

Data Mining: Concepts and Techniques

— Slides for Textbook —
— Chapter 10 —

©Jiawei Han and Micheline Kamber
Department of Computer Science
University of Illinois at Urbana-Champaign

www.cs.uiuc.edu/~hanj

Chapter 10: Applications and Trends in Data Mining

- Data mining applications
- Data mining system products and research prototypes
- Additional themes on data mining
- Social impacts of data mining
- Trends in data mining
- Summary

Data Mining Applications

- Data mining is a young discipline with wide and diverse applications
 - There is still a nontrivial gap between general principles of data mining and domain-specific, effective data mining tools for particular applications
- Some application domains (covered in this chapter)
 - Biomedical and DNA data analysis
 - Financial data analysis
 - Retail industry
 - Telecommunication industry

Biomedical and DNA Data Analysis

- DNA sequences: 4 basic building blocks (nucleotides): adenine (A), cytosine (C), guanine (G), and thymine (T).
- Gene: a sequence of hundreds of individual nucleotides arranged in a particular order
- Humans have around 30,000 genes
- Tremendous number of ways that the nucleotides can be ordered and sequenced to form distinct genes
- Semantic integration of heterogeneous, distributed genome databases
 - Current: highly distributed, uncontrolled generation and use of a wide variety of DNA data
 - Data cleaning and data integration methods developed in data mining will help

DNA Analysis: Examples

- Similarity search and comparison among DNA sequences
 - Compare the frequently occurring patterns of each class (e.g., diseased and healthy)
 - Identify gene sequence patterns that play roles in various diseases
- Association analysis: identification of co-occurring gene sequences
 - Most diseases are not triggered by a single gene but by a combination of genes acting together
 - Association analysis may help determine the kinds of genes that are likely to co-occur together in target samples
- Path analysis: linking genes to different disease development stages
 - Different genes may become active at different stages of the disease
 - Develop pharmaceutical interventions that target the different stages separately
- Visualization tools and genetic data analysis

Data Mining for Financial Data Analysis

- Financial data collected in banks and financial institutions are often relatively complete, reliable, and of high quality
- Design and construction of data warehouses for multidimensional data analysis and data mining
 - View the debt and revenue changes by month, by region, by sector, and by other factors
 - Access statistical information such as max, min, total, average, trend, etc.
- Loan payment prediction/consumer credit policy analysis
 - feature selection and attribute relevance ranking
 - Loan payment performance
 - Consumer credit rating

Financial Data Mining

- Classification and clustering of customers for targeted marketing
 - multidimensional segmentation by nearest-neighbor, classification, decision trees, etc. to identify customer groups or associate a new customer to an appropriate customer group
- Detection of money laundering and other financial crimes
 - integration of from multiple DBs (e.g., bank transactions, federal/state crime history DBs)
 - Tools: data visualization, linkage analysis, classification, clustering tools, outlier analysis, and sequential pattern analysis tools (find unusual access sequences)

Data Mining for Retail Industry

- Retail industry: huge amounts of data on sales, customer shopping history, etc.
- Applications of retail data mining
 - Identify customer buying behaviors
 - Discover customer shopping patterns and trends
 - Improve the quality of customer service
 - Achieve better customer retention and satisfaction
 - Enhance goods consumption ratios
 - Design more effective goods transportation and distribution policies

Data Mining in Retail Industry: Examples

- Design and construction of data warehouses based on the benefits of data mining
 - Multidimensional analysis of sales, customers, products, time, and region
- Analysis of the effectiveness of sales campaigns
- Customer retention: Analysis of customer loyalty
 - Use customer loyalty card information to register sequences of purchases of particular customers
 - Use sequential pattern mining to investigate changes in customer consumption or loyalty
 - Suggest adjustments on the pricing and variety of goods
- Purchase recommendation and cross-reference of items

Data Mining for Telecomm. Industry (1)

- A rapidly expanding and highly competitive industry and a great demand for data mining
 - Understand the business involved
 - Identify telecommunication patterns
 - Catch fraudulent activities
 - Make better use of resources
 - Improve the quality of service
- Multidimensional analysis of telecommunication data
 - Intrinsically multidimensional: calling-time, duration, location of caller, location of callee, type of call, etc.

