

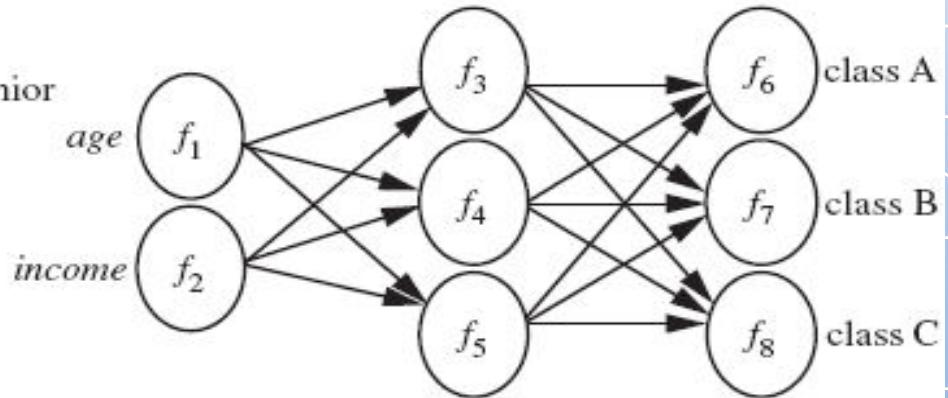
(a)

age(X, "youth") AND income(X, "high") → class(X, "A")
age(X, "youth") AND income(X, "low") → class(X, "B")
age(X, "middle_aged") → class(X, "C")
age(X, "senior") → class(X, "C")

(b)

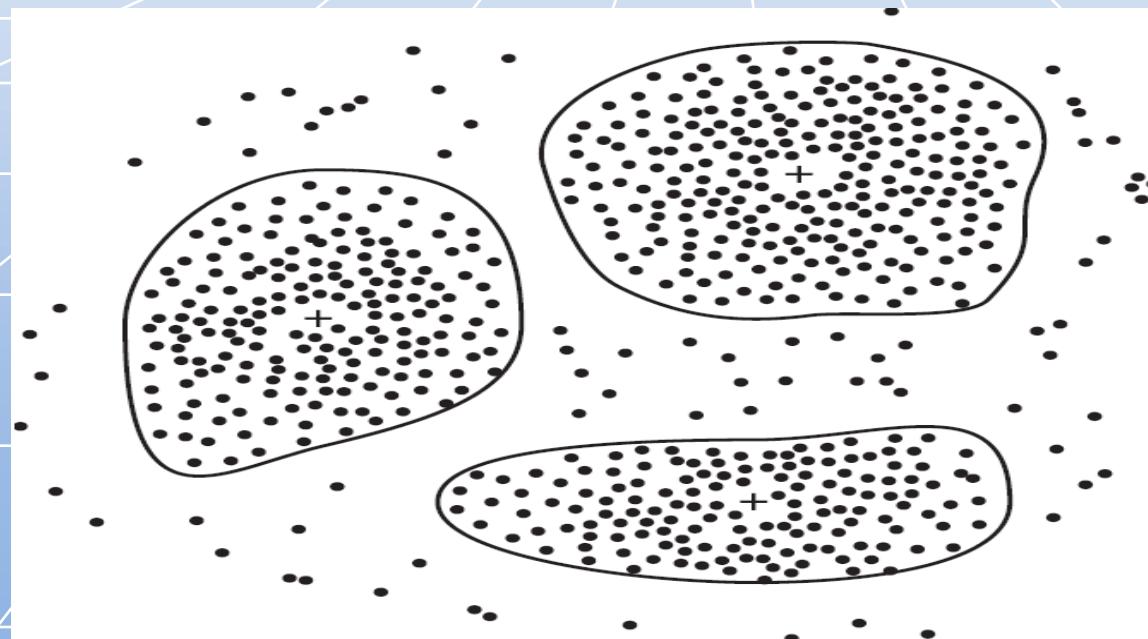


(c)



Data Mining Functionalities (2)

- Cluster Analysis
 - Class label is unknown: group data to form new classes
 - Clusters of objects are formed based on the principle of *maximizing intra-class similarity & minimizing interclass similarity*
 - E.g. Identify homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing.



Data Mining Functionalities (2)

- **Outlier Analysis**
 - Data that do not comply with the general behavior or model.
 - Outliers are usually discarded as noise or exceptions.
 - Useful for fraud detection.
 - E.g. Detect purchases of extremely large amounts
- **Evolution Analysis**
 - Describes and models regularities or trends for objects whose behavior changes over time.
 - E.g. Identify stock evolution regularities for overall stocks and for the stocks of particular companies.

1.5 Are All of the Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting
- A pattern is **interesting** if it is
 - **easily understood** by humans
 - **valid** on new or test data with some degree of certainty,
 - **potentially useful**
 - **novel**
 - **validates some hypothesis** that a user seeks to confirm
- An interesting measure represents ***knowledge*** !

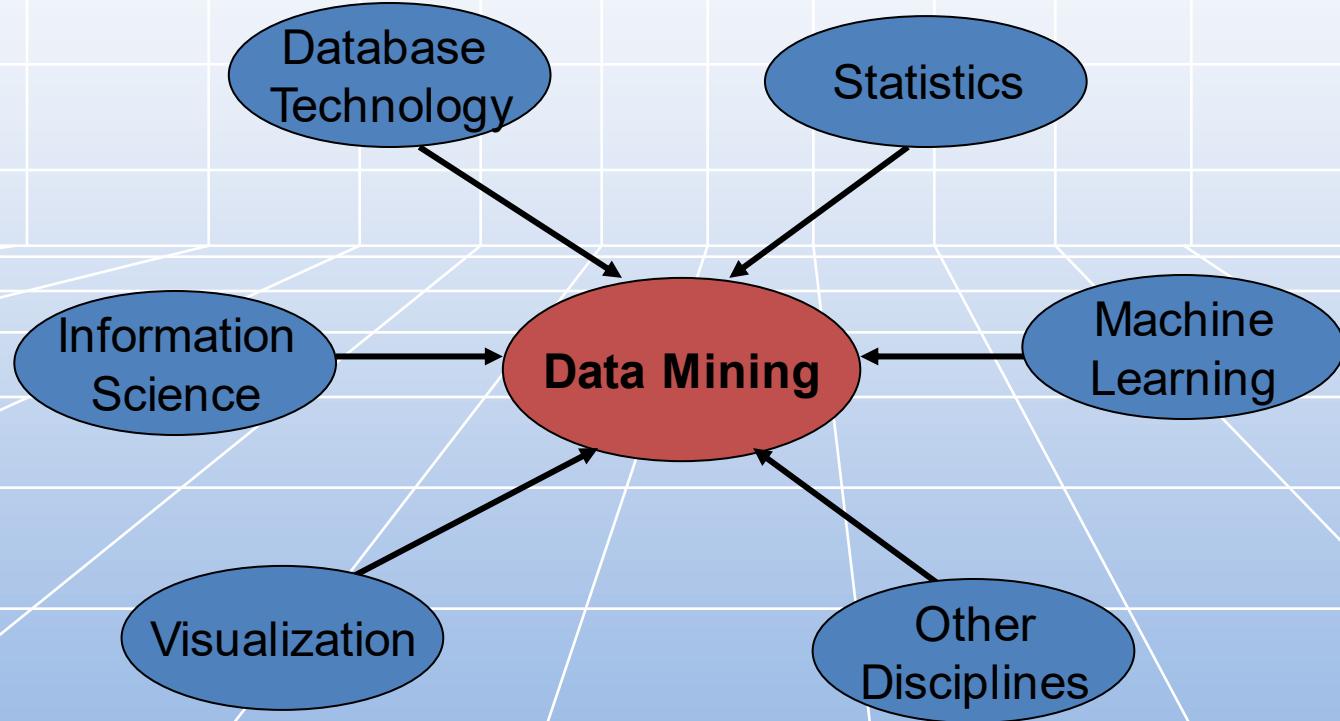
1.5 Are All of the Patterns Interesting?

- Objective measures
 - Based on **statistics and structures of patterns**, e.g., support, confidence, etc.
(Rules that do not satisfy a threshold are considered uninteresting.)
- Subjective measures
 - Reflect the **needs and interests** of a particular user.
 - E.g. A marketing manager is only interested in characteristics of customers who shop frequently.
 - Based on **user's belief** in the data.
 - e.g., Patterns are interesting if they are unexpected, or can be used for strategic planning, etc
- Objective and subjective measures need to be combined.

1.5 Are All of the Patterns Interesting?

- Find **all** the interesting patterns: **Completeness**
 - Unrealistic and inefficient
 - User-provided constraints and interestingness measures should be used
- Search for **only** interesting patterns: An optimization problem
 - Highly desirable
 - No need to search through the generated patterns to identify truly interesting ones.
 - Measures can be used to rank the discovered patterns according their interestingness.

1.6 Classification of data mining systems



1.6 Classification of data mining systems

- **Database**
 - Relational, data warehouse, transactional, stream, object-oriented/relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW
- **Knowledge**
 - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
 - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
 - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, etc.
- **Applications adapted**
 - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

1.7 Data Mining Task Primitives

- How to construct a data mining query?
 - The primitives allow the user to interactively communicate with the data mining system during discovery to direct the mining process, or examine the findings

1.7 Data Mining Task Primitives

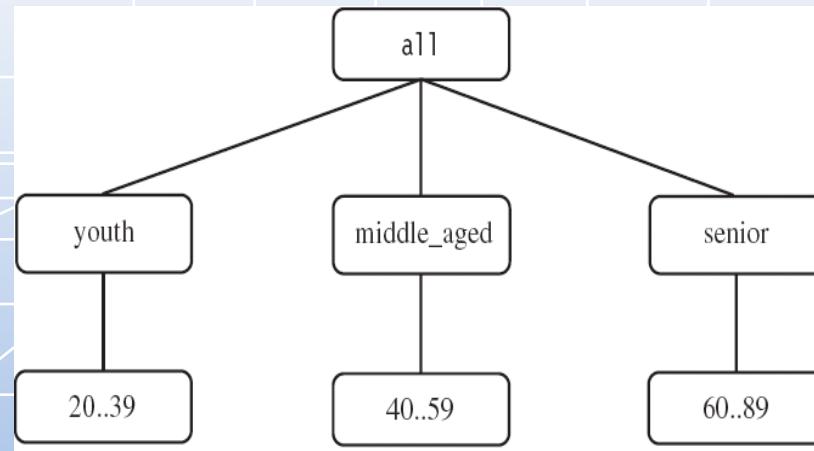
- The primitives specify:
 - (1) The set of task-relevant data – which portion of the database to be used
 - Database or data warehouse name
 - Database tables or data warehouse cubes
 - Condition for data selection
 - Relevant attributes or dimensions
 - Data grouping criteria

1.7 Data Mining Task Primitives

- The primitives specify:
 - (2) The kind of knowledge to be mined – what DB functions to be performed
 - Characterization
 - Discrimination
 - Association
 - Classification/prediction
 - Clustering
 - Outlier analysis
 - Other data mining tasks

1.7 Data Mining Task Primitives

(3) The background knowledge to be used – what domain knowledge, concept hierarchies, etc.



(4) Interestingness measures and thresholds – support, confidence, etc.

(5) Visualization methods – what form to display the result, e.g. rules, tables, charts, graphs, ...

1.7 Data Mining Task Primitives

- DMQL – Data Mining Query Language
 - Designed to incorporate these primitives
 - Allow user to interact with DM systems
 - Providing a **standardized language** like SQL

An Example Query in DMQL

Example 1.11 Mining classification rules. Suppose, as a marketing manager of *AllElectronics*, you would like to classify customers based on their buying patterns. You are especially interested in those customers whose salary is no less than \$40,000, and who have bought more than \$1,000 worth of items, each of which is priced at no less than \$100. In particular, you are interested in the customer's age, income, the types of items purchased, the purchase location, and where the items were made. You would like to view the resulting classification in the form of rules. This data mining query is expressed in DMQL³ as follows, where each line of the query has been enumerated to aid in our discussion.

```
use database AllElectronics_db      (1)
use hierarchy location_hierarchy for T.branch, age_hierarchy for C.age (3)
mine classification as promising_customers (2)
in relevance to C.age, C.income, I.type, I.place_made, T.branch (1)
from customer C, item I, transaction T (1)
where I.item_ID = T.item_ID and C.cust_ID = T.cust_ID
      and C.income ≥ 40,000 and I.price ≥ 100 (1)
group by T.cust_ID (2)
having sum(I.price) ≥ 1,000 (1)
display as rules (5)
```

Why Data Mining Query Language?

- Automated vs. query-driven?
 - Finding all the patterns autonomously in a database?—unrealistic because the patterns could be too many but uninteresting
- Data mining should be an interactive process
 - User directs what to be mined
- Users must be provided with a set of **primitives** to be used to communicate with the data mining system
- Incorporating these primitives in a **data mining query language**
 - More flexible user interaction
 - Foundation for design of graphical user interface
 - Standardization of data mining industry and practice

1.8 Integration of Data Mining and Data Warehousing

- No coupling
 - Flat file processing, no utilization of any functions of a DB/DW system
 - Not recommended
- Loose coupling
 - Fetching data from DB/DW
 - Does not explore data structures and query optimization methods provided by DB/DW system
 - Difficult to achieve high scalability and good performance with large data sets

1.8 Integration of Data Mining and Data Warehousing

- Semi-tight
 - Efficient implementations of a few essential data mining primitives in a DB/DW system are provided, e.g., sorting, indexing, aggregation, histogram analysis, multiway join, precomputation of some stat functions
 - Enhanced DM performance
- Tight
 - DM is smoothly integrated into a DB/DW system, mining query is optimized based on mining query analysis, data structures, indexing, query processing methods of a DB/DW system
 - A uniform information processing environment, highly desirable

1.9 Major Issues in Data Mining

- Mining methodology and User interaction
 - Mining different kinds of knowledge
 - DM should cover a wide spectrum of data analysis and knowledge discovery tasks
 - Enable to use the database in different ways
 - Require the development of numerous data mining techniques
 - Interactive mining of knowledge at multiple levels of abstraction
 - Difficult to know exactly what will be discovered
 - Allow users to focus the search, refine data mining requests
 - Incorporation of background knowledge
 - Guide the discovery process
 - Allow discovered patterns to be expressed in concise terms and different levels of abstraction
 - Data mining query languages and ad hoc data mining
 - High-level query languages need to be developed
 - Should be integrated with a DB/DW query language

1.9 Major Issues in Data Mining

- Presentation and visualization of results
 - Knowledge should be easily understood and directly usable
 - High level languages, visual representations or other expressive forms
 - Require the DM system to adopt the above techniques
- Handling noisy or incomplete data
 - Require data cleaning methods and data analysis methods that can handle noise
- Pattern evaluation – the interestingness problem
 - How to develop techniques to access the interestingness of discovered patterns, especially with subjective measures bases on user beliefs or expectations

1.9 Major Issues in Data Mining

- Performance Issues
 - Efficiency and scalability
 - Huge amount of data
 - Running time must be predictable and acceptable
 - Parallel, distributed and incremental mining algorithms
 - Divide the data into partitions and processed in parallel
 - Incorporate database updates without having to mine the entire data again from scratch
- Diversity of Database Types
 - Other database that contain complex data objects, multimedia data, spatial data, etc.
 - Expect to have different DM systems for different kinds of data
 - Heterogeneous databases and global information systems
 - Web mining becomes a very challenging and fast-evolving field in data mining

Data Mining: Concepts and Techniques

— Chapter 2 —

Jiawei Han, Micheline Kamber, and Jian Pei

University of Illinois at Urbana-Champaign

Simon Fraser University

©2011 Han, Kamber, and Pei. All rights reserved.

Chapter 2: Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary



Types of Data Sets

- Record
 - Relational records
 - Data matrix, e.g., numerical matrix, crosstabs
 - Document data: text documents: term-frequency vector
 - Transaction data
- Graph and network
 - World Wide Web
 - Social or information networks
 - Molecular Structures
- Ordered
 - Video data: sequence of images
 - Temporal data: time-series
 - Sequential Data: transaction sequences
 - Genetic sequence data
- Spatial, image and multimedia:
 - Spatial data: maps
 - Image data:
 - Video data:

	team	coach	pla y	ball	score	game	wi n	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Important Characteristics of Structured Data

- Dimensionality
 - Curse of dimensionality
- Sparsity
 - Only presence counts
- Resolution
 - Patterns depend on the scale
- Distribution
 - Centrality and dispersion

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
 - sales database: customers, store items, sales
 - medical database: patients, treatments
 - university database: students, professors, courses
- Also called *samples*, *examples*, *instances*, *data points*, *objects*, *tuples*.
- Data objects are described by **attributes**.
- Database rows -> data objects; columns -> attributes.

Attributes

- **Attribute** (or **dimensions, features, variables**):
a data field, representing a characteristic or feature
of a data object.
 - *E.g., customer_ID, name, address*
- Types:
 - Nominal
 - Binary
 - Numeric: quantitative
 - Interval-scaled
 - Ratio-scaled

Attribute Types

- **Nominal:** categories, states, or “names of things”
 - $\text{Hair_color} = \{\text{auburn}, \text{black}, \text{blond}, \text{brown}, \text{grey}, \text{red}, \text{white}\}$
 - marital status, occupation, ID numbers, zip codes
- **Binary**
 - Nominal attribute with only 2 states (0 and 1)
 - Symmetric binary: both outcomes equally important
 - e.g., gender
 - Asymmetric binary: outcomes not equally important.
 - e.g., medical test (positive vs. negative)
 - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**
 - Values have a meaningful order (ranking) but magnitude between successive values is not known.
 - $\text{Size} = \{\text{small}, \text{medium}, \text{large}\}$, grades, army rankings

Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval**
 - Measured on a scale of **equal-sized units**
 - Values have order
 - E.g., *temperature in C° or F°, calendar dates*
 - No true zero-point
- **Ratio**
 - Inherent **zero-point**
 - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K° is twice as high as 5 K°).
 - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Discrete vs. Continuous Attributes

■ Discrete Attribute

- Has only a finite or countably infinite set of values
 - E.g., zip codes, profession, or the set of words in a collection of documents
- Sometimes, represented as integer variables
- Note: Binary attributes are a special case of discrete attributes

■ Continuous Attribute

- Has real numbers as attribute values
 - E.g., temperature, height, or weight
- Practically, real values can only be measured and represented using a finite number of digits
- Continuous attributes are typically represented as floating-point variables

Chapter 2: Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data 
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

Basic Statistical Descriptions of Data

- 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):

Note: n is sample size and N is population size.