Data Mining for Telecomm. Industry (2)

- Fraudulent pattern analysis and the identification of unusual patterns
 - Identify potentially fraudulent users and their atypical usage patterns
 - Detect attempts to gain fraudulent entry to customer accounts
 - Discover unusual patterns which may need special attention
- Multidimensional association and sequential pattern analysis
 - Find usage patterns for a set of communication services by customer group, by month, etc.
 - Promote the sales of specific services
 - Improve the availability of particular services in a region
- Use of visualization tools in telecommunication data analysis

Chapter 10: Applications and Trends in Data Mining

- Data mining applications
- Data mining system products and research prototypes
- Additional themes on data mining
- Social impact of data mining
- Trends in data mining
- Summary

How to Choose a Data Mining System?

- Commercial data mining systems have little in common
 - Different data mining functionality or methodology
 - May even work with completely different kinds of data sets
- Need multiple dimensional view in selection
- Data types: relational, transactional, text, time sequence, spatial?
- System issues
 - running on only one or on several operating systems?
 - a client/server architecture?
 - Provide Web-based interfaces and allow XML data as input and/or output?

How to Choose a Data Mining System? (2)

- Data sources
 - ASCII text files, multiple relational data sources
 - support ODBC connections (OLE DB, JDBC)?
- Data mining functions and methodologies
 - One vs. multiple data mining functions
 - One vs. variety of methods per function
 - More data mining functions and methods per function provide the user with greater flexibility and analysis power
- Coupling with DB and/or data warehouse systems
 - Four forms of coupling: no coupling, loose coupling, semitight coupling, and tight coupling
 - Ideally, a data mining system should be tightly coupled with a database system

How to Choose a Data Mining System? (3)

- Scalability
 - Row (or database size) scalability
 - Column (or dimension) scalability
 - Curse of dimensionality: it is much more challenging to make a system column scalable than row scalable
- Visualization tools
 - “A picture is worth a thousand words”
 - Visualization categories: data visualization, mining result visualization, mining process visualization, and visual data mining
- Data mining query language and graphical user interface
 - Easy-to-use and high-quality graphical user interface
 - Essential for user-guided, highly interactive data mining

Examples of Data Mining Systems (1)

- **IBM Intelligent Miner**
 - A wide range of data mining algorithms
 - Scalable mining algorithms
 - Toolkits: neural network algorithms, statistical methods, data preparation, and data visualization tools
 - Tight integration with IBM's DB2 relational database system
- **SAS Enterprise Miner**
 - A variety of statistical analysis tools
 - Data warehouse tools and multiple data mining algorithms
- **Microsoft SQLServer 2000**
 - Integrate DB and OLAP with mining
 - Support OLEDB for DM standard

Examples of Data Mining Systems (2)

- **SGI MineSet**
 - Multiple data mining algorithms and advanced statistics
 - Advanced visualization tools
- **Clementine (SPSS)**
 - An integrated data mining development environment for end-users and developers
 - Multiple data mining algorithms and visualization tools
- **DBMiner (DBMiner Technology Inc.)**
 - Multiple data mining modules: discovery-driven OLAP analysis, association, classification, and clustering
 - Efficient, association and sequential-pattern mining functions, and visual classification tool

Chapter 10: Applications and Trends in Data Mining

- Data mining applications
- Data mining system products and research prototypes
- **Additional themes on data mining**
- Social impact of data mining
- Trends in data mining
- Summary

Visual Data Mining

- **Visualization**: use of computer graphics to create visual images which aid in the understanding of complex, often massive representations of data
- **Visual Data Mining**: the process of discovering implicit but useful knowledge from large data sets using visualization techniques