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

- 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:

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

age	frequency
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

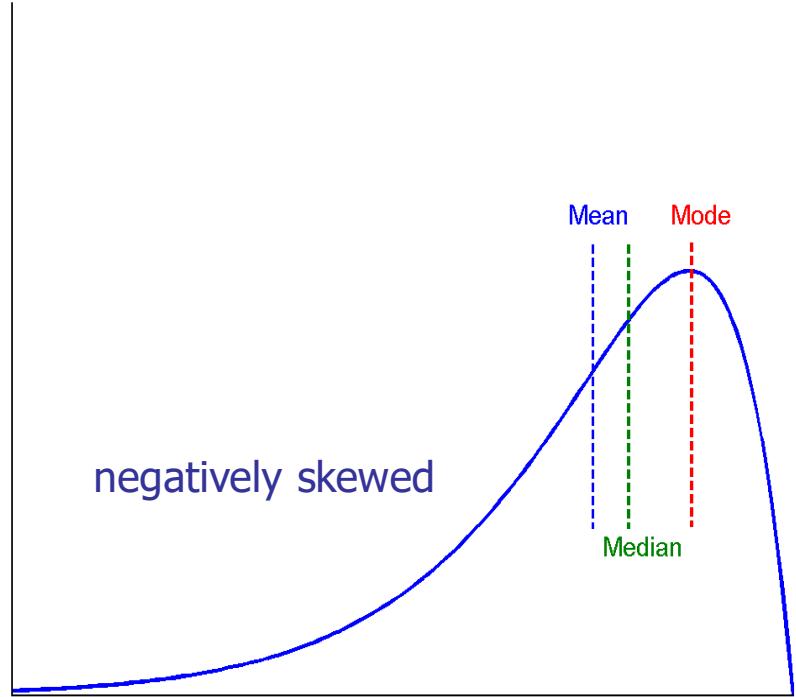
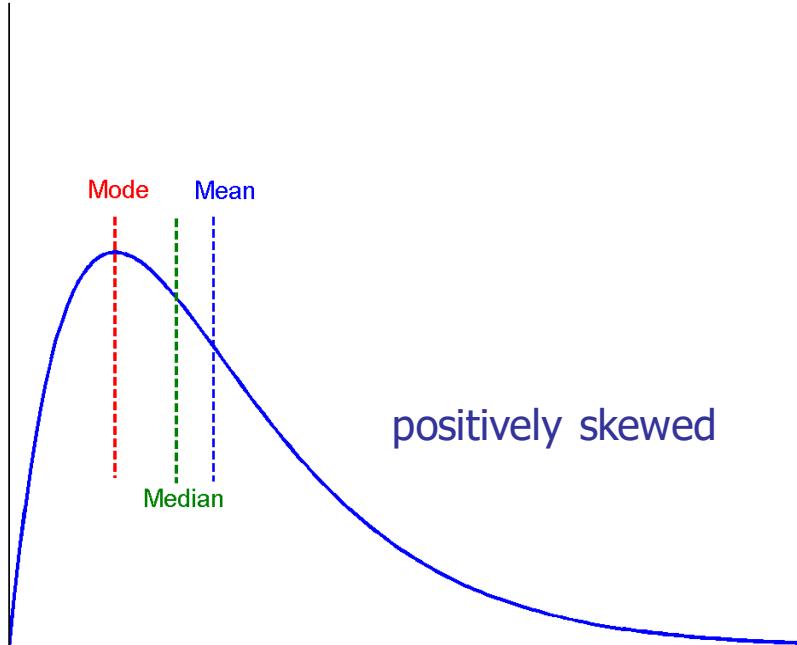
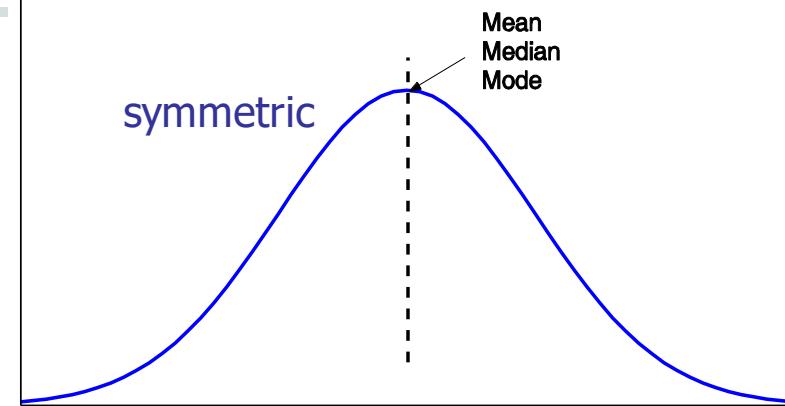
$$median = L_1 + \left(\frac{n/2 - (\sum freq)l}{freq_{median}} \right) width$$

- 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



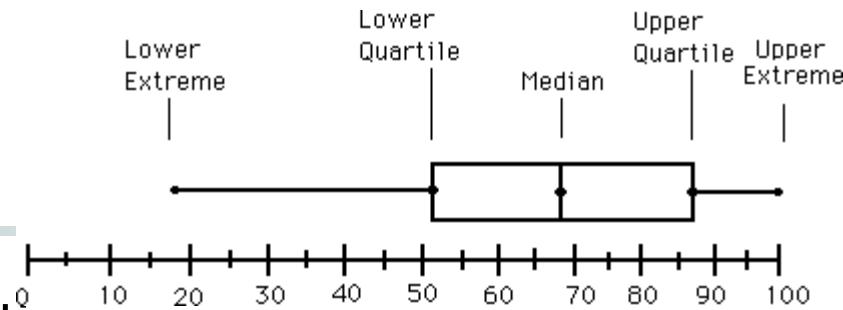
Measuring the Dispersion of Data

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

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

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

Boxplot Analysis

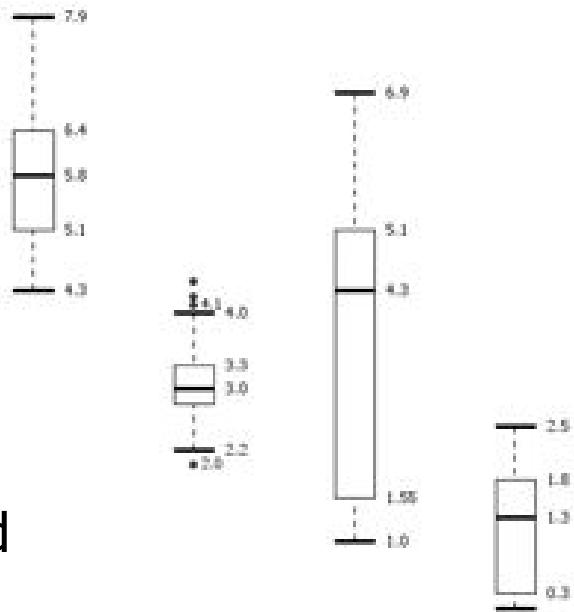


- **Five-number summary** of a distribution

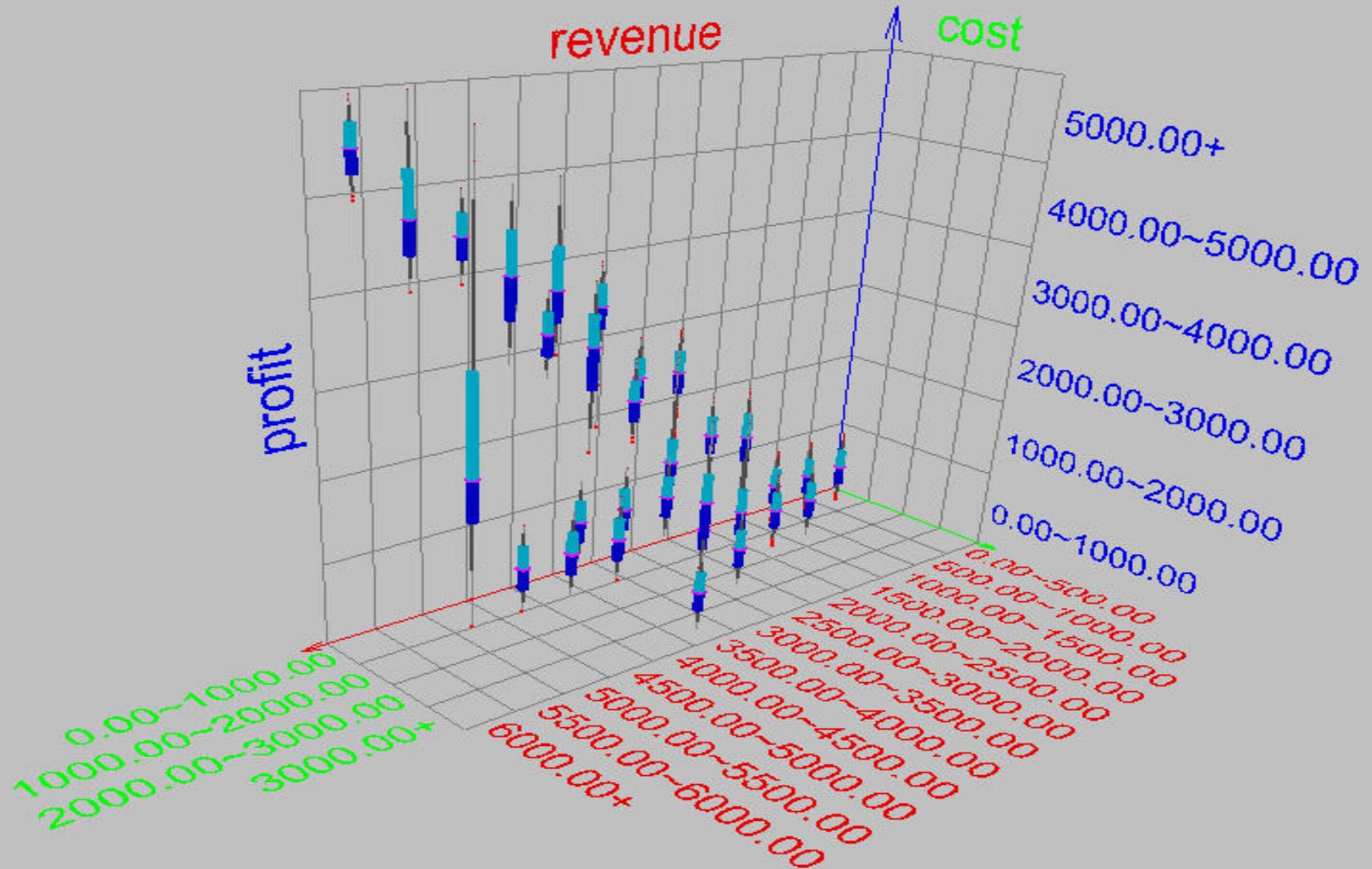
- Minimum, Q1, Median, Q3, Maximum

- **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 extended to Minimum and Maximum
 - Outliers: points beyond a specified outlier threshold, plotted individually

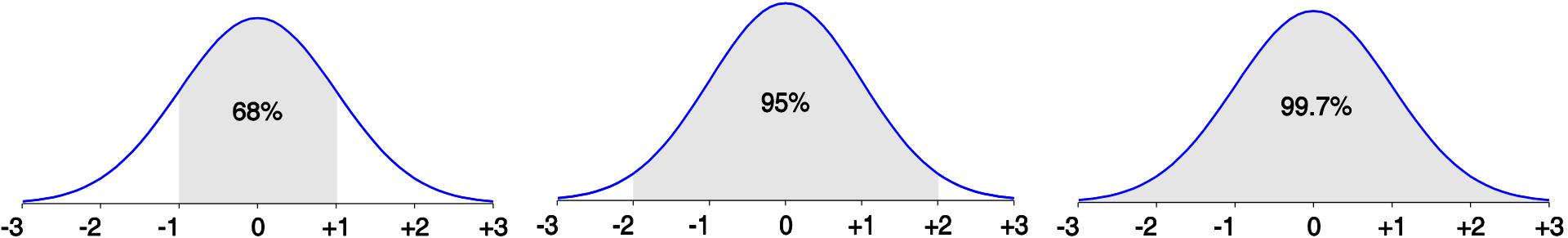


Visualization of Data Dispersion: 3-D Boxplots



Properties of Normal Distribution Curve

- The normal (distribution) curve
 - From $\mu-\sigma$ to $\mu+\sigma$: contains about 68% of the measurements (μ : mean, σ : 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

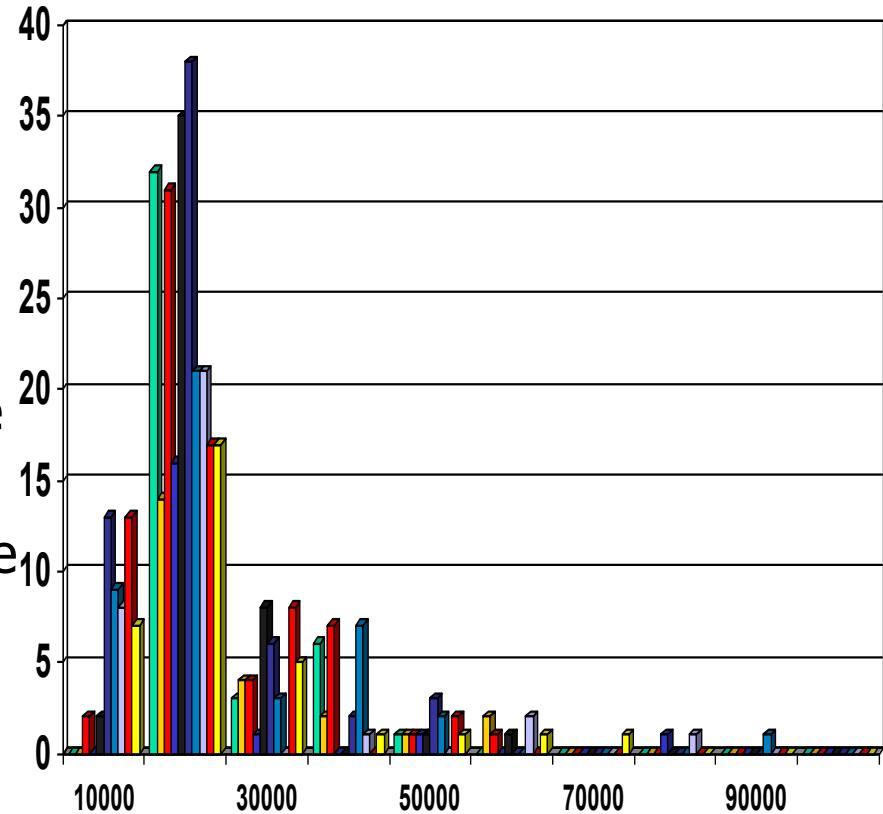


Graphic Displays of Basic Statistical Descriptions

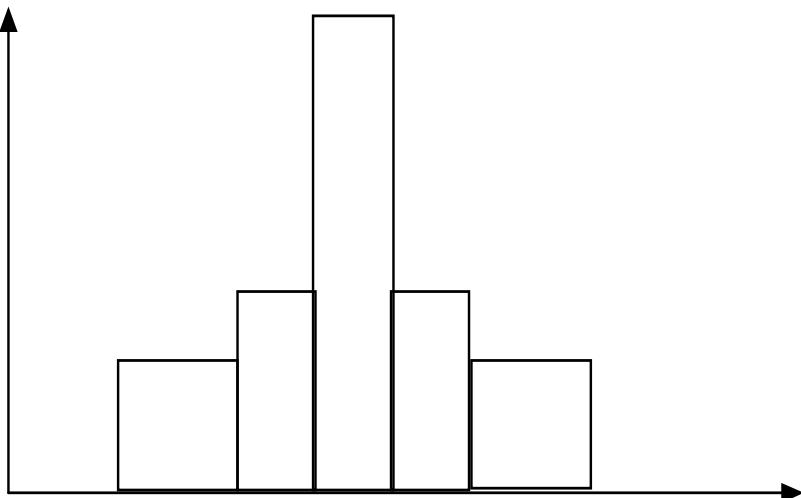
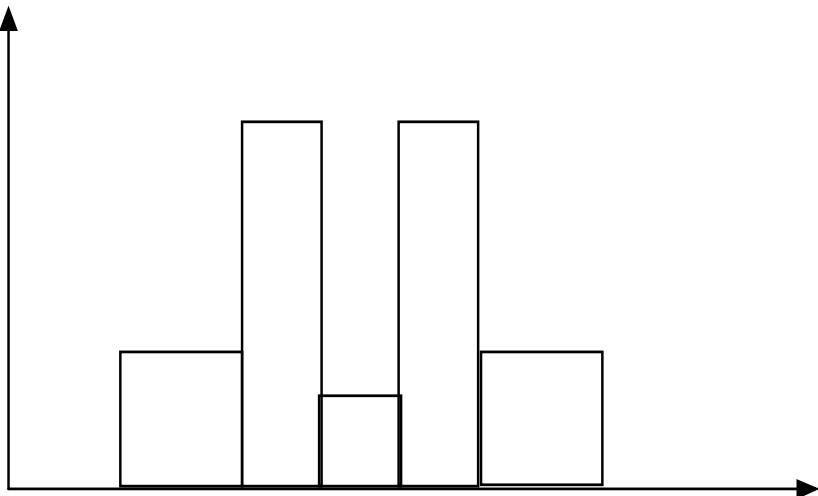
- **Boxplot:** graphic display of five-number summary
- **Histogram:** x-axis are values, y-axis repres. frequencies
- **Quantile plot:** each value x_i is paired with f_i indicating that approximately $100 f_i\%$ of data are $\leq x_i$
- **Quantile-quantile (q-q) plot:** graphs the quantiles of one 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

Histogram Analysis

- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- Differs from a bar chart in that it is the *area* of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent



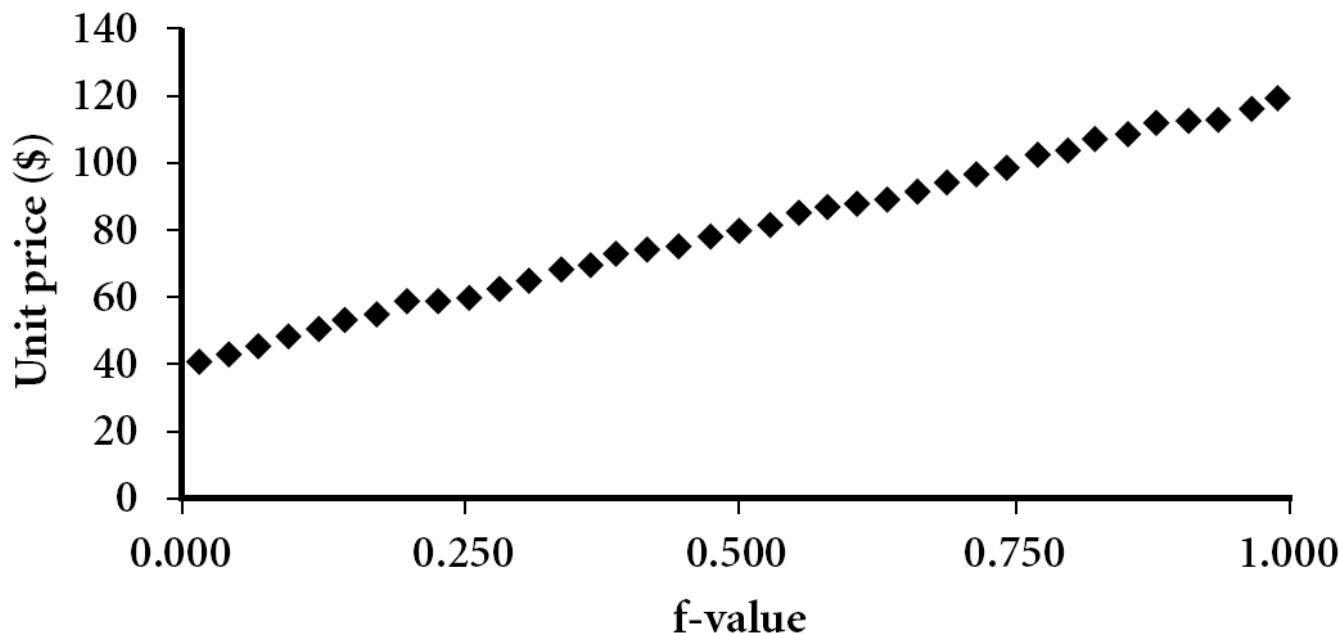
Histograms Often Tell More than Boxplots



- The two histograms shown in the left may have the same boxplot representation
 - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

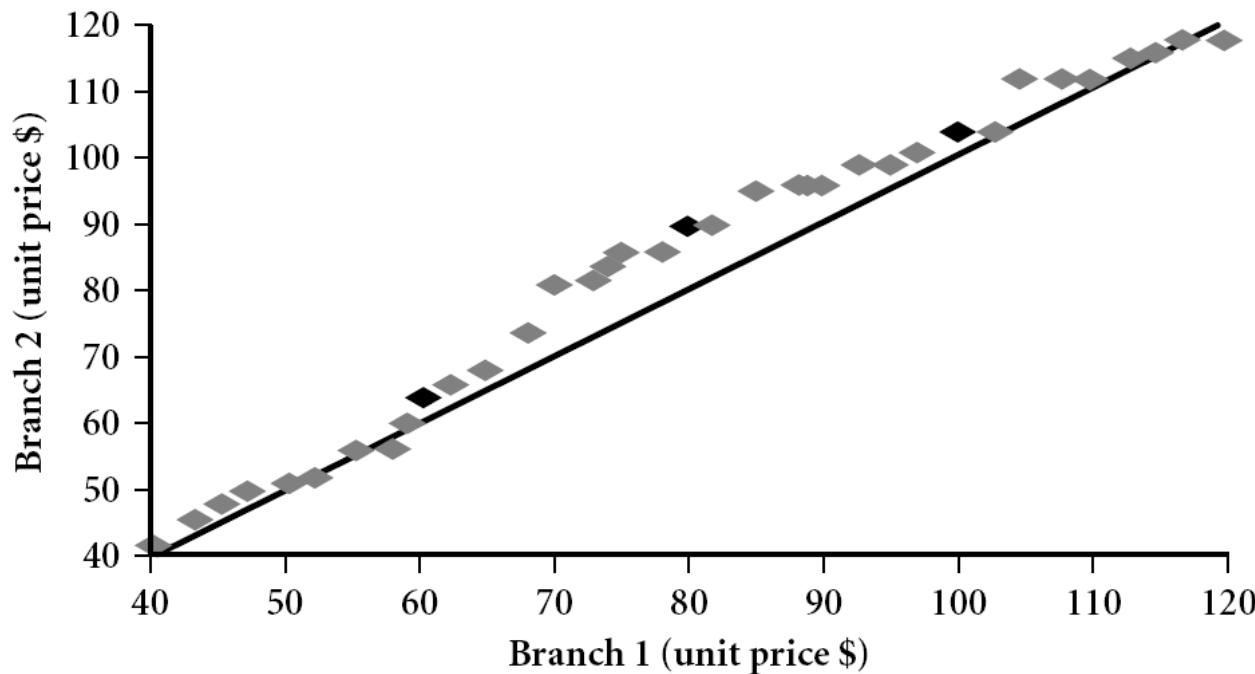
Quantile Plot

- 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



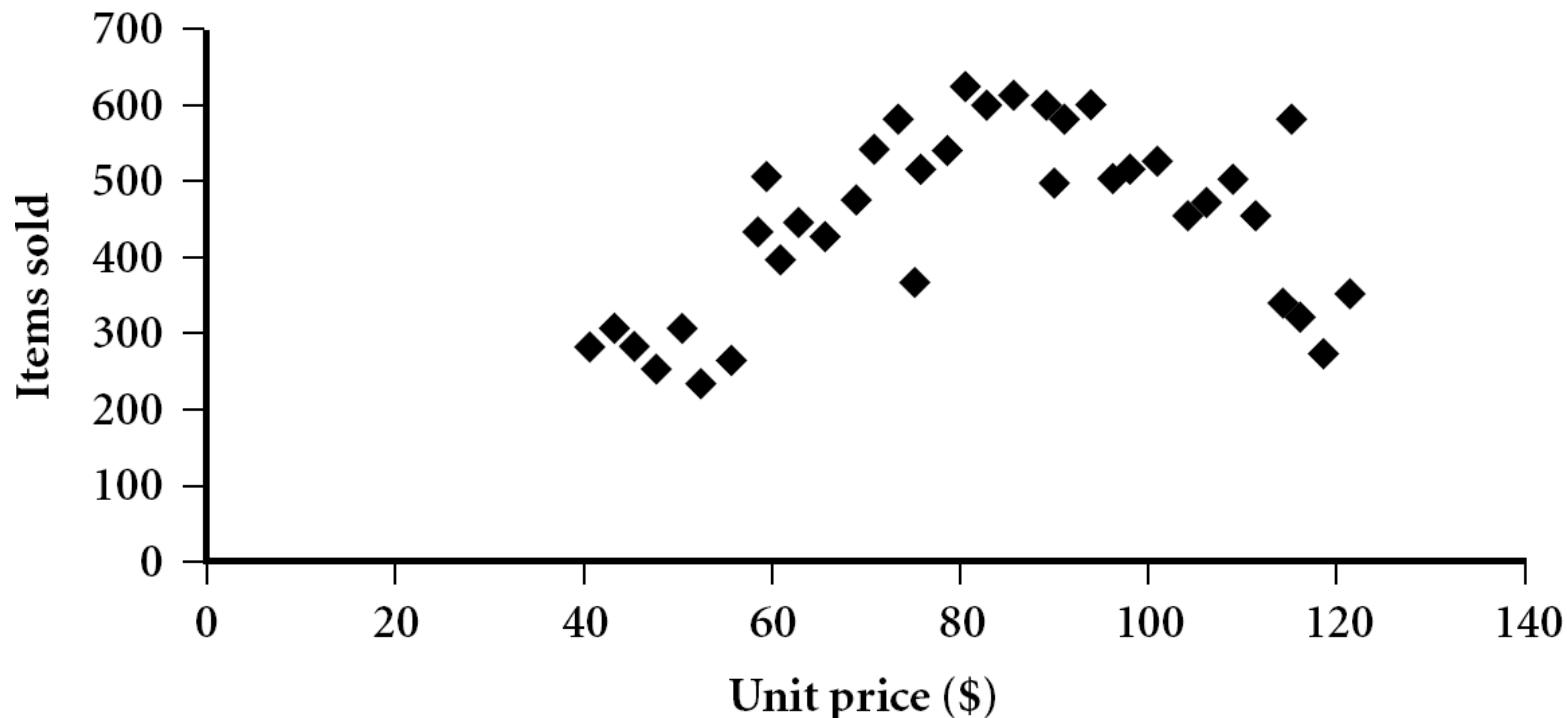
Quantile-Quantile (Q-Q) Plot

- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- View: Is there is a shift in going from one distribution to another?
- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.