Computer
Graphics

Multimedia
Systems

Human
Computer
Interfaces

High
Performance
Computing

Pattern
Recognition

Visualization

- Purpose of 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

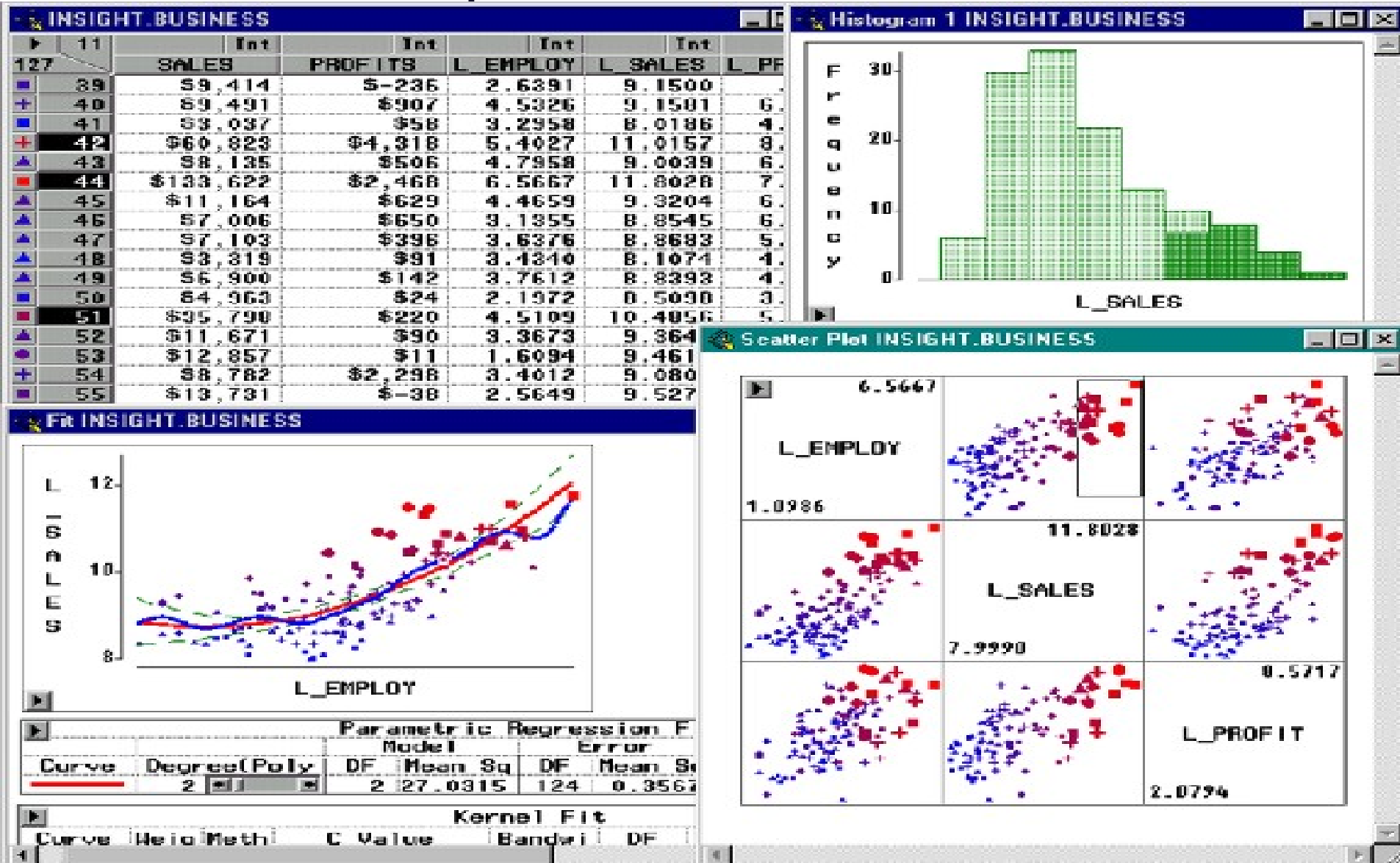
Visual Data Mining & Data Visualization

- Integration of visualization and data mining
 - data visualization
 - data mining result visualization
 - data mining process visualization
 - interactive visual data mining
- Data visualization
 - Data in a database or data warehouse can be viewed
 - at different levels of abstraction
 - as different combinations of attributes or dimensions
 - Data can be presented in various visual forms

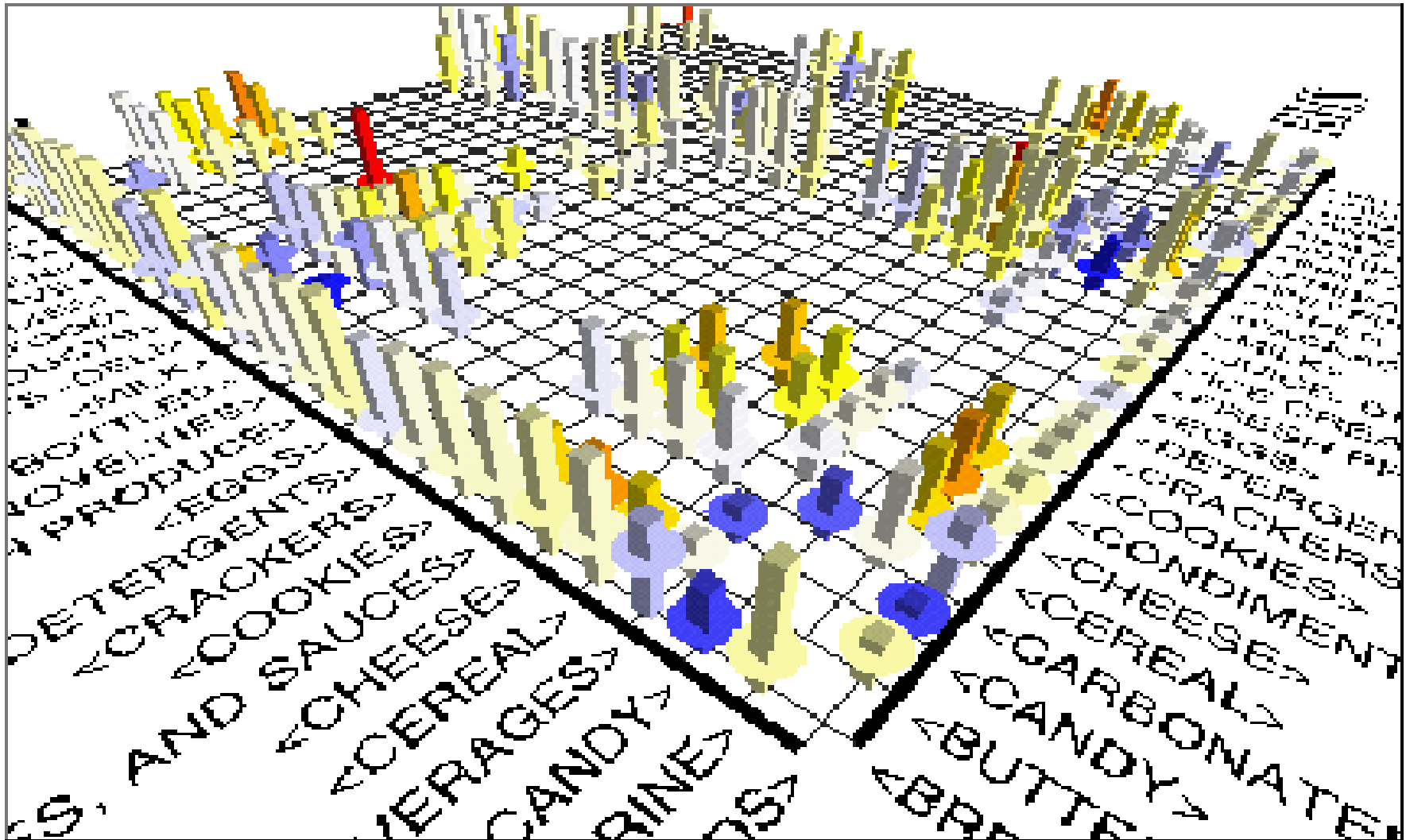
Data Mining Result Visualization

- Presentation of the **results or knowledge** obtained from data mining in visual forms
- Examples
 - Scatter plots and boxplots (obtained from descriptive data mining)
 - Decision trees
 - Association rules
 - Clusters
 - Outliers
 - Generalized rules