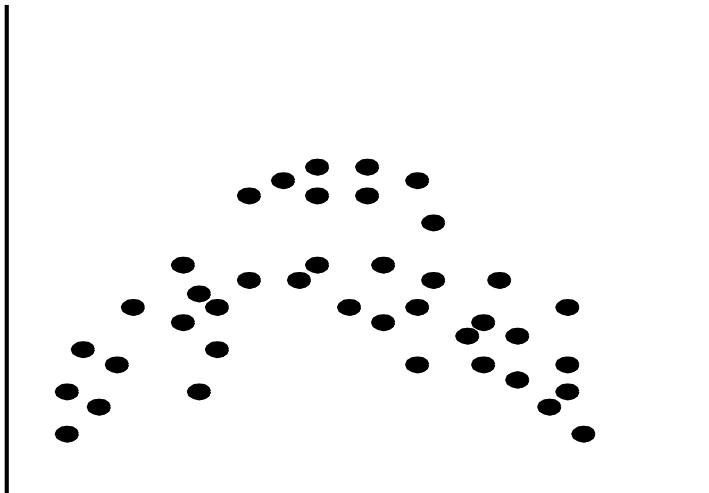
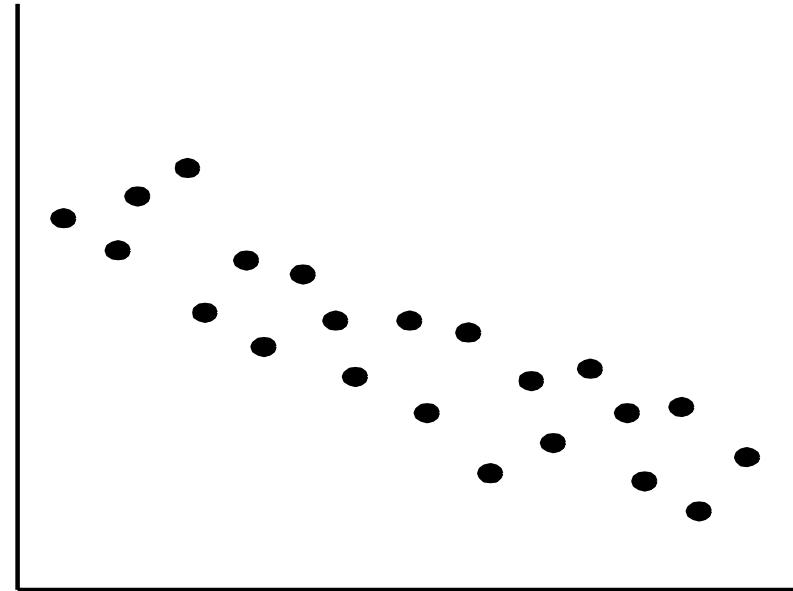
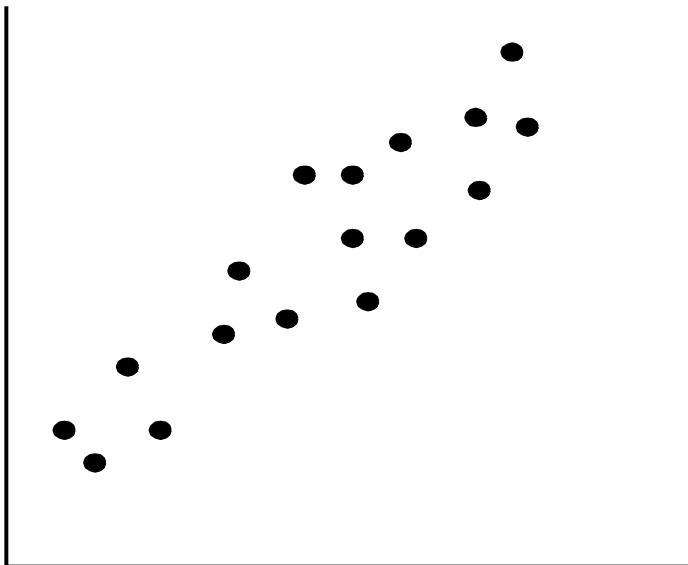


Scatter plot

- 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

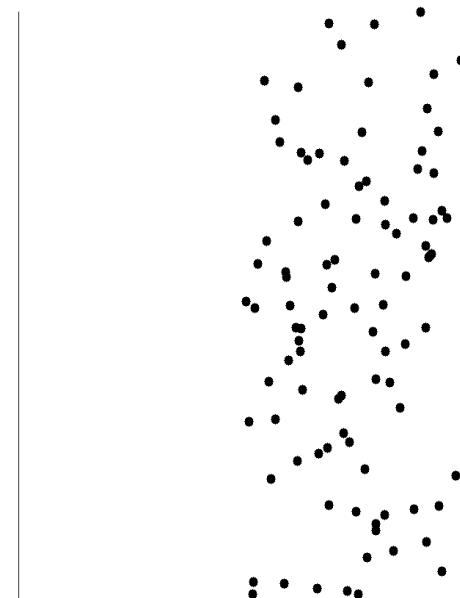
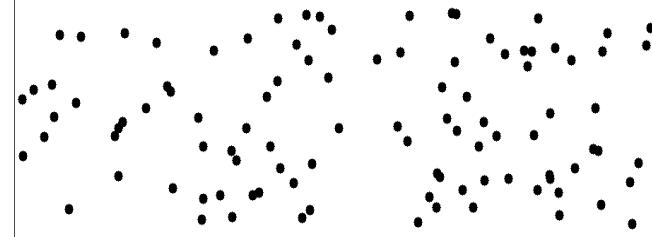
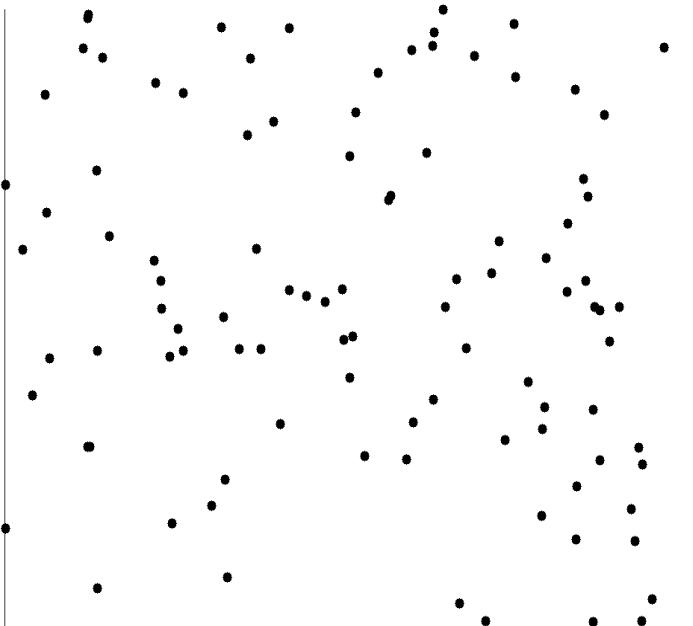


Positively and Negatively Correlated Data



- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data



Chapter 2: Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

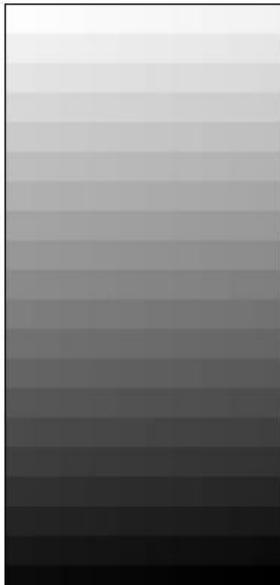


Data Visualization

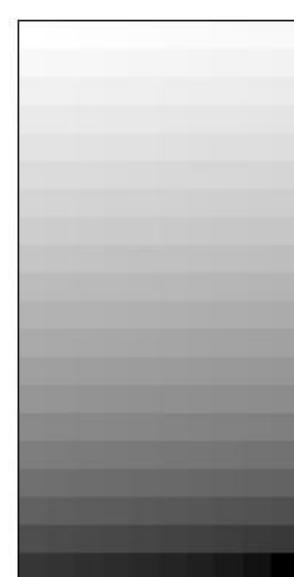
- Why data visualization?
 - Gain insight into an information space by mapping data onto graphical primitives
 - Provide qualitative overview of large data sets
 - Search for patterns, trends, structure, irregularities, relationships among data
 - Help find interesting regions and suitable parameters for further quantitative analysis
 - Provide a visual proof of computer representations derived
- Categorization of visualization methods:
 - Pixel-oriented visualization techniques
 - Geometric projection visualization techniques
 - Icon-based visualization techniques
 - Hierarchical visualization techniques
 - Visualizing complex data and relations

Pixel-Oriented Visualization Techniques

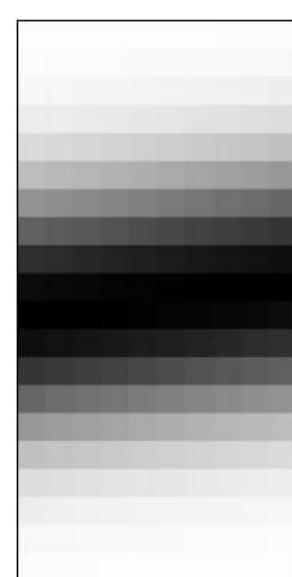
- For a data set of m dimensions, create m windows on the screen, one for each dimension
- The m dimension values of a record are mapped to m pixels at the corresponding positions in the windows
- The colors of the pixels reflect the corresponding values



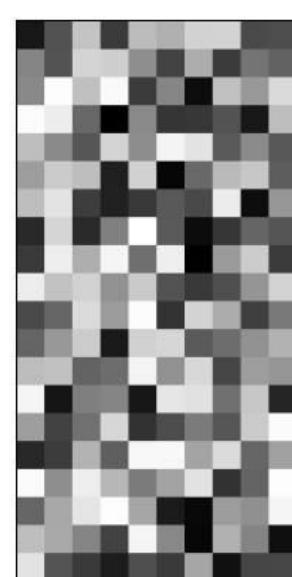
(a) Income



(b) Credit Limit



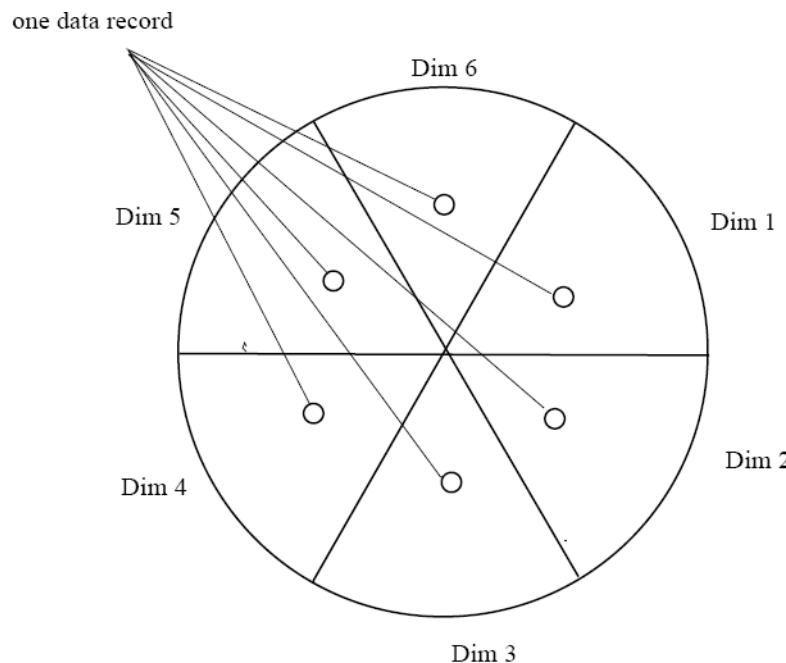
(c) transaction volume



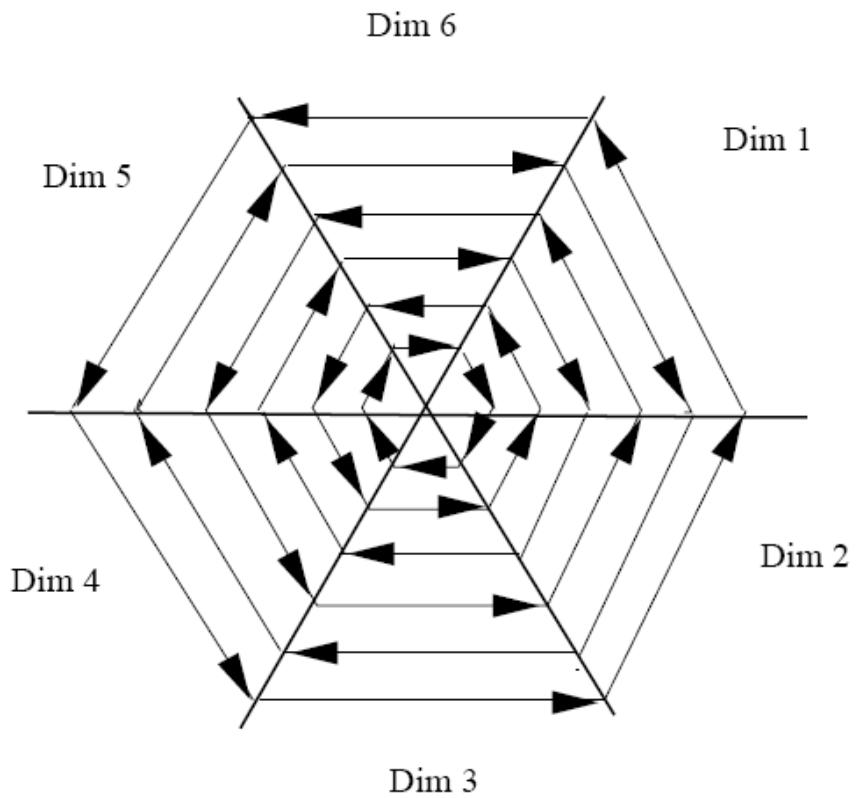
(d) age

Laying Out Pixels in Circle Segments

- To save space and show the connections among multiple dimensions, space filling is often done in a circle segment



(a) Representing a data record
in circle segment



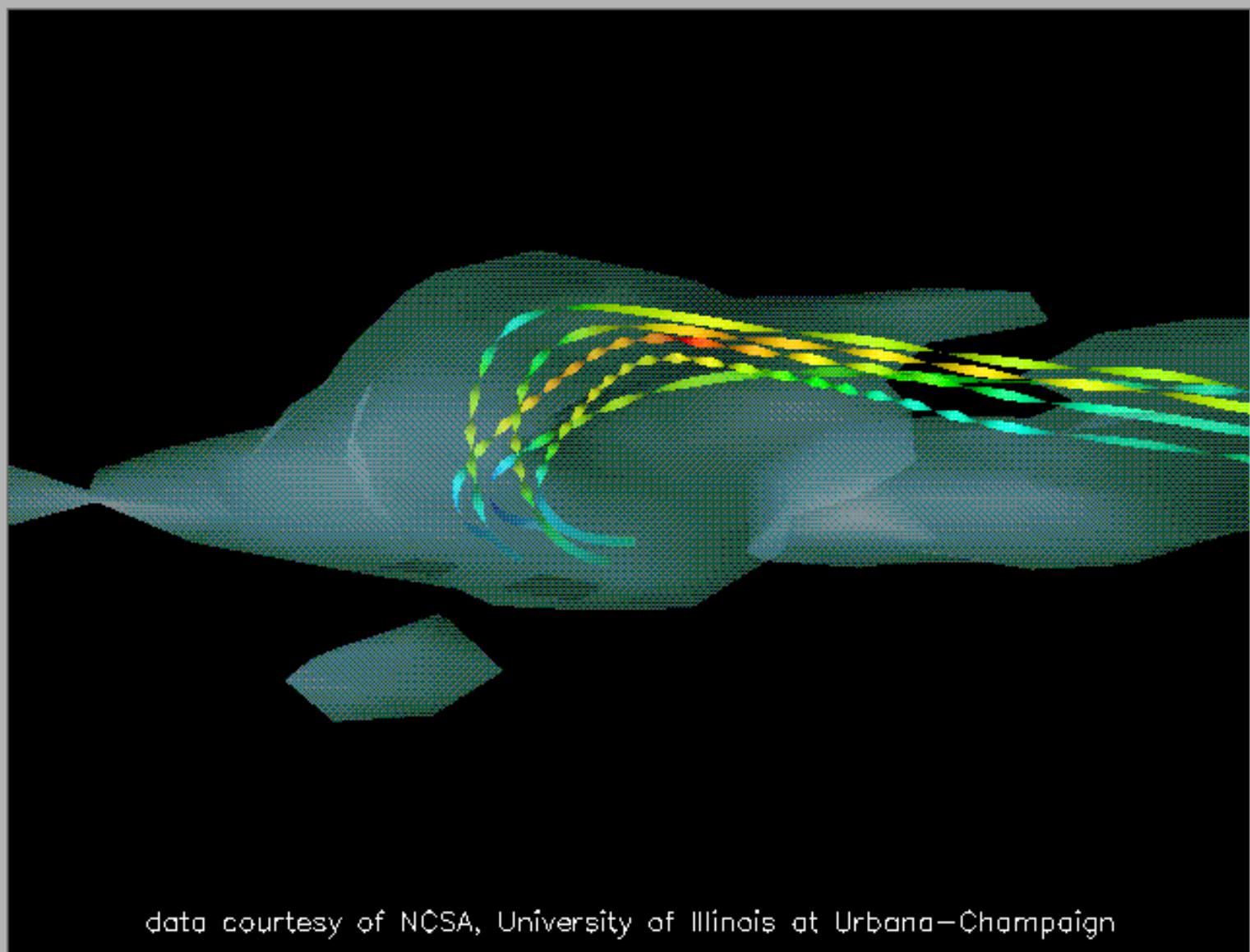
(b) Laying out pixels in circle segment

Geometric Projection Visualization Techniques

- Visualization of geometric transformations and projections of the data
- Methods
 - Direct visualization
 - Scatterplot and scatterplot matrices
 - Landscapes
 - Projection pursuit technique: Help users find meaningful projections of multidimensional data
 - Prosection views
 - Hyperslice
 - Parallel coordinates