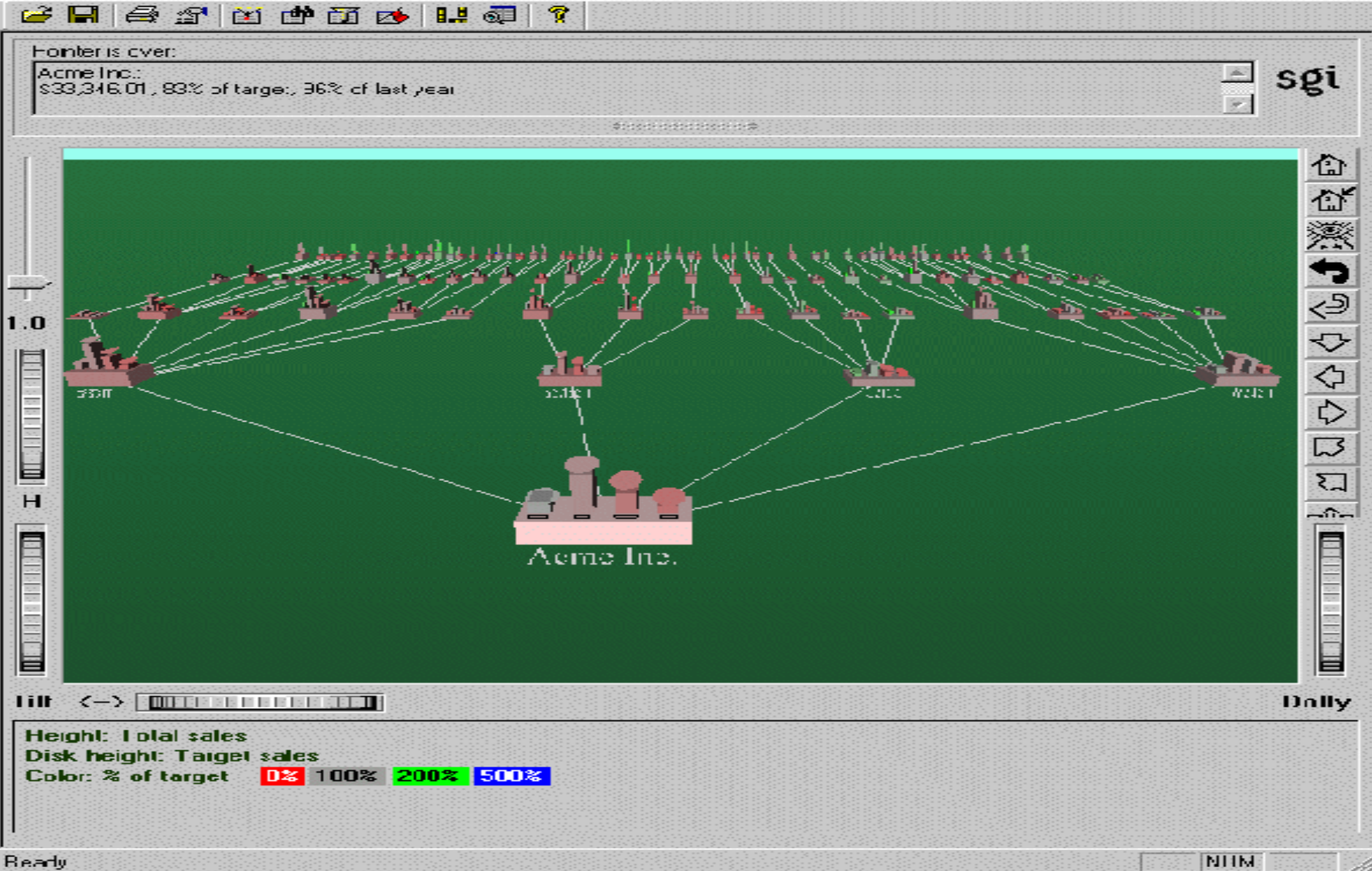
Visualization of Data Mining Results in SAS Enterprise Miner: Scatter Plots



Visualization of Association Rules in SGI/MineSet 3.0



Visualization of a Decision Tree in SGI/MineSet 3.0



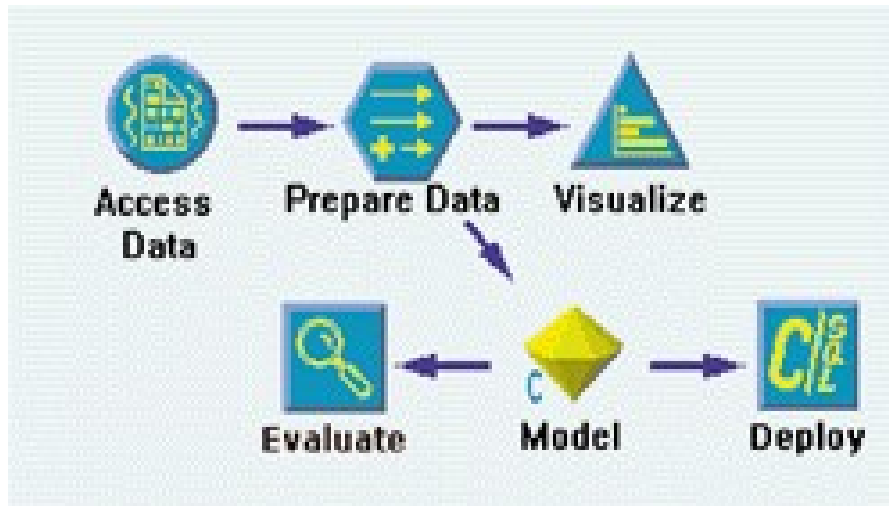
Visualization of Cluster Grouping in IBM Intelligent Miner



Data Mining Process Visualization

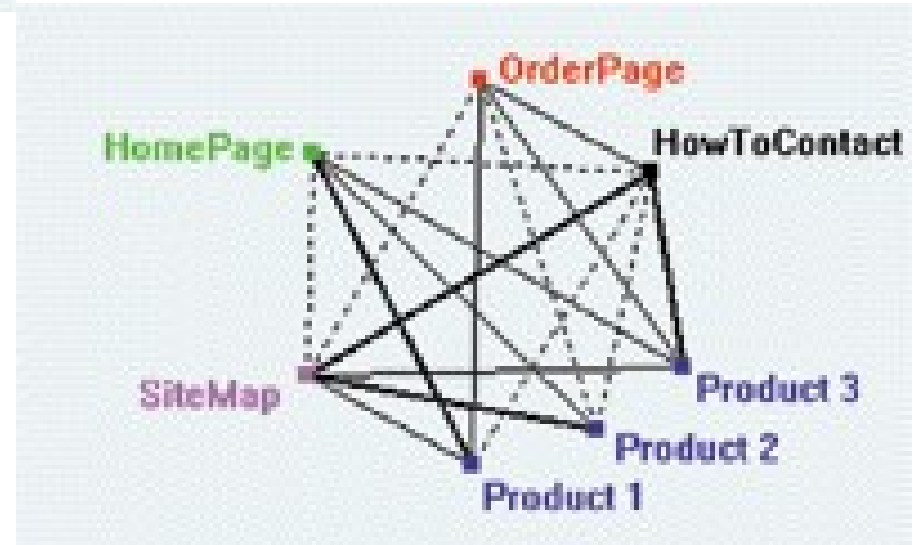
- Presentation of the various processes of data mining in visual forms so that users can see
 - Data extraction process
 - Where the data is extracted
 - How the data is cleaned, integrated, preprocessed, and mined
 - Method selected for data mining
 - Where the results are stored
 - How they may be viewed