Direct Data Visualization

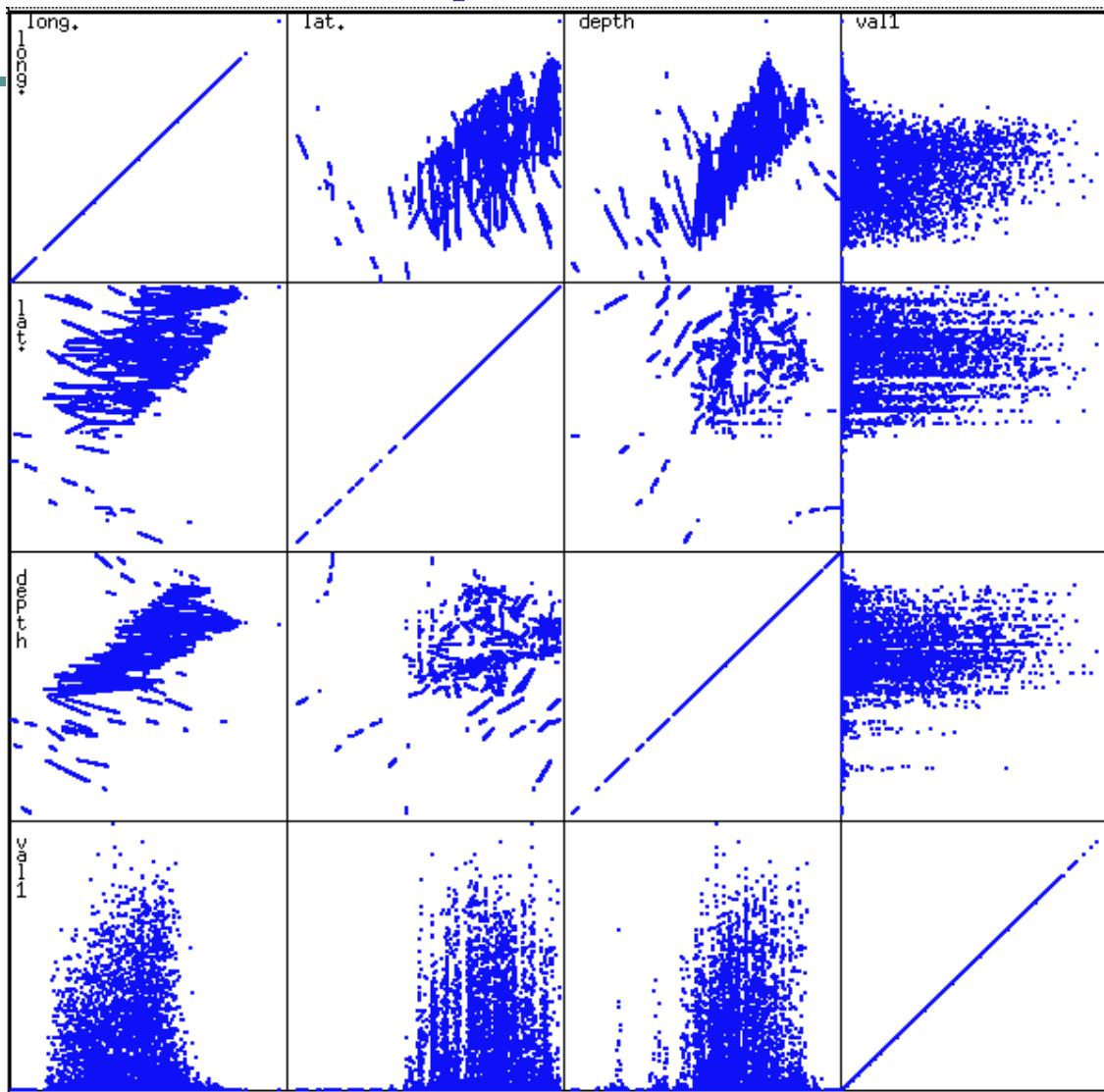
Ribbons with Twists Based on Vorticity



data courtesy of NCSA, University of Illinois at Urbana-Champaign

Scatterplot Matrices

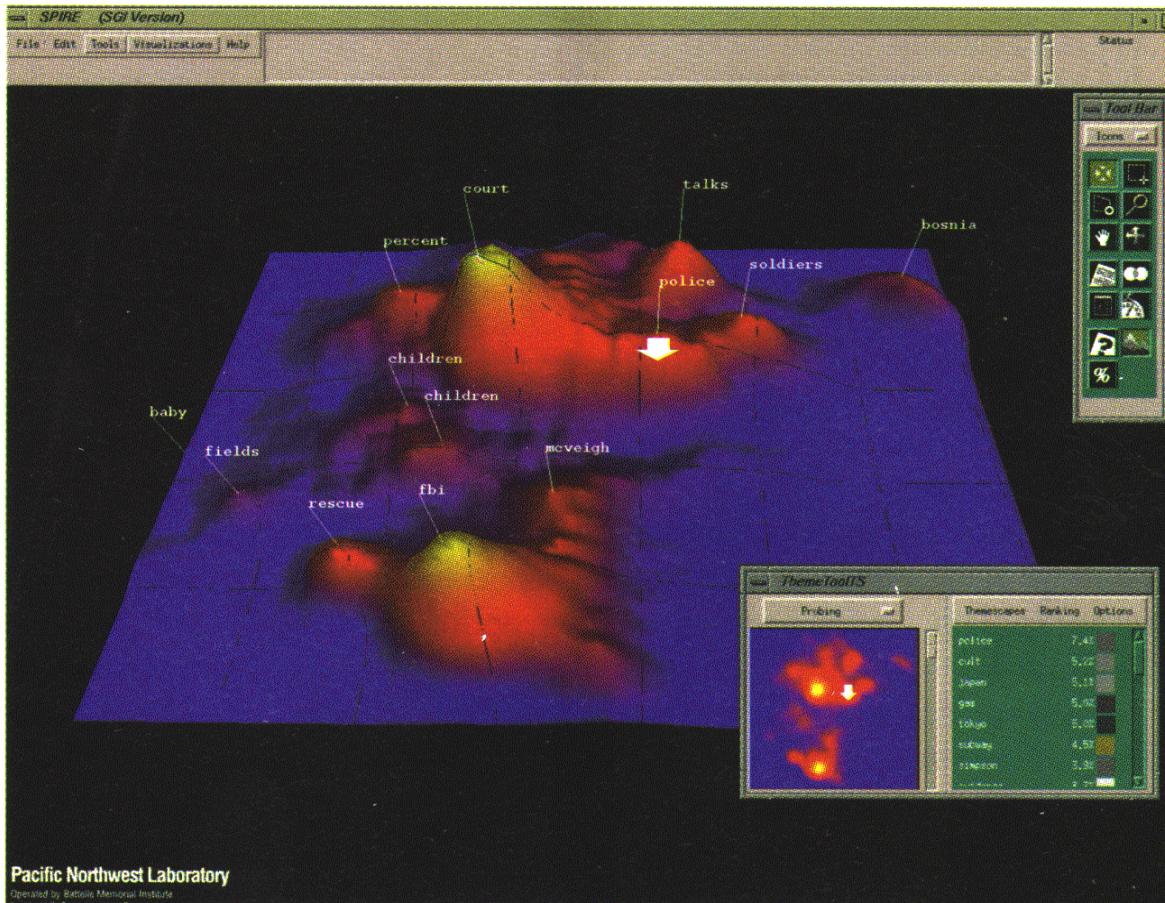
Used by permission of M. Ward, Worcester Polytechnic Institute



Matrix of scatterplots (x-y-diagrams) of the k-dim. data [total of $(k^2/2 - k)$ scatterplots]

Landscapes

Used by permission of B. Wright, Visible Decisions Inc.

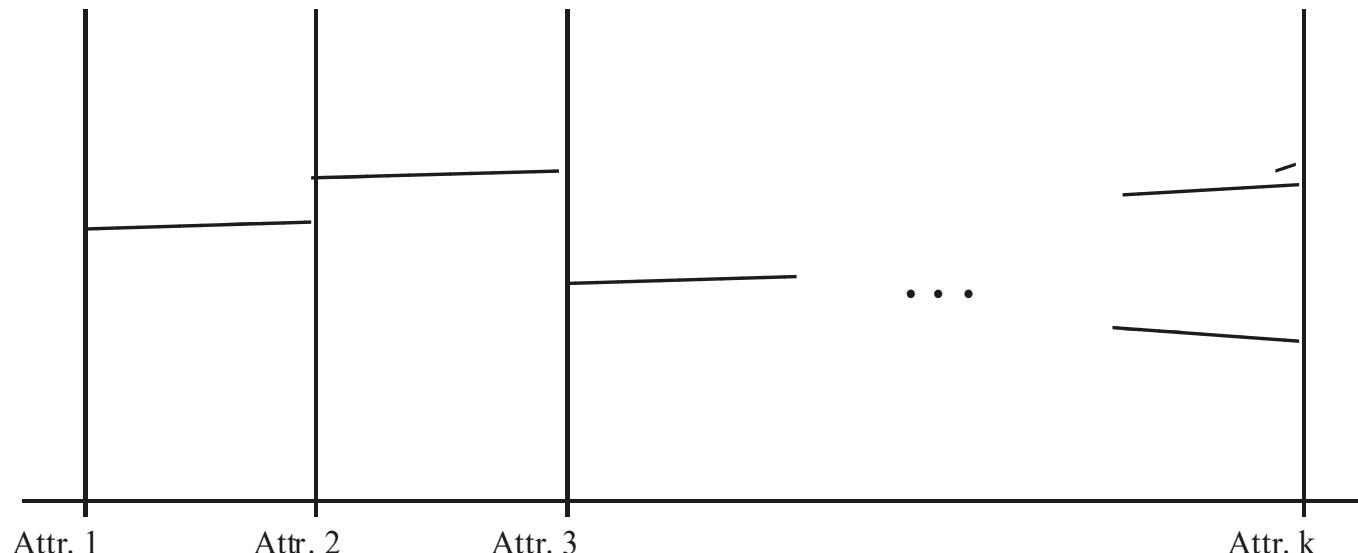


news articles
visualized as
a landscape

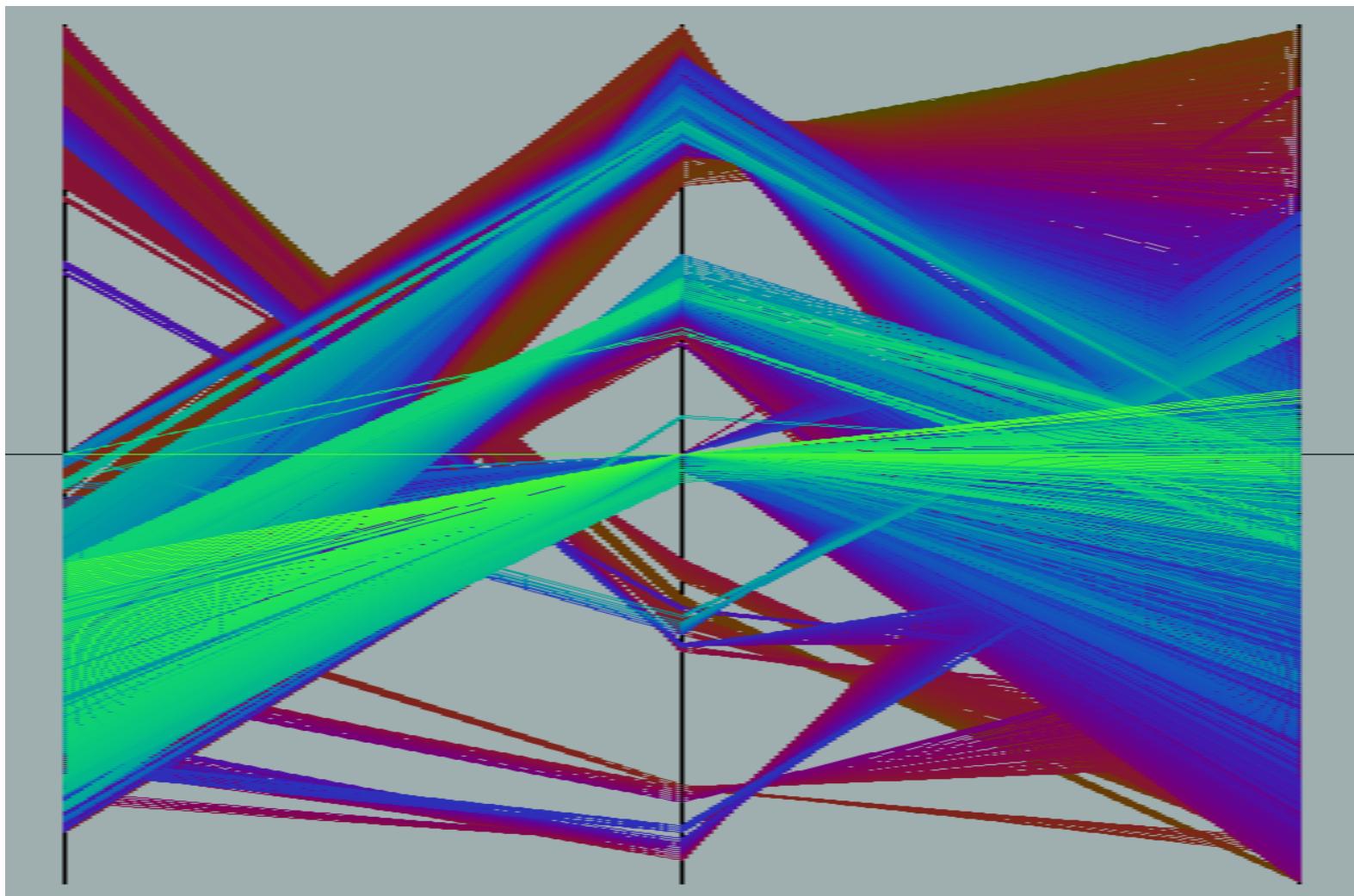
- Visualization of the data as perspective landscape
- The data needs to be transformed into a (possibly artificial) 2D spatial representation which preserves the characteristics of the data

Parallel Coordinates

- n equidistant axes which are parallel to one of the screen axes and correspond to the attributes
- The axes are scaled to the [minimum, maximum]: range of the corresponding attribute
- Every data item corresponds to a polygonal line which intersects each of the axes at the point which corresponds to the value for the attribute



Parallel Coordinates of a Data Set



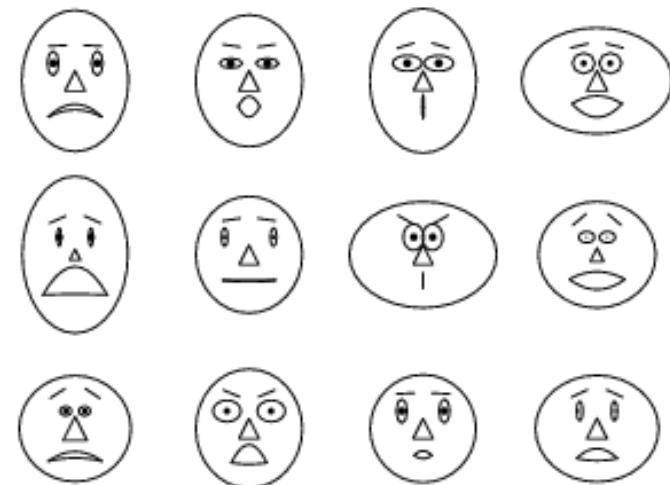
Icon-Based Visualization Techniques

- Visualization of the data values as features of icons
- Typical visualization methods
 - Chernoff Faces
 - Stick Figures
- General techniques
 - Shape coding: Use shape to represent certain information encoding
 - Color icons: Use color icons to encode more information
 - Tile bars: Use small icons to represent the relevant feature vectors in document retrieval

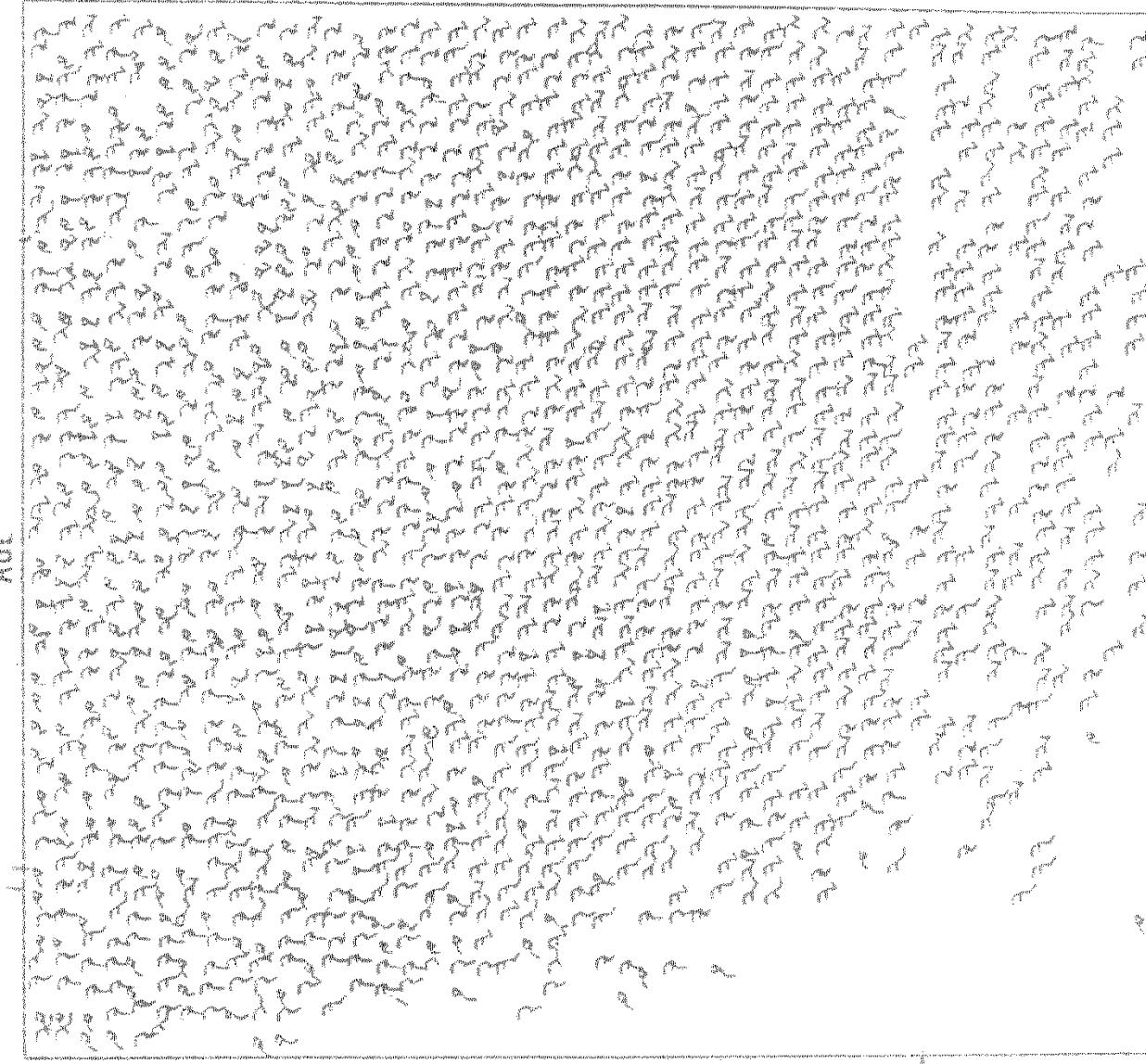
Chernoff Faces

- A way to display variables on a two-dimensional surface, e.g., let x be eyebrow slant, y be eye size, z be nose length, etc.
- The figure shows faces produced using 10 characteristics--head eccentricity, eye size, eye spacing, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, mouth size, and mouth opening): Each assigned one of 10 possible values, generated using *Mathematica* (S. Dickson)

- REFERENCE: Gonick, L. and Smith, W. *The Cartoon Guide to Statistics*, New York: Harper Perennial, p. 212, 1993
- Weisstein, Eric W. "Chernoff Face." From *MathWorld--A Wolfram Web Resource*.
mathworld.wolfram.com/ChernoffFace.html



Stick Figure



INCOME

Two attributes mapped to axes, remaining attributes mapped to angle or length of limbs". Look at texture pattern

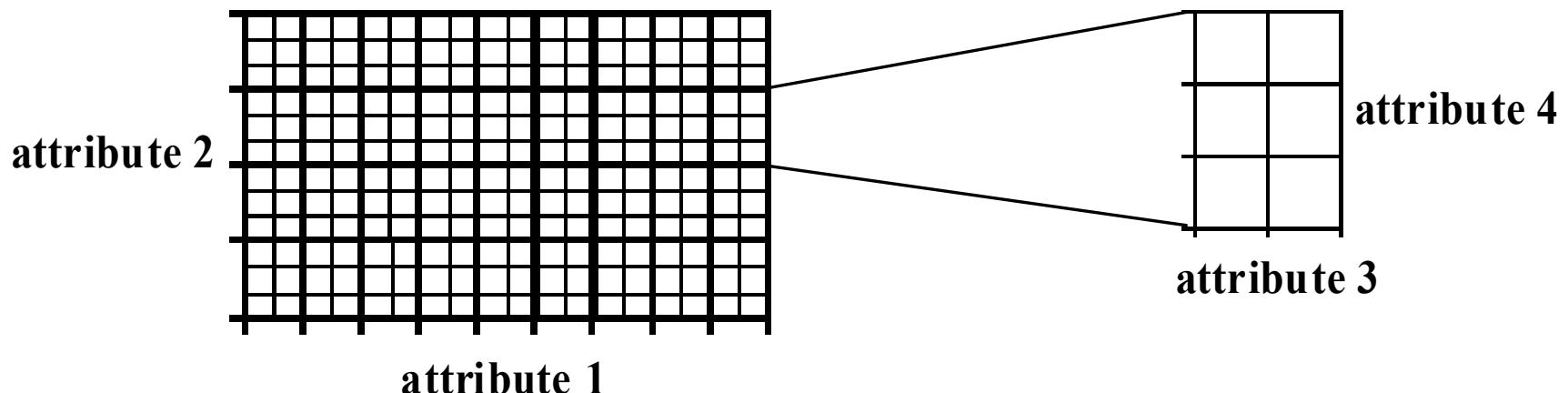
A census data figure showing age, income, gender, education, etc.

A 5-piece stick figure (1 body and 4 limbs w. different angle/length)

Hierarchical Visualization Techniques

- Visualization of the data using a hierarchical partitioning into subspaces
- Methods
 - Dimensional Stacking
 - Worlds-within-Worlds
 - Tree-Map
 - Cone Trees
 - InfoCube

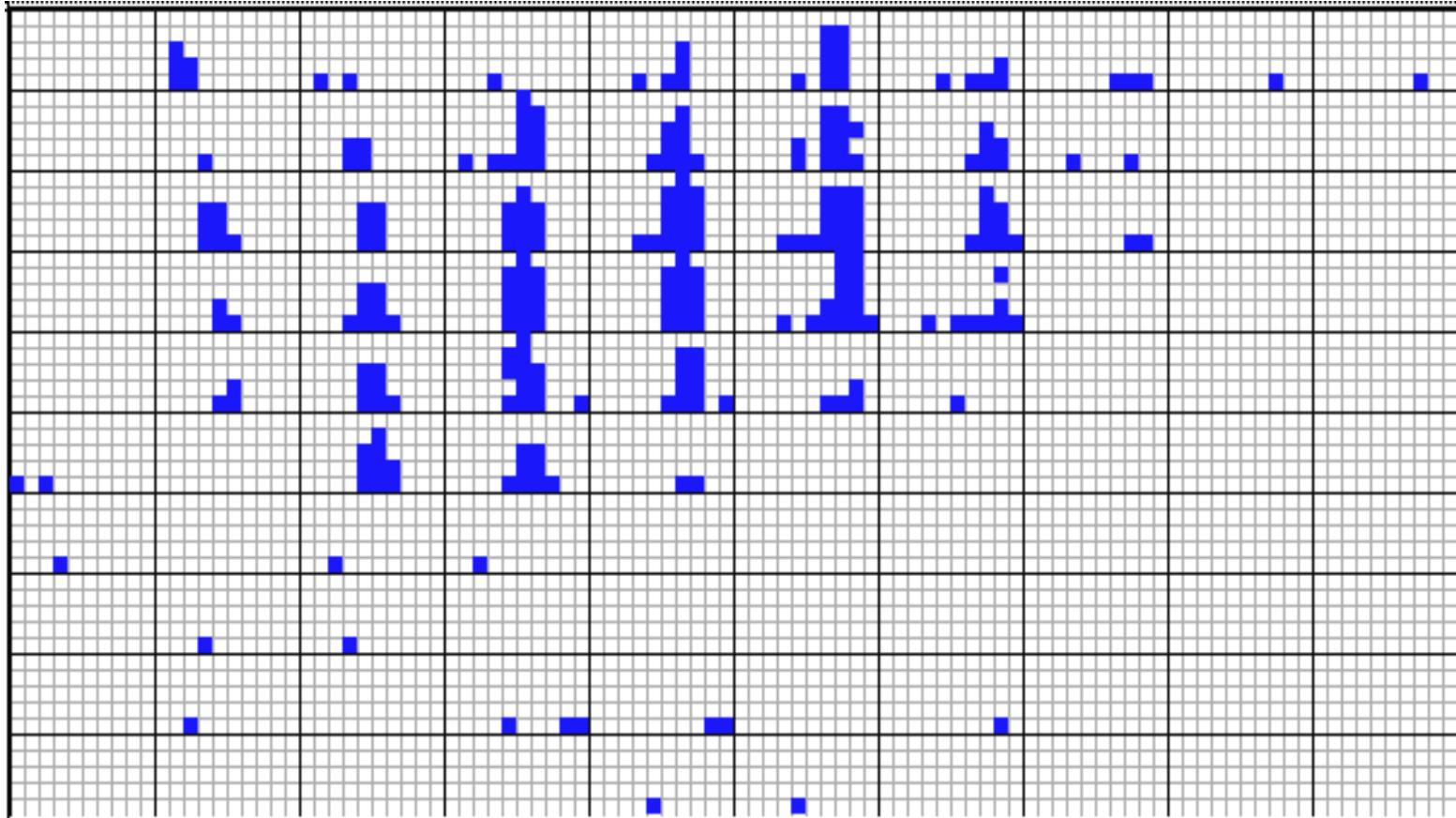
Dimensional Stacking



- Partitioning of the n-dimensional attribute space in 2-D subspaces, which are 'stacked' into each other
- Partitioning of the attribute value ranges into classes. The important attributes should be used on the outer levels.
- Adequate for data with ordinal attributes of low cardinality
- But, difficult to display more than nine dimensions
- Important to map dimensions appropriately

Dimensional Stacking

Used by permission of M. Ward, Worcester Polytechnic Institute

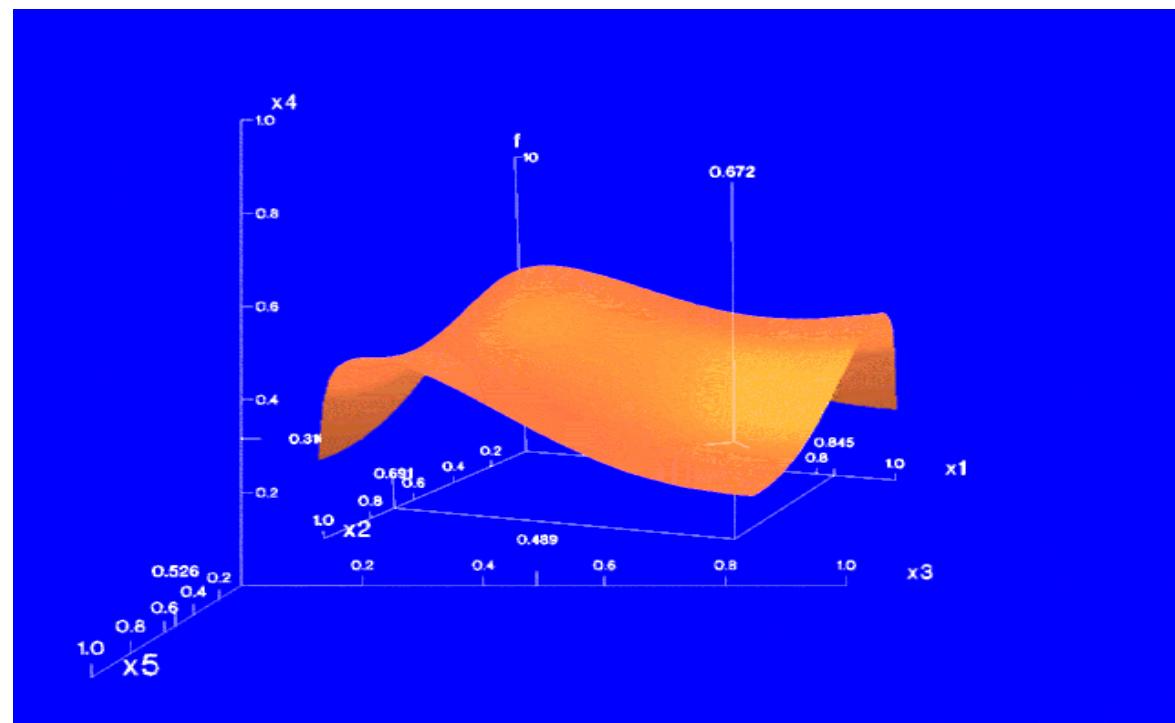


Visualization of oil mining data with longitude and latitude mapped to the outer x-, y-axes and ore grade and depth mapped to the inner x-, y-axes

Worlds-within-Worlds

- Assign the function and two most important parameters to innermost world
- Fix all other parameters at constant values - draw other (1 or 2 or 3 dimensional worlds choosing these as the axes)
- Software that uses this paradigm

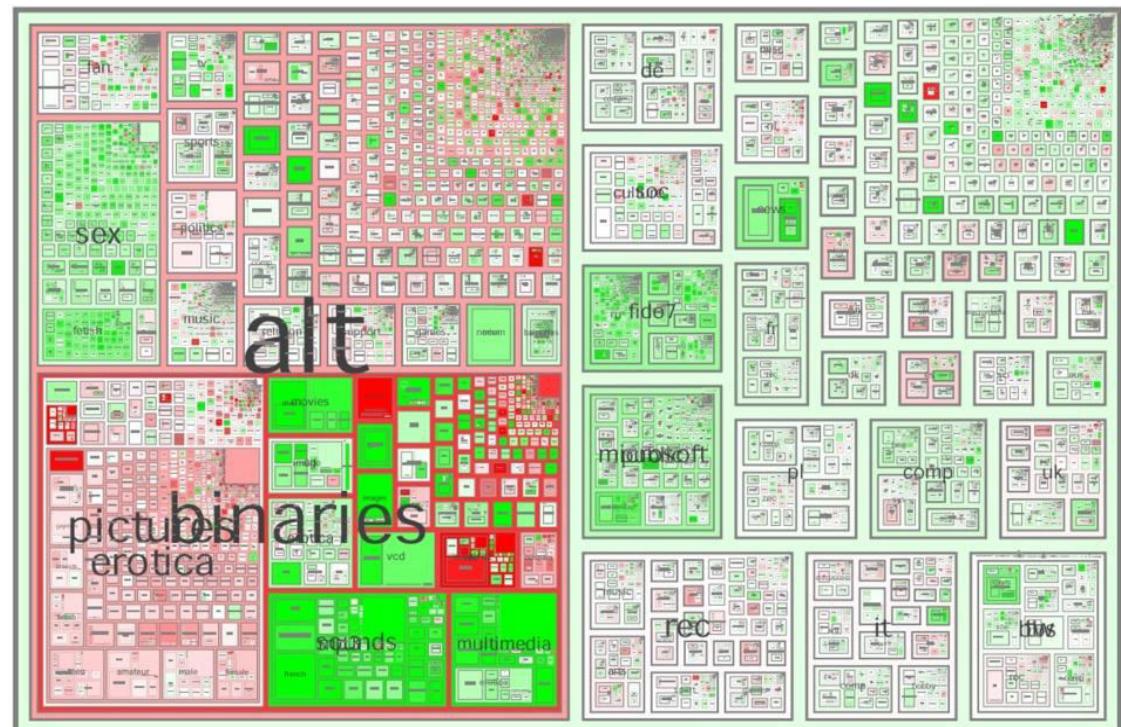
- N-vision: Dynamic interaction through data glove and stereo displays, including rotation, scaling (inner) and translation (inner/outer)
- Auto Visual: Static interaction by means of queries



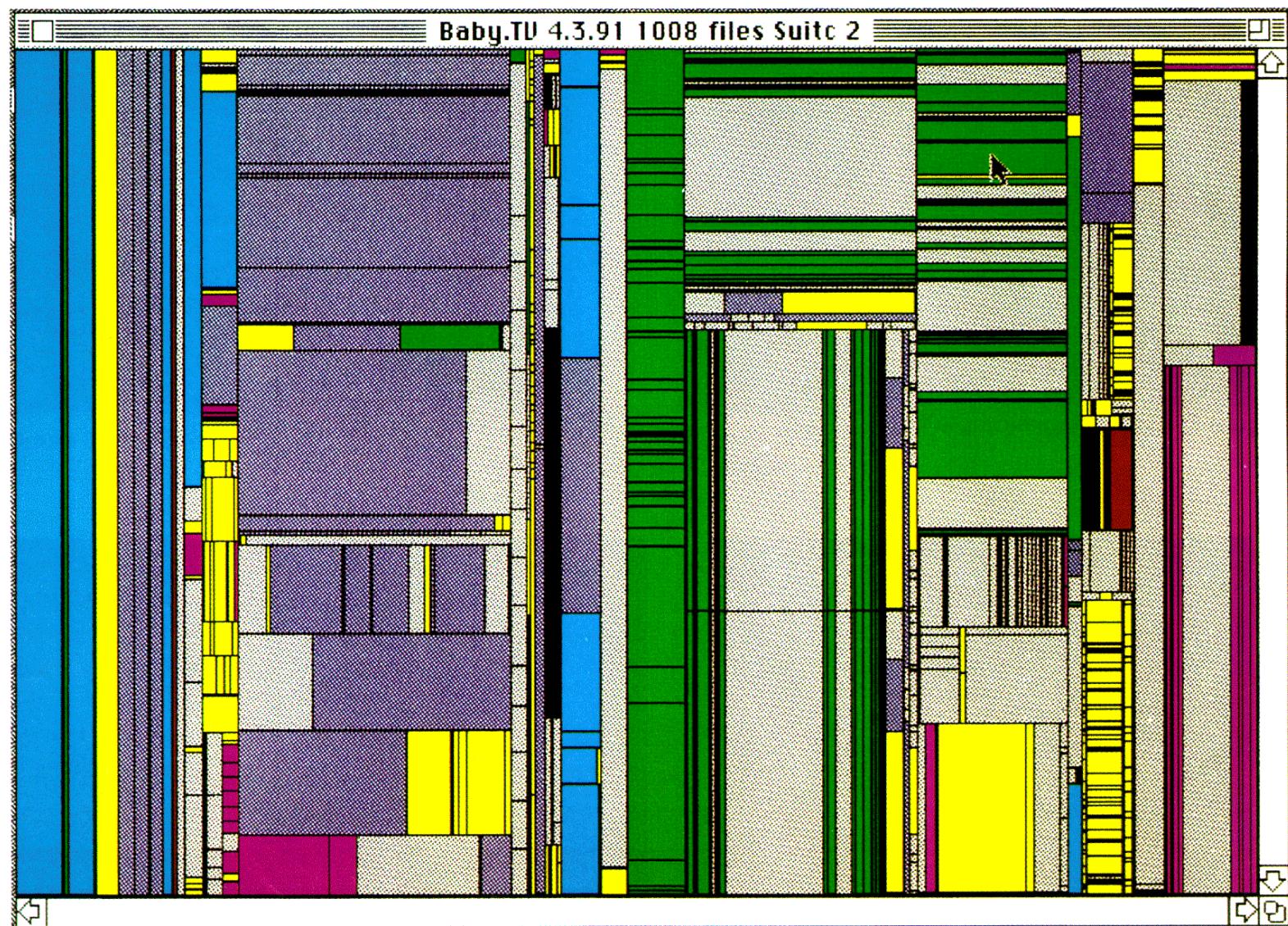
Tree-Map

- Screen-filling method which uses a hierarchical partitioning of the screen into regions depending on the attribute values
 - The x- and y-dimension of the screen are partitioned alternately according to the attribute values (classes)

MSR Netscan Image

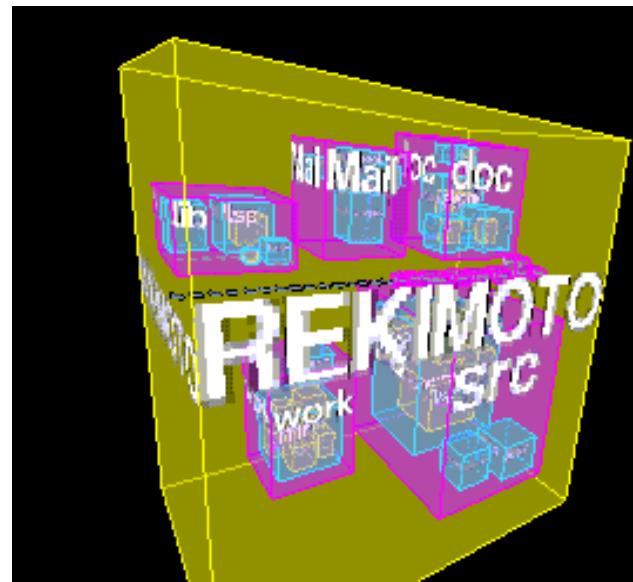


Tree-Map of a File System (Schneiderman)



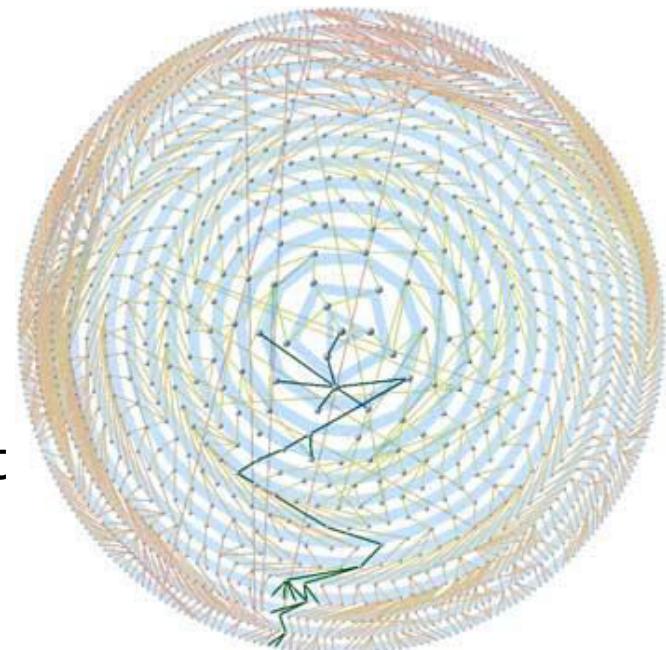
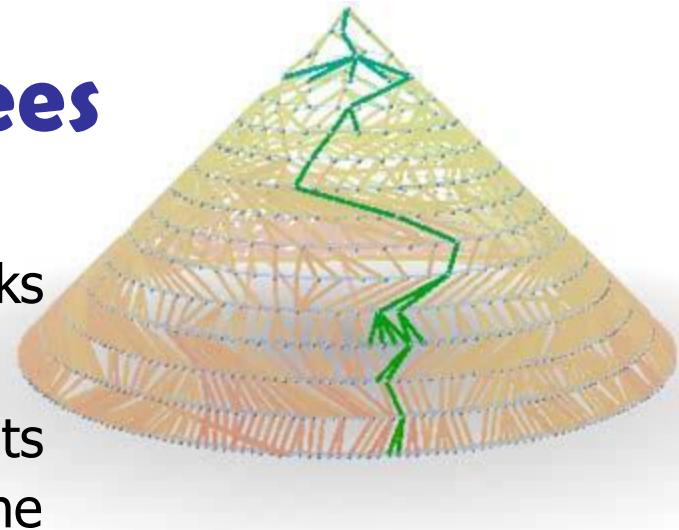
InfoCube

- A 3-D visualization technique where hierarchical information is displayed as nested semi-transparent cubes
- The outermost cubes correspond to the top level data, while the subnodes or the lower level data are represented as smaller cubes inside the outermost cubes, and so on



Three-D Cone Trees

- *3D cone tree* visualization technique works well for up to a thousand nodes or so
- First build a *2D circle tree* that arranges its nodes in concentric circles centered on the root node
- Cannot avoid overlaps when projected to 2D
- G. Robertson, J. Mackinlay, S. Card. "Cone Trees: Animated 3D Visualizations of Hierarchical Information", *ACM SIGCHI'91*
- Graph from Nadeau Software Consulting website: Visualize a social network data set that models the way an infection spreads from one person to the next



Visualizing Complex Data and Relations

- Visualizing non-numerical data: text and social networks
- Tag cloud: visualizing user-generated tags
 - The importance of tag is represented by font size/color
- Besides text data, there are also methods to visualize relationships, such as visualizing social networks



Newsmap: Google News Stories in 2005

Chapter 2: Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary



Similarity and Dissimilarity

- **Similarity**
 - Numerical measure of how alike two data objects are
 - Value is higher when objects are more alike
 - Often falls in the range [0,1]
- **Dissimilarity** (e.g., distance)
 - Numerical measure of how different two data objects are
 - Lower when objects are more alike
 - Minimum dissimilarity is often 0
 - Upper limit varies
- **Proximity** refers to a similarity or dissimilarity

Data Matrix and Dissimilarity Matrix

■ Data matrix

- n data points with p dimensions
- Two modes

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

■ Dissimilarity matrix

- n data points, but registers only the distance
- A triangular matrix
- Single mode

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \ddots & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Proximity Measure for Nominal Attributes

- Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)
- Method 1: Simple matching
 - m : # of matches, p : total # of variables

$$d(i, j) = \frac{p - m}{p}$$

- Method 2: Use a large number of binary attributes
 - creating a new binary attribute for each of the M nominal states