Visualization of **Data Mining Processes** by Clementine



See your solution
discovery
process clearly

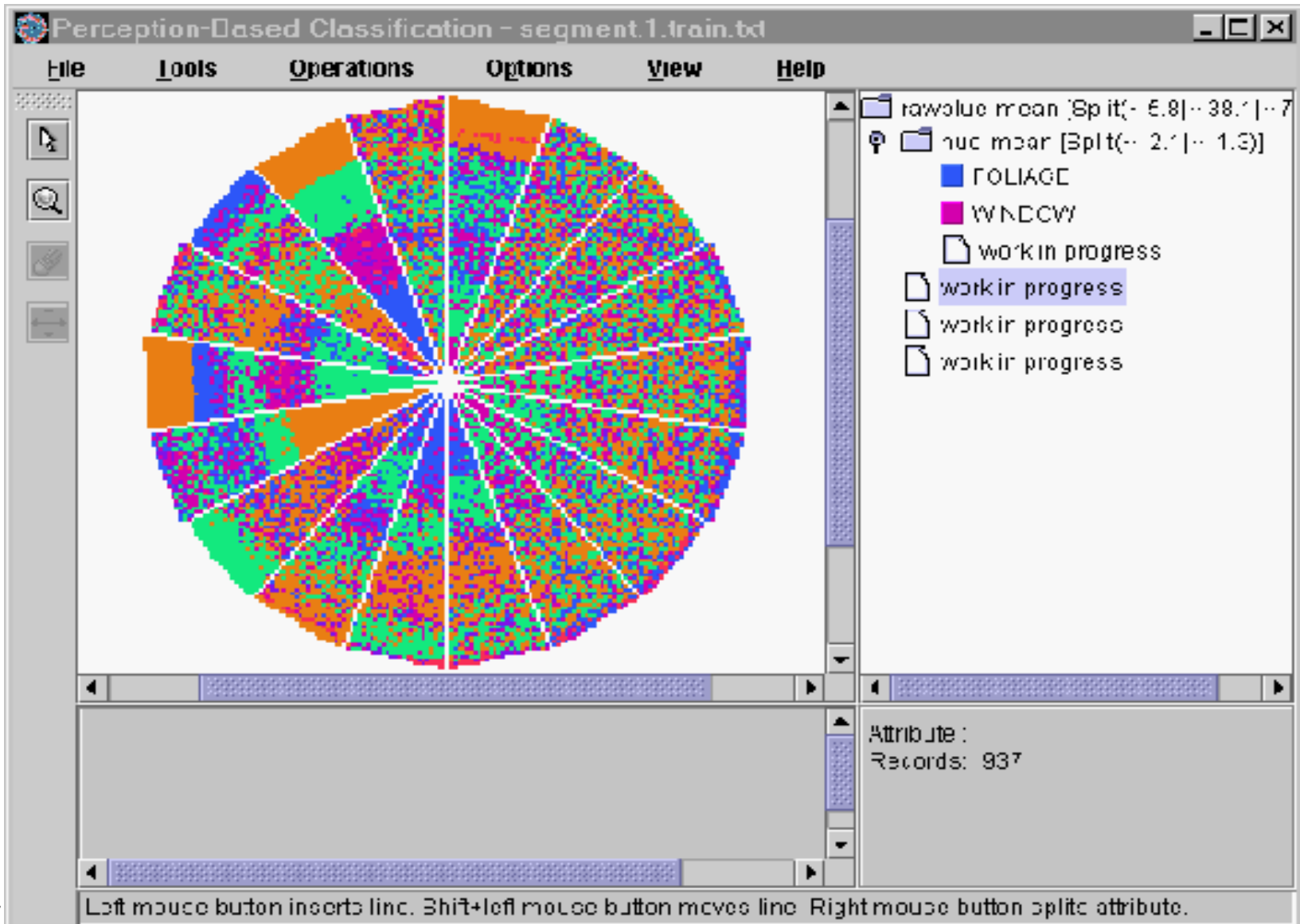
Understand
variations with
visualized data



Interactive Visual Data Mining

- Using visualization tools in the data mining process to help users make smart data mining decisions
- Example
 - Display the data distribution in a set of attributes using colored sectors or columns (depending on whether the whole space is represented by either a circle or a set of columns)
 - Use the display to which sector should first be selected for classification and where a good split point for this sector may be

Interactive Visual Mining by Perception-Based Classification (PBC)