Proximity Measure for Binary Attributes

- A contingency table for binary data

		Object <i>j</i>		
		1	0	sum
Object <i>i</i>	1	<i>q</i>	<i>r</i>	<i>q</i> + <i>r</i>
	0	<i>s</i>	<i>t</i>	<i>s</i> + <i>t</i>
sum		<i>q</i> + <i>s</i>	<i>r</i> + <i>t</i>	<i>p</i>

- Distance measure for symmetric binary variables:

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

- Distance measure for asymmetric binary variables:

$$d(i, j) = \frac{r + s}{q + r + s}$$

- Jaccard coefficient (*similarity* measure for *asymmetric* binary variables):

$$\text{sim}_{\text{Jaccard}}(i, j) = \frac{q}{q + r + s}$$

- Note: Jaccard coefficient is the same as “coherence”:

$$\text{coherence}(i, j) = \frac{\text{sup}(i, j)}{\text{sup}(i) + \text{sup}(j) - \text{sup}(i, j)} = \frac{q}{(q + r) + (q + s) - q}$$

Dissimilarity between Binary Variables

■ Example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is a symmetric attribute
- The remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N 0

$$d(jack, mary) = \frac{0 + 1}{2 + 0 + 1} = 0.33$$

$$d(jack, jim) = \frac{1 + 1}{1 + 1 + 1} = 0.67$$

$$d(jim, mary) = \frac{1 + 2}{1 + 1 + 2} = 0.75$$

Standardizing Numeric Data

- Z-score:
$$z = \frac{x - \mu}{\sigma}$$
 - X: raw score to be standardized, μ : mean of the population, σ : standard deviation
 - the distance between the raw score and the population mean in units of the standard deviation
 - negative when the raw score is below the mean, "+" when above
- An alternative way: Calculate the mean absolute deviation

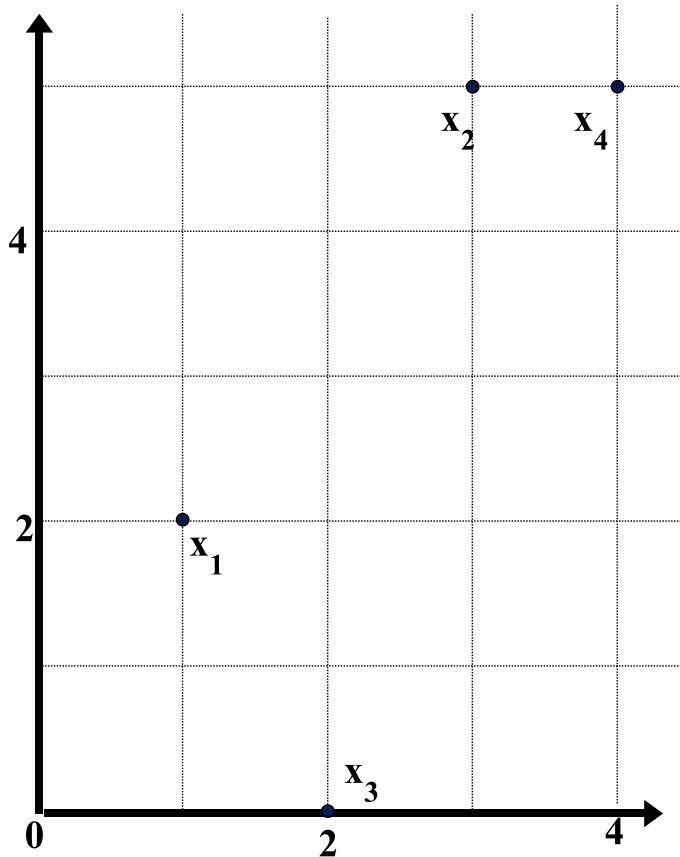
$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$

where $m_f = \frac{1}{n}(x_{1f} + x_{2f} + \dots + x_{nf})$.

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

- standardized measure (*z-score*):
- Using mean absolute deviation is more robust than using standard deviation

Example: Data Matrix and Dissimilarity Matrix



Data Matrix

point	attribute1	attribute2
$x1$	1	2
$x2$	3	5
$x3$	2	0
$x4$	4	5

Dissimilarity Matrix

(with Euclidean Distance)

	$x1$	$x2$	$x3$	$x4$
$x1$	0			
$x2$	3.61	0		
$x3$	5.1	5.1	0	
$x4$	4.24	1	5.39	0

Distance on Numeric Data: Minkowski Distance

- *Minkowski distance*: A popular distance measure

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

where $i = (x_{i1}, x_{i2}, \dots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ are two p -dimensional data objects, and h is the order (the distance so defined is also called L- h norm)

- Properties
 - $d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$ (Positive definiteness)
 - $d(i, j) = d(j, i)$ (Symmetry)
 - $d(i, j) \leq d(i, k) + d(k, j)$ (Triangle Inequality)
- A distance that satisfies these properties is a **metric**

Special Cases of Minkowski Distance

- $h = 1$: Manhattan (city block, L_1 norm) distance
 - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$

- $h = 2$: (L_2 norm) Euclidean distance

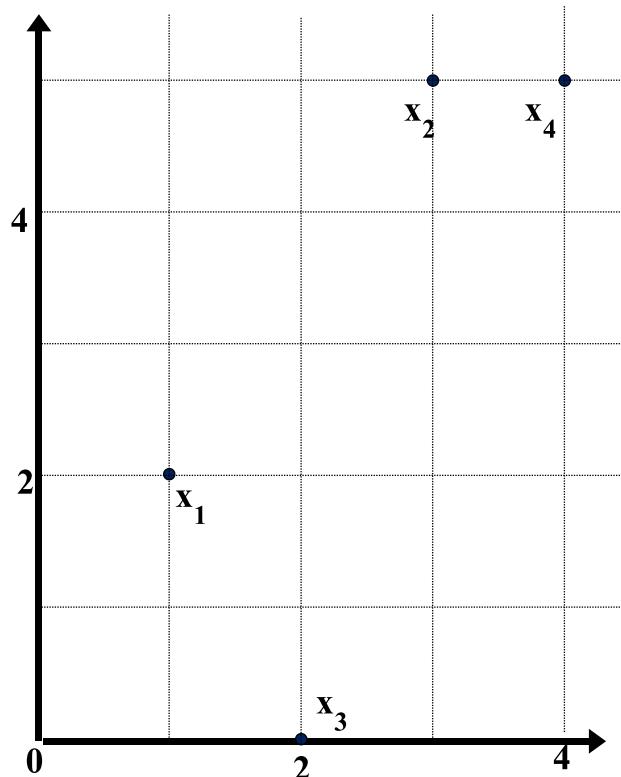
$$d(i, j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2)}$$

- $h \rightarrow \infty$. “supremum” (L_{\max} norm, L_∞ norm) distance.
 - This is the maximum difference between any component (attribute) of the vectors

$$d(i, j) = \lim_{h \rightarrow \infty} \left(\sum_{f=1}^p |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_f^p |x_{if} - x_{jf}|$$

Example: Minkowski Distance

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Dissimilarity Matrices

Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank
- Can be treated like interval-scaled
 - replace x_{if} by their rank $r_{if} \in \{1, \dots, M_f\}$
 - map the range of each variable onto [0, 1] by replacing i -th object in the f -th variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

- compute the dissimilarity using methods for interval-scaled variables

Attributes of Mixed Type

- A database may contain all attribute types
 - Nominal, symmetric binary, asymmetric binary, numeric, ordinal
- One may use a weighted formula to combine their effects

$$d(i, j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$$

- f is binary or nominal:
 $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$, or $d_{ij}^{(f)} = 1$ otherwise
- f is numeric: use the normalized distance
- f is ordinal
 - Compute ranks r_{if} and
 - Treat z_{if} as interval-scaled

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

Cosine Similarity

- A **document** can be represented by thousands of attributes, each recording the *frequency* of a particular word (such as keywords) or phrase in the document.

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- Other vector objects: gene features in micro-arrays, ...
- Applications: information retrieval, biologic taxonomy, gene feature mapping, ...
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / \|d_1\| \|d_2\| ,$$

where \bullet indicates vector dot product, $\|d\|$: the length of vector d

Example: Cosine Similarity

- $\cos(d_1, d_2) = (d_1 \bullet d_2) / \|d_1\| \|d_2\|$,
where \bullet indicates vector dot product, $\|d\|$: the length of vector d
- Ex: Find the **similarity** between documents 1 and 2.

$$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$$

$$d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$$

$$d_1 \bullet d_2 = 5*3 + 0*0 + 3*2 + 0*0 + 2*1 + 0*1 + 0*1 + 2*1 + 0*0 + 0*1 = 25$$

$$\|d_1\| = \sqrt{(5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2)} = \sqrt{42} \approx 6.481$$

$$\|d_2\| = \sqrt{(3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2)} = \sqrt{17} \approx 4.12$$

$$\cos(d_1, d_2) = 0.94$$

Chapter 2: Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary



Summary

- Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled
- Many types of data sets, e.g., numerical, text, graph, Web, image.
- Gain insight into the data by:
 - Basic statistical data description: central tendency, dispersion, graphical displays
 - Data visualization: map data onto graphical primitives
 - Measure data similarity
- Above steps are the beginning of data preprocessing.
- Many methods have been developed but still an active area of research.

References

- W. Cleveland, Visualizing Data, Hobart Press, 1993
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- H. V. Jagadish, et al., Special Issue on Data Reduction Techniques. Bulletin of the Tech. Committee on Data Eng., 20(4), Dec. 1997
- D. A. Keim. Information visualization and visual data mining, IEEE trans. on Visualization and Computer Graphics, 8(1), 2002
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- S. Santini and R. Jain," Similarity measures", IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999
- E. R. Tufte. The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001
- C. Yu , et al., Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009

Data Mining:

Concepts and Techniques

(3rd ed.)

— Chapter 3 —

Jiawei Han, Micheline Kamber, and Jian Pei
University of Illinois at Urbana-Champaign &
Simon Fraser University

©2011 Han, Kamber & Pei. All rights reserved.

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview 

 - Data Quality
 - Major Tasks in Data Preprocessing

- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Data Quality: Why Preprocess the Data?

- Measures for data quality: A multidimensional view
 - Accuracy: correct or wrong, accurate or not
 - Completeness: not recorded, unavailable, ...
 - Consistency: some modified but some not, dangling, ...
 - Timeliness: timely update?
 - Believability: how trustable the data are correct?
 - Interpretability: how easily the data can be understood?

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 reduction**
 - Dimensionality reduction
 - Numerosity reduction
 - Data compression
- **Data transformation and data discretization**
 - Normalization
 - Concept hierarchy generation

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
 - Data Quality
 - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary



Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
 - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
 - e.g., *Occupation*=“ ” (missing data)
 - noisy: containing noise, errors, or outliers
 - e.g., *Salary*=“-10”(an error)
 - inconsistent: containing discrepancies in codes or names, e.g.,
 - *Age*=“42”, *Birthday*=“03/07/2010”
 - Was rating “1, 2, 3”, now rating “A, B, C”
 - discrepancy between duplicate records
 - Intentional (e.g., *disguised missing* data)
 - Jan. 1 as everyone’s birthday?

Incomplete (Missing) Data

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

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % 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 be due to
 - faulty data collection instruments
 - data entry problems
 - data transmission problems
 - technology limitation
 - inconsistency in naming convention
- **Other data problems** which require data cleaning
 - duplicate records
 - incomplete data
 - inconsistent data

How to Handle Noisy Data?

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

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 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)

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
 - Data Quality
 - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- 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 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* and *covariance analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Correlation Analysis (Nominal Data)

■ **X² (chi-square) test**

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

- The larger the X² value, the more likely the variables are related
- The cells that contribute the most to the X² 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

- χ^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

Correlation Analysis (Numeric Data)

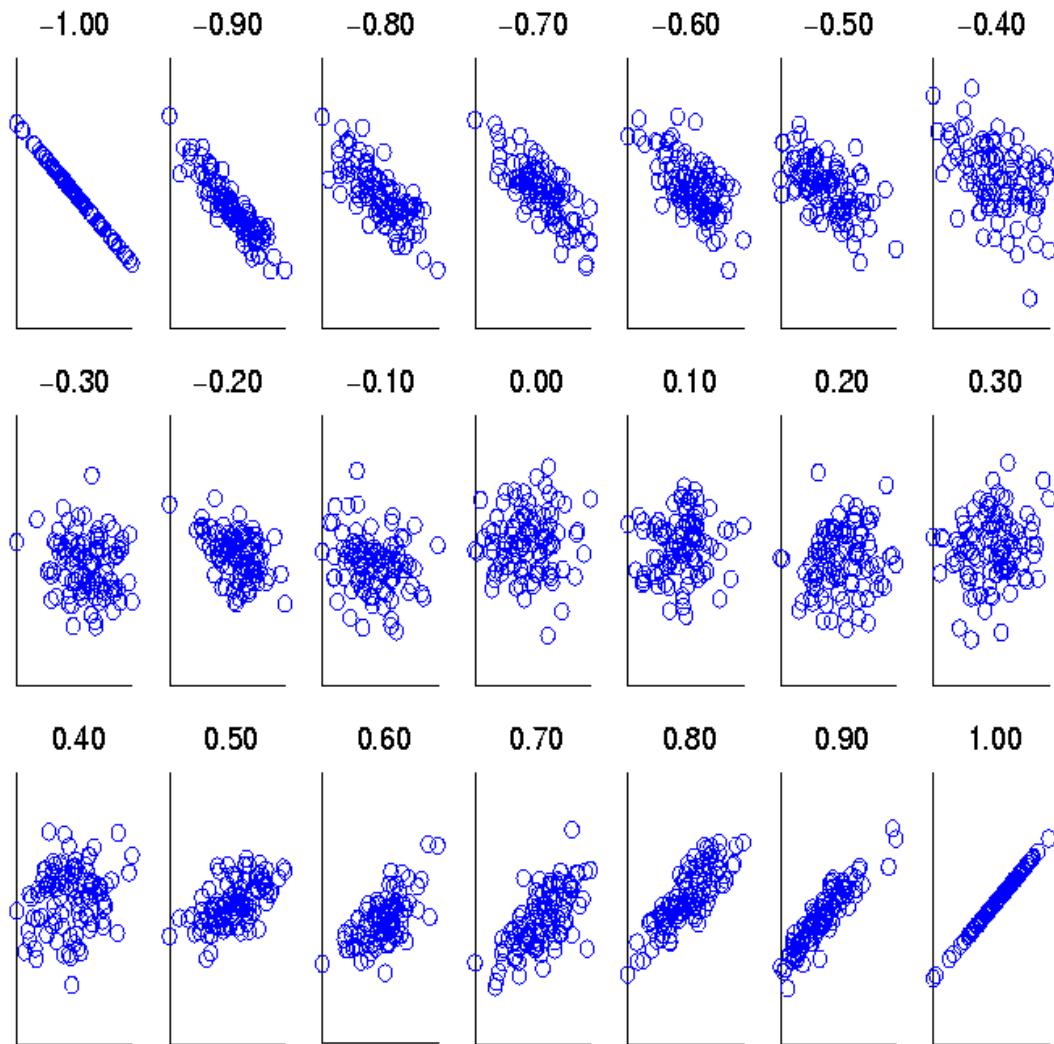
- Correlation coefficient (also called Pearson's product moment coefficient)

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

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 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; $r_{AB} < 0$: negatively correlated

Visually Evaluating Correlation



**Scatter plots
showing the
similarity from
-1 to 1.**

Correlation (viewed as linear relationship)

- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, A and B, and then take their dot product

$$a'_k = (a_k - \text{mean}(A)) / \text{std}(A)$$

$$b'_k = (b_k - \text{mean}(B)) / \text{std}(B)$$

$$\text{correlation}(A, B) = A' \bullet B'$$

Covariance (Numeric Data)

- Covariance is similar to correlation

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n}$$

Correlation coefficient: $r_{A,B} = \frac{Cov(A, B)}{\sigma_A \sigma_B}$

where n is the number of tuples, \bar{A} and \bar{B} are the respective mean or **expected values** of A and B, σ_A and σ_B are the respective standard deviation of A and B.

- **Positive covariance:** If $Cov_{A,B} > 0$, then A and B both tend to be larger than their expected values.
- **Negative covariance:** If $Cov_{A,B} < 0$ then if A is larger than its expected value, B is likely to be smaller than its expected value.
- **Independence:** $Cov_{A,B} = 0$ but the converse is not true:
 - Some pairs of random variables may have a covariance of 0 but are not independent. Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence

Co-Variance: An Example

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n}$$

- It can be simplified in computation as

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}$$

- Suppose two stocks A and B have the following values in one week:
(2, 5), (3, 8), (5, 10), (4, 11), (6, 14).
- Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?
 - $E(A) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4$
 - $E(B) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6$
 - $Cov(A,B) = (2 \times 5 + 3 \times 8 + 5 \times 10 + 4 \times 11 + 6 \times 14) /5 - 4 \times 9.6 = 4$
- Thus, A and B rise together since $Cov(A, B) > 0$.

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
 - Data Quality
 - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction 
- Data Transformation and Data Discretization
- Summary

Data Reduction Strategies

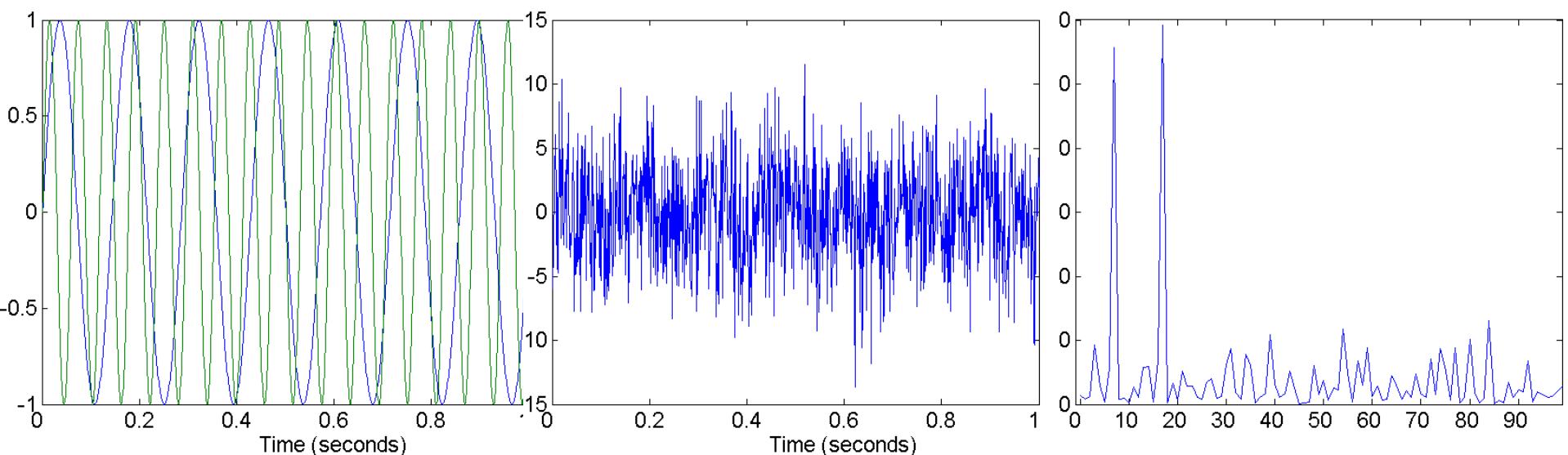
- **Data reduction:** Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.
- Data reduction strategies
 - Dimensionality reduction, e.g., remove unimportant attributes
 - Wavelet transforms
 - Principal Components Analysis (PCA)
 - Feature subset selection, feature creation
 - Numerosity reduction (some simply call it: Data Reduction)
 - Regression and Log-Linear Models
 - Histograms, clustering, sampling
 - Data cube aggregation
 - Data compression

Data Reduction 1: Dimensionality Reduction

- **Curse of dimensionality**
 - When dimensionality increases, data becomes increasingly sparse
 - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
 - The possible combinations of subspaces will grow exponentially
- **Dimensionality reduction**
 - Avoid the curse of dimensionality
 - Help eliminate irrelevant features and reduce noise
 - Reduce time and space required in data mining
 - Allow easier visualization
- **Dimensionality reduction techniques**
 - Wavelet transforms
 - Principal Component Analysis
 - Supervised and nonlinear techniques (e.g., feature selection)

Mapping Data to a New Space

- Fourier transform
- Wavelet transform



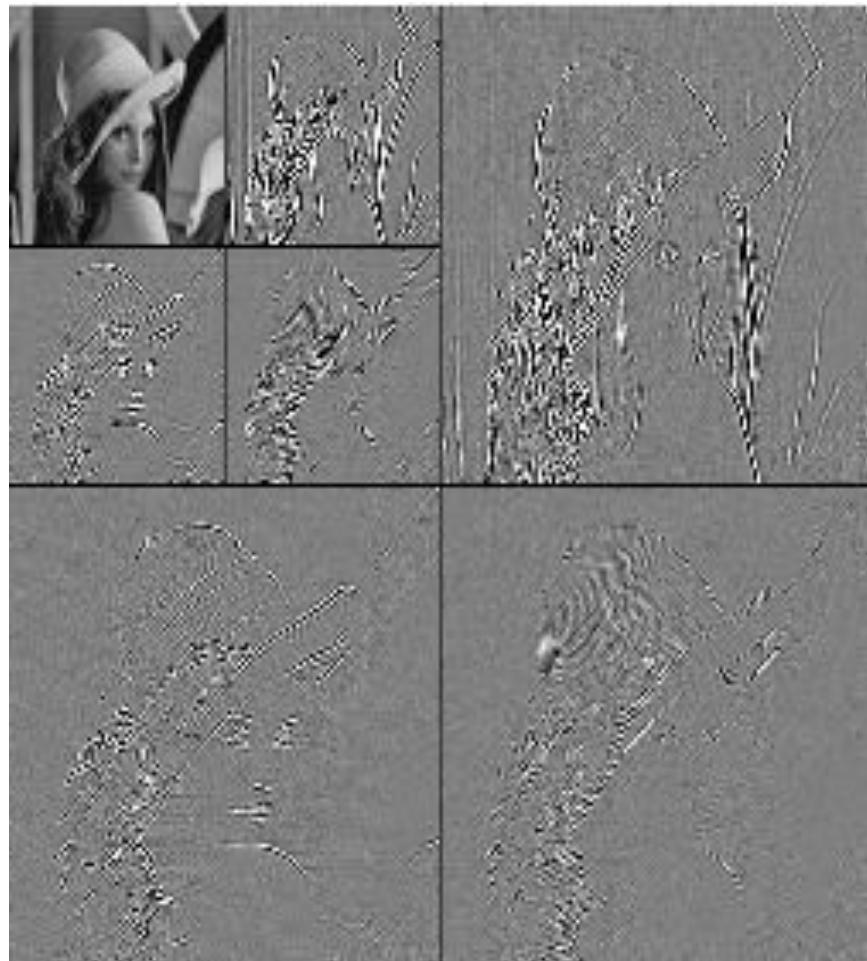
Two Sine Waves

Two Sine Waves + Noise

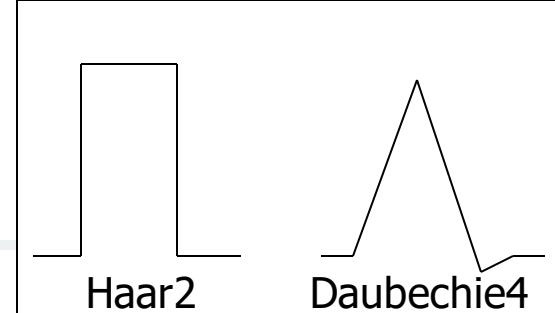
Frequency

What Is Wavelet Transform?

- Decomposes a signal into different frequency subbands
 - Applicable to n-dimensional signals
- Data are transformed to preserve relative distance between objects at different levels of resolution
- Allow natural clusters to become more distinguishable
- Used for image compression



Wavelet Transformation



- Discrete wavelet transform (DWT) for linear signal processing, multi-resolution analysis
- Compressed approximation: store only a small fraction of the strongest of the 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 set of data of length $L/2$
 - Applies two functions recursively, until reaches the desired length

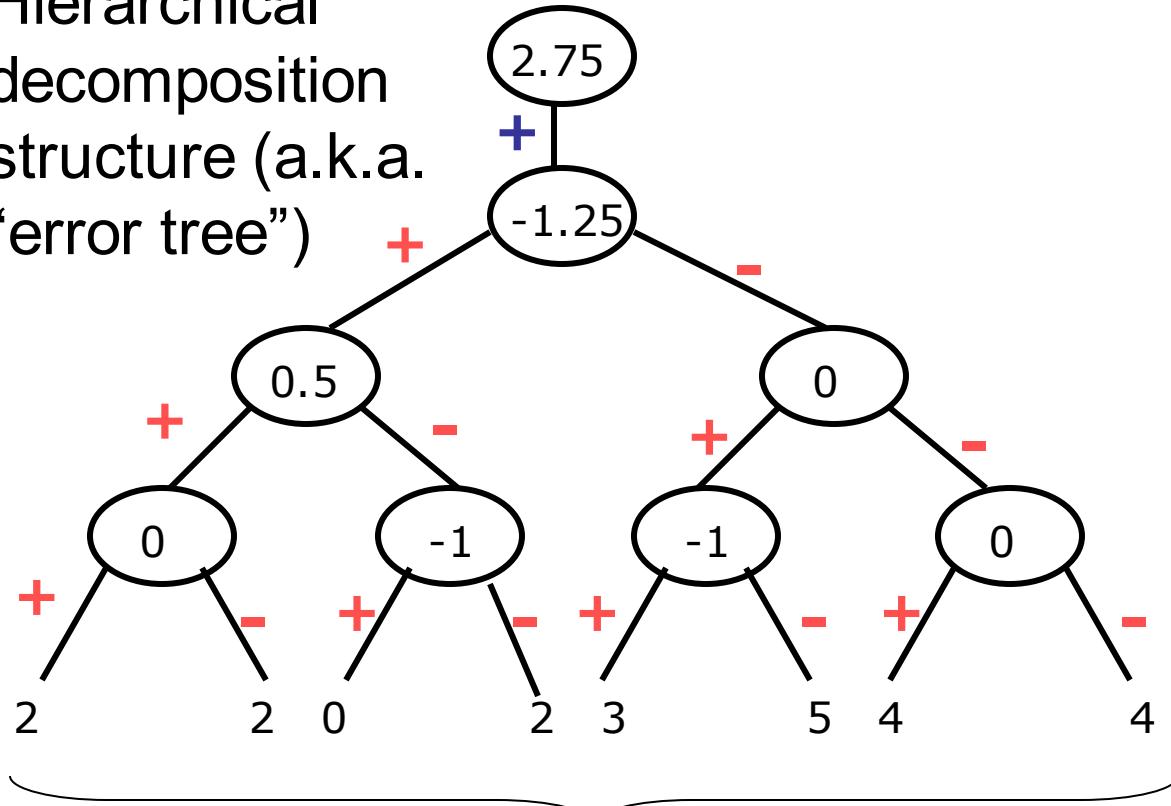
Wavelet Decomposition

- Wavelets: A math tool for space-efficient hierarchical decomposition of functions
- $S = [2, 2, 0, 2, 3, 5, 4, 4]$ can be transformed to $S_\wedge = [2^3/4, -1^1/4, 1/2, 0, 0, -1, -1, 0]$
- Compression: many small detail coefficients can be replaced by 0's, and only the significant coefficients are retained

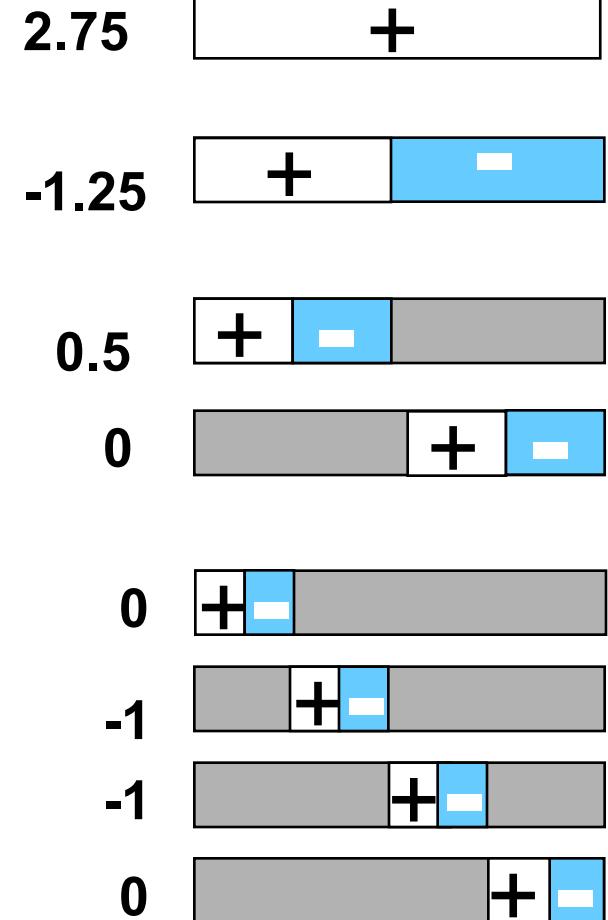
Resolution	Averages	Detail Coefficients
8	$[2, 2, 0, 2, 3, 5, 4, 4]$	
4	$[2, 1, 4, 4]$	$[0, -1, -1, 0]$
2	$[1\frac{1}{2}, 4]$	$[\frac{1}{2}, 0]$
1	$[2\frac{3}{4}]$	$[-1\frac{1}{4}]$

Haar Wavelet Coefficients

Hierarchical decomposition structure (a.k.a.
“error tree”)



Coefficient “Supports”

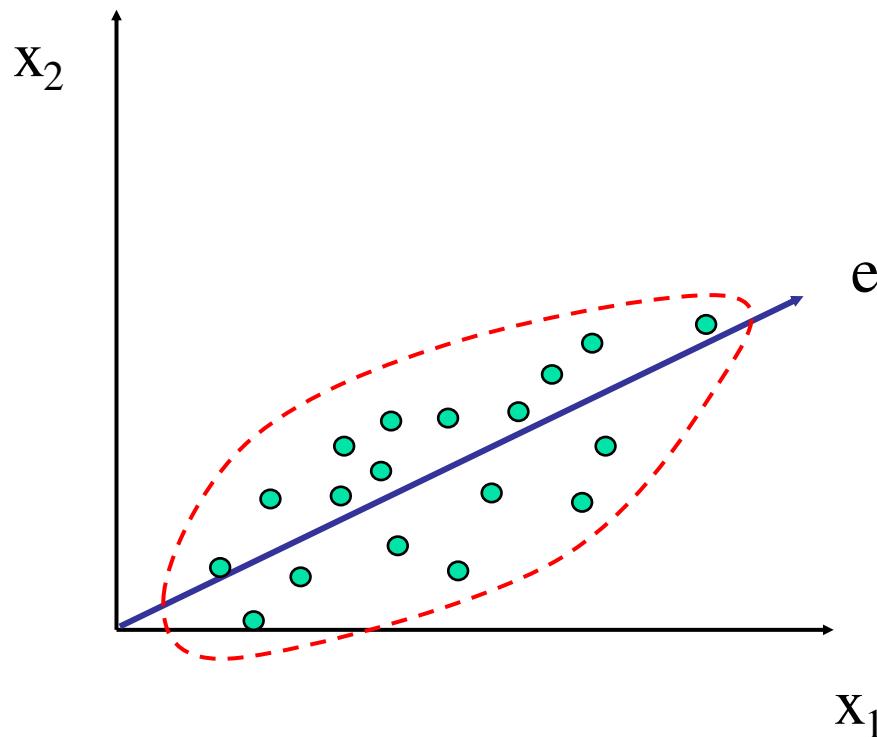


Why Wavelet Transform?

- Use hat-shape filters
 - Emphasize region where points cluster
 - Suppress weaker information in their boundaries
- Effective removal of outliers
 - Insensitive to noise, insensitive to input order
- Multi-resolution
 - Detect arbitrary shaped clusters at different scales
- Efficient
 - Complexity $O(N)$
- Only applicable to low dimensional data

Principal Component Analysis (PCA)

- Find a projection that captures the largest amount of variation in data
- The original data are projected onto a much smaller space, resulting in dimensionality reduction. We find the eigenvectors of the covariance matrix, and these eigenvectors define the new space



Principal Component Analysis (Steps)

- Given N data vectors from n -dimensions, find $k \leq n$ orthogonal vectors (*principal components*) that can be best used to represent data
 - Normalize input data: Each attribute falls within the same range
 - Compute k orthonormal (unit) 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)
- Works for numeric data only

Attribute Subset Selection

- Another way to reduce dimensionality of data
- Redundant attributes
 - Duplicate much or all of the information contained in one or more other attributes
 - E.g., purchase price of a product and the amount of sales tax paid
- Irrelevant attributes
 - Contain no information that is useful for the data mining task at hand
 - E.g., students' ID is often irrelevant to the task of predicting students' GPA

Heuristic Search in Attribute Selection

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

Attribute Creation (Feature Generation)

- Create new attributes (features) that can capture the important information in a data set more effectively than the original ones
- Three general methodologies
 - Attribute extraction
 - Domain-specific
 - Mapping data to new space (see: data reduction)
 - E.g., Fourier transformation, wavelet transformation, manifold approaches (not covered)
 - Attribute construction
 - Combining features (see: discriminative frequent patterns in Chapter 7)
 - Data discretization

Data Reduction 2: Numerosity Reduction

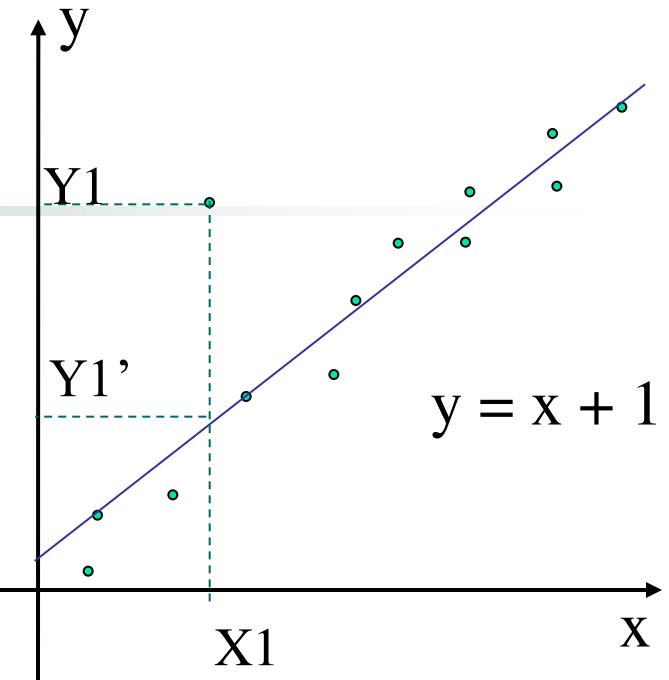
- Reduce data volume by choosing alternative, *smaller forms* of data representation
- **Parametric methods** (e.g., regression)
 - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
 - Ex.: Log-linear models—obtain value at a point in m -D space as the product on appropriate marginal subspaces
- **Non-parametric** methods
 - Do not assume models
 - Major families: histograms, clustering, sampling, ...

Parametric Data Reduction: Regression and Log-Linear Models

- **Linear regression**
 - Data modeled to fit a straight line
 - Often uses the least-square method to fit the line
- **Multiple regression**
 - Allows a response variable Y to be modeled as a linear function of multidimensional feature vector
- **Log-linear model**
 - Approximates discrete multidimensional probability distributions

Regression Analysis

- Regression analysis: A collective name for techniques for the modeling and analysis of numerical data consisting of values of a **dependent variable** (also called **response variable** or *measurement*) and of one or more *independent variables* (aka. **explanatory variables** or **predictors**)



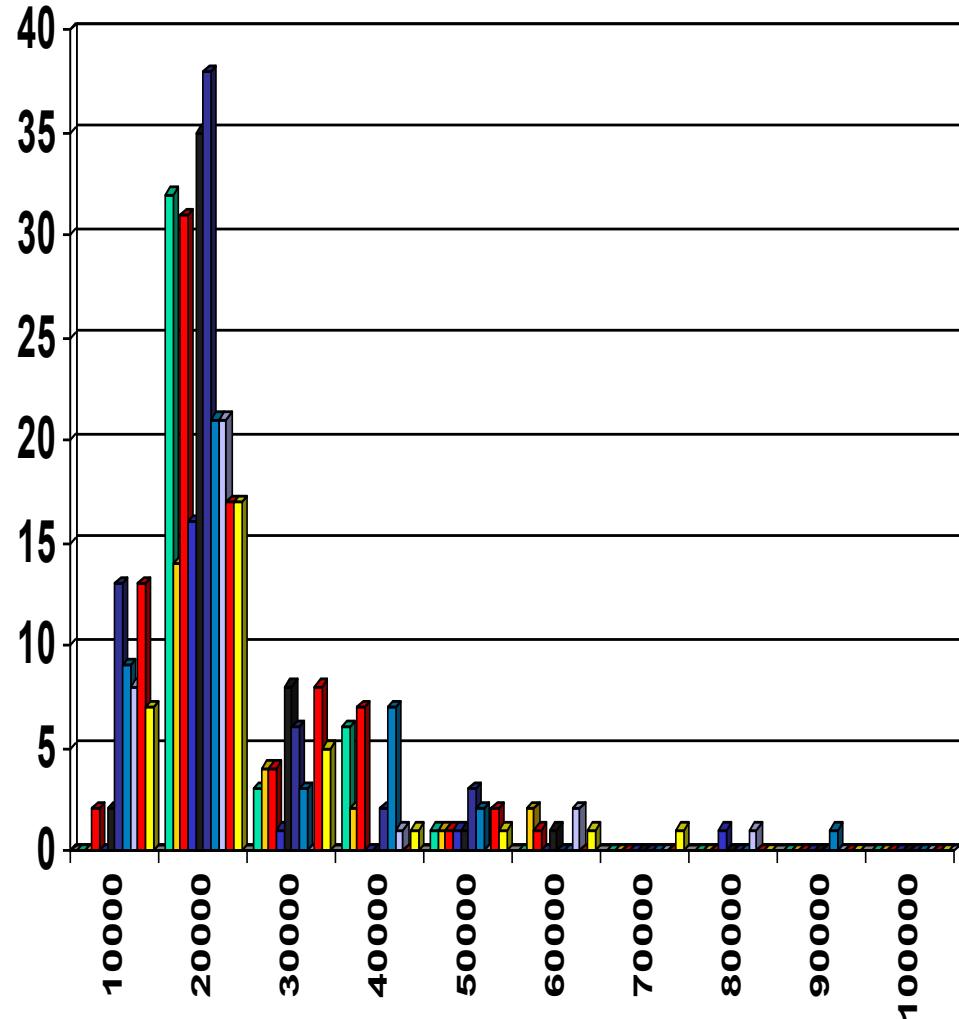
- The parameters are estimated so as to give a "**best fit**" of the data
- Most commonly the best fit is evaluated by using the **least squares method**, but other criteria have also been used
- Used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships

Régress Analysis and Log-Linear Models

- Linear regression: $Y = wX + b$
 - Two regression coefficients, w and b , specify the line and are to be estimated by using the data at hand
 - Using the least squares criterion to the known values of $Y_1, Y_2, \dots, X_1, X_2, \dots$
- Multiple regression: $Y = b_0 + b_1 X_1 + b_2 X_2$
 - Many nonlinear functions can be transformed into the above
- Log-linear models:
 - Approximate discrete multidimensional probability distributions
 - Estimate the probability of each point (tuple) in a multi-dimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations
 - Useful for dimensionality reduction and data smoothing

Histogram Analysis

- Divide data into buckets and store average (sum) for each bucket
- Partitioning rules:
 - Equal-width: equal bucket range
 - Equal-frequency (or equal-depth)



Clustering

- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only
- Can be very effective if data is clustered but not if data is “smeared”
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms
- Cluster analysis will be studied in depth in Chapter 10

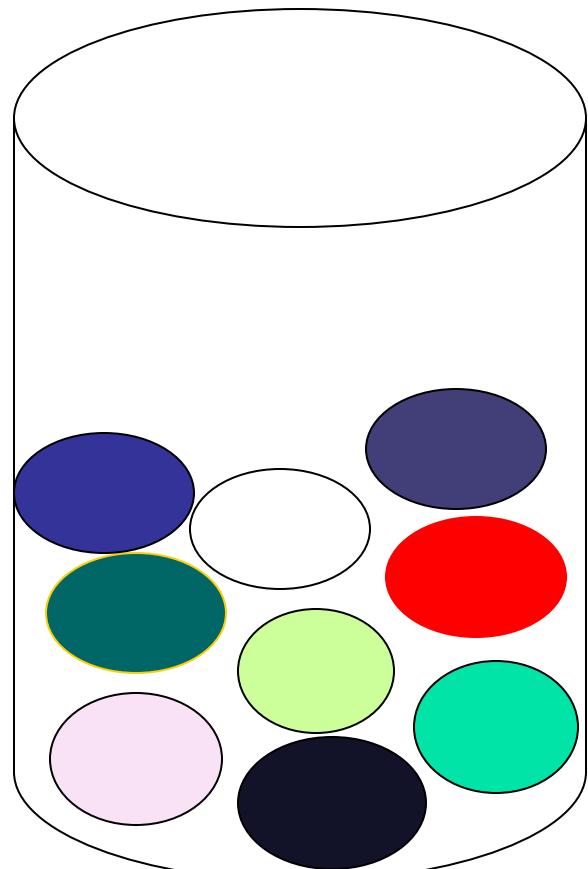
Sampling

- 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
- Key principle: Choose a **representative** subset of the data
 - Simple random sampling may have very poor performance in the presence of skew
 - Develop adaptive sampling methods, e.g., stratified sampling:
- Note: Sampling may not reduce database I/Os (page at a time)

Types of Sampling

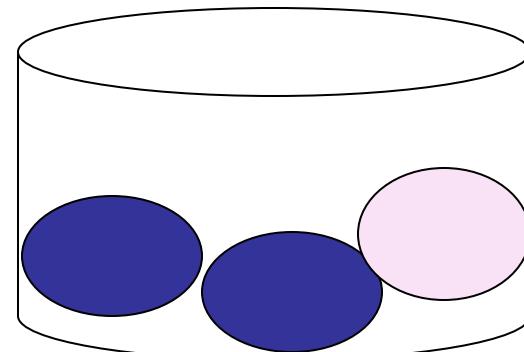
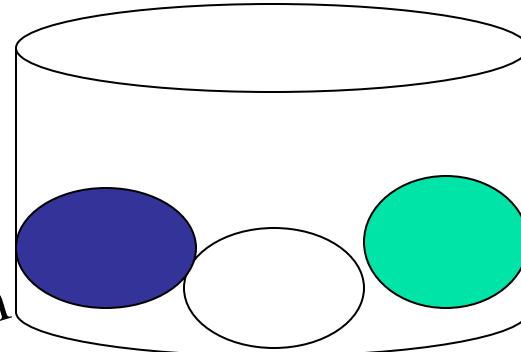
- **Simple random sampling**
 - There is an equal probability of selecting any particular item
- **Sampling without replacement**
 - Once an object is selected, it is removed from the population
- **Sampling with replacement**
 - A selected object is not removed from the population
- **Stratified sampling:**
 - Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
 - Used in conjunction with skewed data

Sampling: With or without Replacement



SRSWOR
(simple random
sample without
replacement)

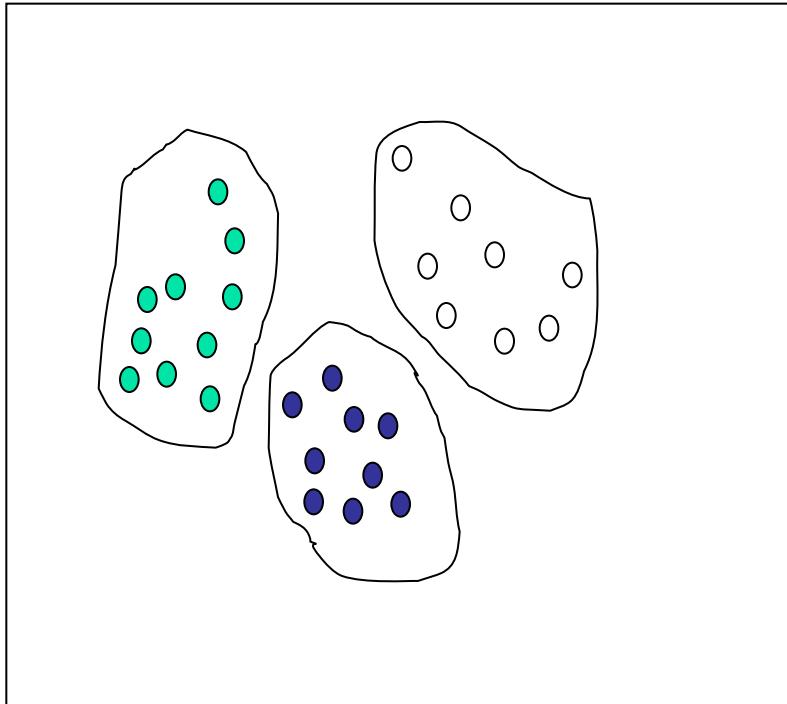
SRSWR



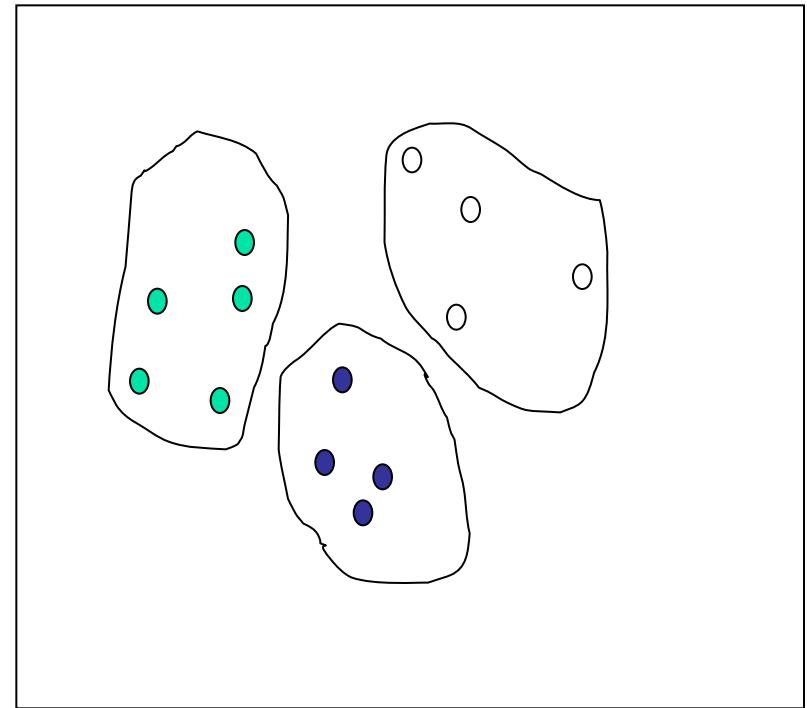
Raw Data

Sampling: Cluster or Stratified Sampling

Raw Data



Cluster/Stratified Sample



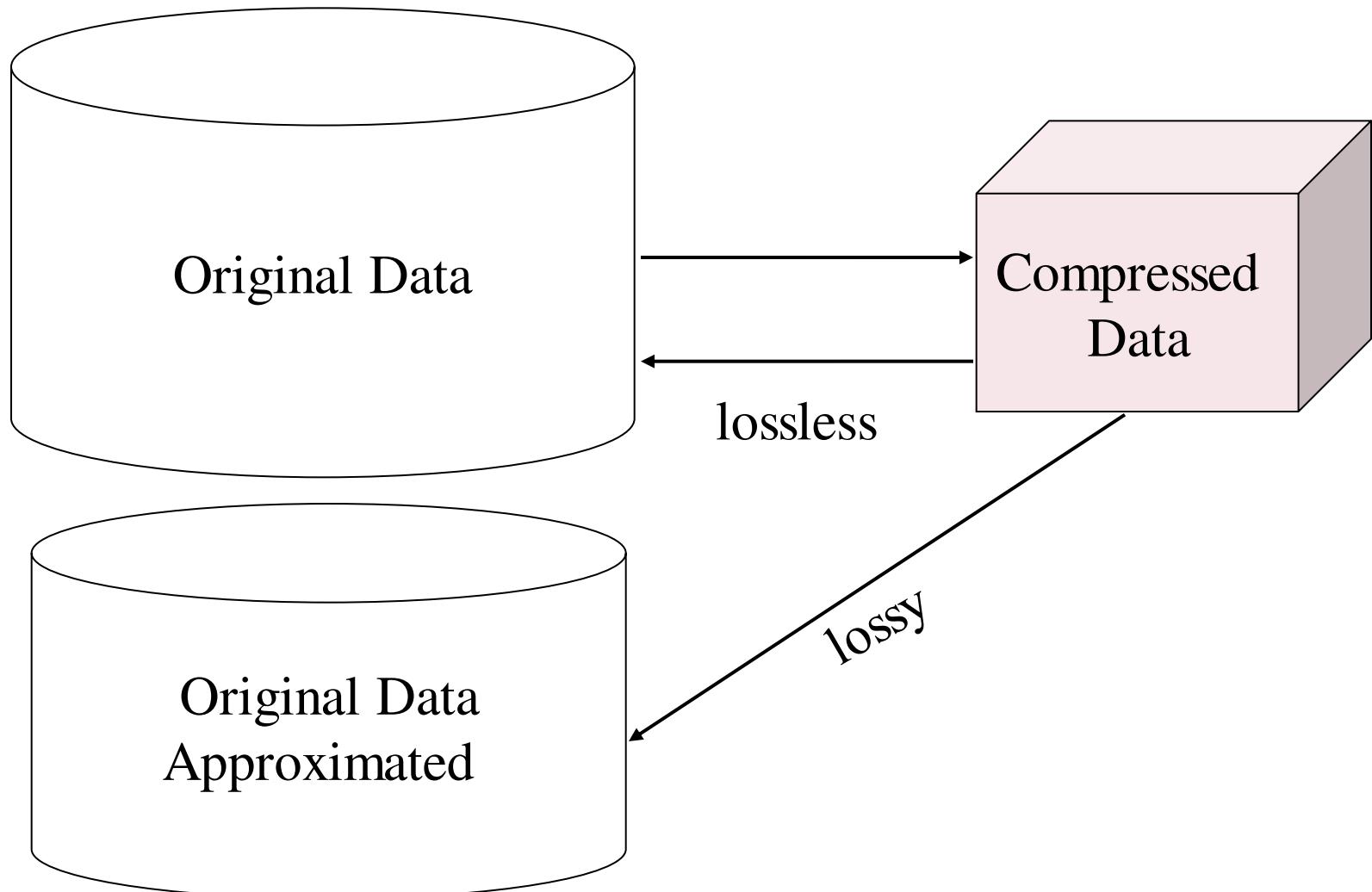
Data Cube Aggregation

- 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

Data Reduction 3: 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
- Dimensionality and numerosity reduction may also be considered as forms of data compression

Data Compression



Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
 - Data Quality
 - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary



Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values s.t. each old value can be identified with one of the new values
- Methods
 - Smoothing: Remove noise from data
 - Attribute/feature construction
 - New attributes constructed from the given ones
 - Aggregation: Summarization, data cube construction
 - Normalization: Scaled to fall within a smaller, specified range
 - min-max normalization
 - z-score normalization
 - normalization by decimal scaling
 - Discretization: Concept hierarchy climbing

Normalization

- **Min-max normalization:** to $[new_min_A, new_max_A]$

$$v' = \frac{v - min_A}{max_A - min_A} (new_max_A - new_min_A) + new_min_A$$

- Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73,600 - 12,000}{98,000 - 12,000} (1.0 - 0) + 0 = 0.716$

- **Z-score normalization** (μ : mean, σ : standard deviation):

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

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

- **Normalization by decimal scaling**

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

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
 - Numeric—real numbers, e.g., integer or real numbers
- Discretization: Divide the range of a continuous attribute into intervals
 - Interval labels can then be used to replace actual data values
 - Reduce data size by discretization
 - Supervised vs. unsupervised
 - Split (top-down) vs. merge (bottom-up)
 - Discretization can be performed recursively on an attribute
 - Prepare for further analysis, e.g., classification

Data Discretization Methods

- Typical methods: All the methods can be applied recursively
 - Binning
 - Top-down split, unsupervised
 - Histogram analysis
 - Top-down split, unsupervised
 - Clustering analysis (unsupervised, top-down split or bottom-up merge)
 - Decision-tree analysis (supervised, top-down split)
 - Correlation (e.g., χ^2) analysis (unsupervised, bottom-up merge)

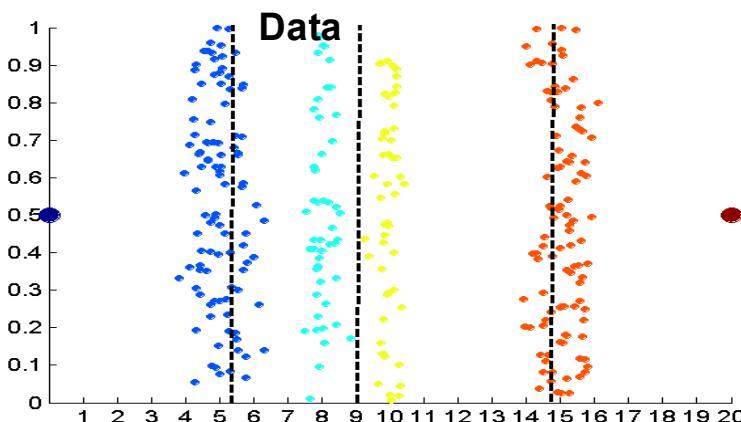
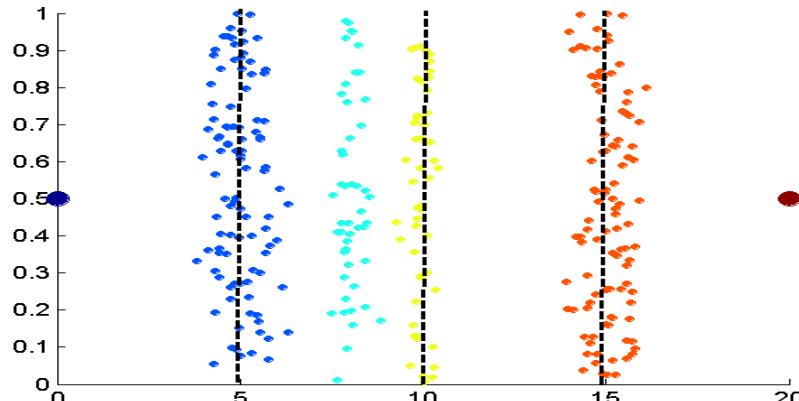
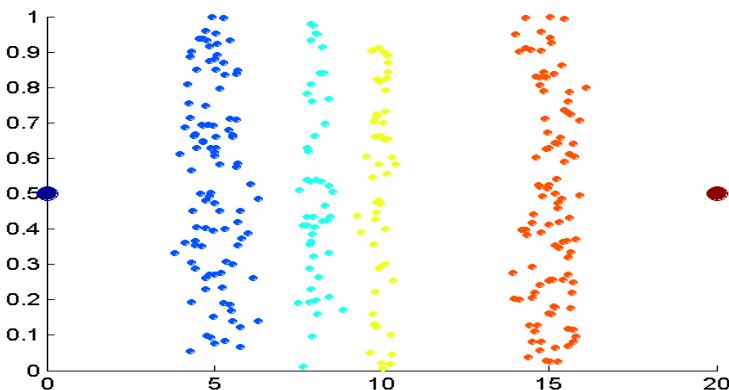
Simple Discretization: Binning

- 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

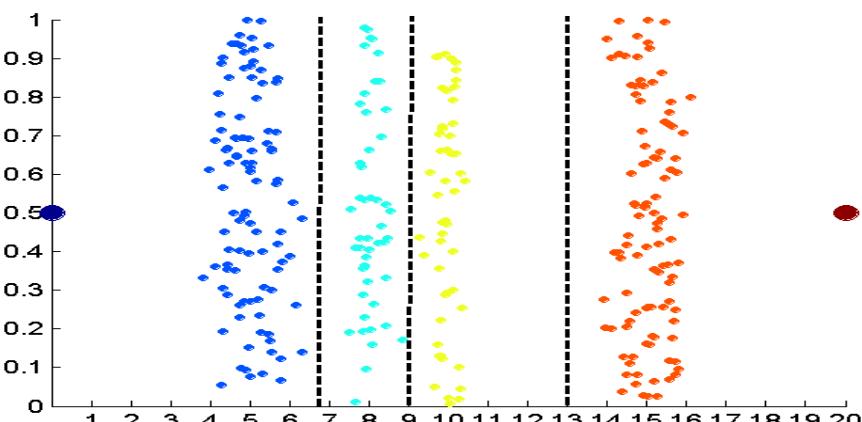
Binning Methods for Data Smoothing

- 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

Discretization Without Using Class Labels (Binning vs. Clustering)



Equal frequency (binning)



K-means clustering leads to better results

Discretization by Classification & Correlation Analysis

- Classification (e.g., decision tree analysis)
 - Supervised: Given class labels, e.g., cancerous vs. benign
 - Using *entropy* to determine split point (discretization point)
 - Top-down, recursive split
 - Details to be covered in Chapter 7
- Correlation analysis (e.g., Chi-merge: χ^2 -based discretization)
 - Supervised: use class information
 - Bottom-up merge: find the best neighboring intervals (those having similar distributions of classes, i.e., low χ^2 values) to merge
 - Merge performed recursively, until a predefined stopping condition

Concept Hierarchy Generation

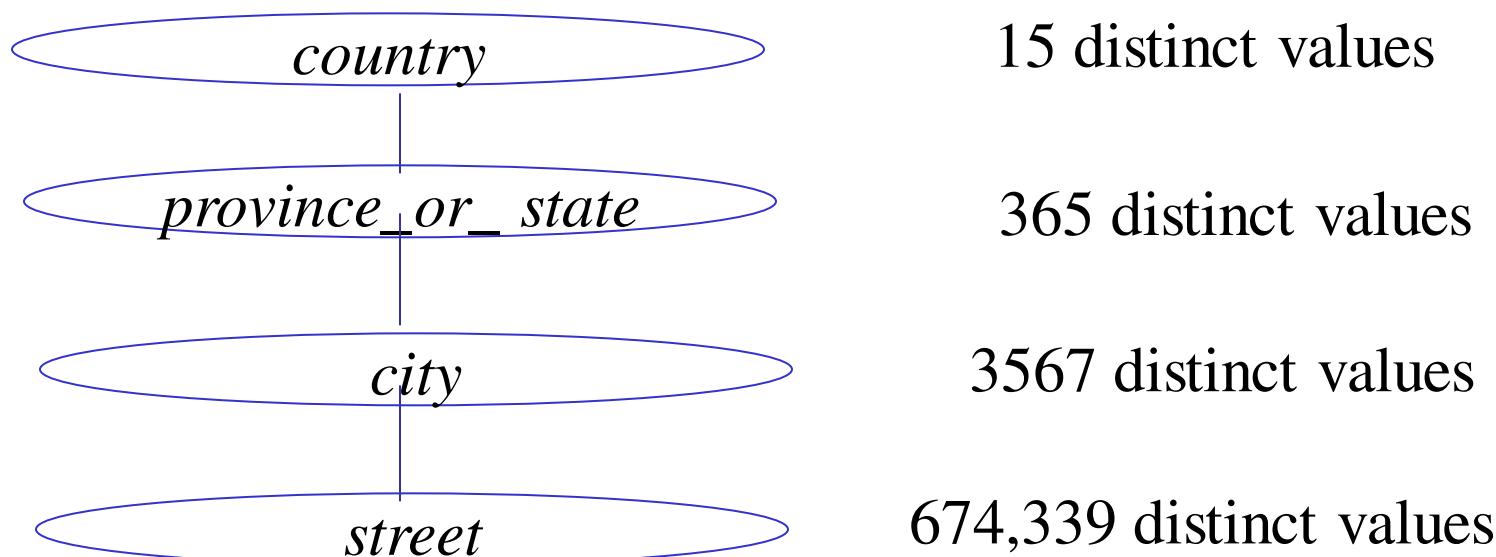
- **Concept hierarchy** organizes concepts (i.e., attribute values) hierarchically and is usually associated with each dimension in a data warehouse
- Concept hierarchies facilitate drilling and rolling in data warehouses to view data in multiple granularity
- Concept hierarchy formation: Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for *age*) by higher level concepts (such as *youth*, *adult*, or *senior*)
- Concept hierarchies can be explicitly specified by domain experts and/or data warehouse designers
- Concept hierarchy can be automatically formed for both numeric and nominal data. For numeric data, use discretization methods shown.

Concept Hierarchy Generation for Nominal Data

- Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
 - $\textit{street} < \textit{city} < \textit{state} < \textit{country}$
- Specification of a hierarchy for a set of values by explicit data grouping
 - $\{\text{Urbana, Champaign, Chicago}\} < \text{Illinois}$
- Specification of only a partial set of attributes
 - E.g., only $\textit{street} < \textit{city}$, 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: $\{\textit{street}, \textit{city}, \textit{state}, \textit{country}\}$

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



Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
 - Data Quality
 - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary 

Summary

- **Data quality:** accuracy, completeness, consistency, timeliness, believability, interpretability
- **Data cleaning:** e.g. missing/noisy values, outliers
- **Data integration** from multiple sources:
 - Entity identification problem
 - Remove redundancies
 - Detect inconsistencies
- **Data reduction**
 - Dimensionality reduction
 - Numerosity reduction
 - Data compression
- **Data transformation and data discretization**
 - Normalization
 - Concept hierarchy generation

References

- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. Comm. of ACM, 42:73-78, 1999
- A. Bruce, D. Donoho, and H.-Y. Gao. Wavelet analysis. *IEEE Spectrum*, Oct 1996
- T. Dasu and T. Johnson. *Exploratory Data Mining and Data Cleaning*. John Wiley, 2003
- J. Devore and R. Peck. *Statistics: The Exploration and Analysis of Data*. Duxbury Press, 1997.
- H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C.-A. Saita. Declarative data cleaning: Language, model, and algorithms. *VLDB'01*
- M. Hua and J. Pei. Cleaning disguised missing data: A heuristic approach. *KDD'07*
- H. V. Jagadish, et al., Special Issue on Data Reduction Techniques. *Bulletin of the Technical Committee on Data Engineering*, 20(4), Dec. 1997
- H. Liu and H. Motoda (eds.). *Feature Extraction, Construction, and Selection: A Data Mining Perspective*. Kluwer Academic, 1998
- J. E. Olson. *Data Quality: The Accuracy Dimension*. Morgan Kaufmann, 2003
- D. Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann, 1999
- V. Raman and J. Hellerstein. *Potters Wheel: An Interactive Framework for Data Cleaning and Transformation*, VLDB'2001
- T. Redman. *Data Quality: The Field Guide*. Digital Press (Elsevier), 2001
- R. Wang, V. Storey, and C. Firth. A framework for analysis of data quality research. *IEEE Trans. Knowledge and Data Engineering*, 7:623-640, 1995