Audio Data Mining

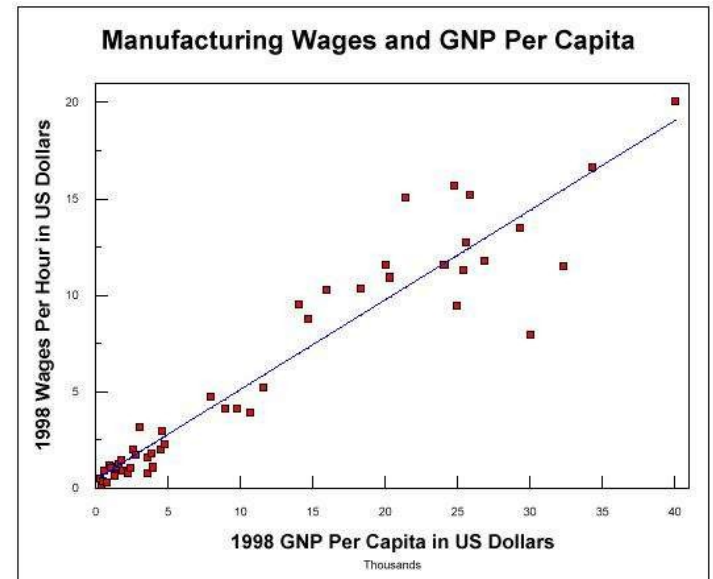
- Uses audio signals to indicate the patterns of data or the features of data mining results
- An interesting alternative to visual mining
- An inverse task of mining audio (such as music) databases which is to find patterns from audio data
- Visual data mining may disclose interesting patterns using graphical displays, but requires users to concentrate on watching patterns
- Instead, transform patterns into sound and music and listen to **pitches, rhythms, tune, and melody** in order to identify anything interesting or unusual

Scientific and Statistical Data Mining (1)

- There are many well-established statistical techniques for data analysis, particularly for numeric data
 - applied extensively to data from scientific experiments and data from economics and the social sciences

- **Regression**

- predict the value of a **response** (dependent) variable from one or more **predictor** (independent) variables where the variables are numeric
- forms of regression: linear, multiple, weighted, polynomial, nonparametric, and robust



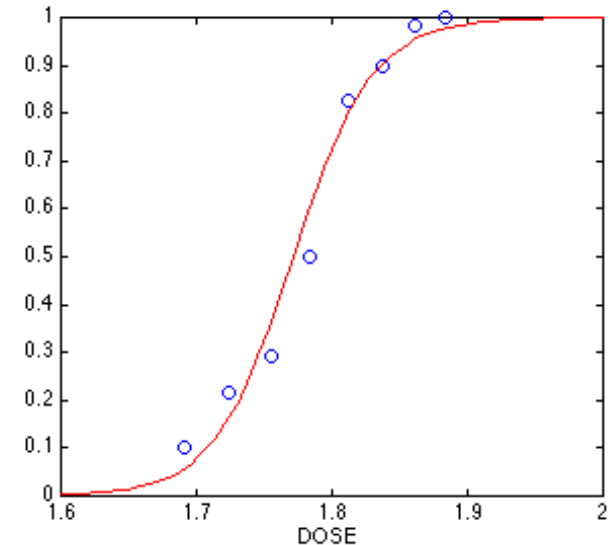
Scientific and Statistical Data Mining (2)

■ Generalized linear models

- allow a categorical response variable (or some transformation of it) to be related to a set of predictor variables
- similar to the modeling of a numeric response variable using linear regression
- include logistic regression and Poisson regression

■ Mixed-effect models

- For analyzing **grouped data**, i.e. data that can be classified according to one or more grouping variables
- Typically describe relationships between a response variable and some covariates in data grouped according to one or more factors



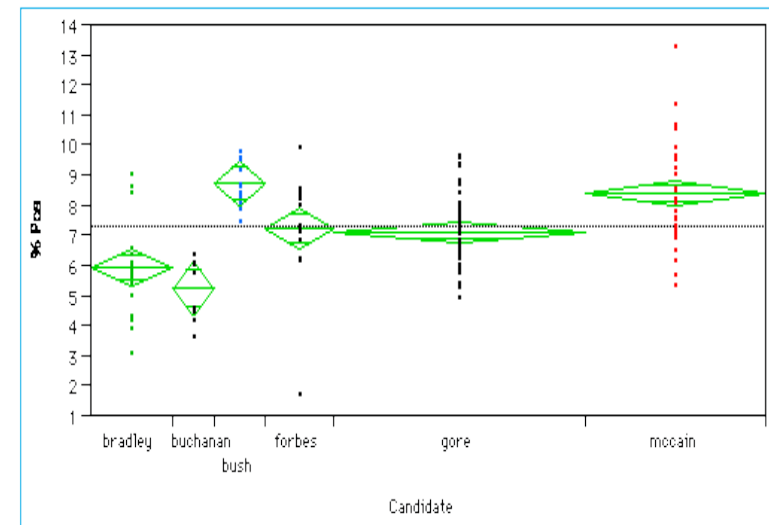
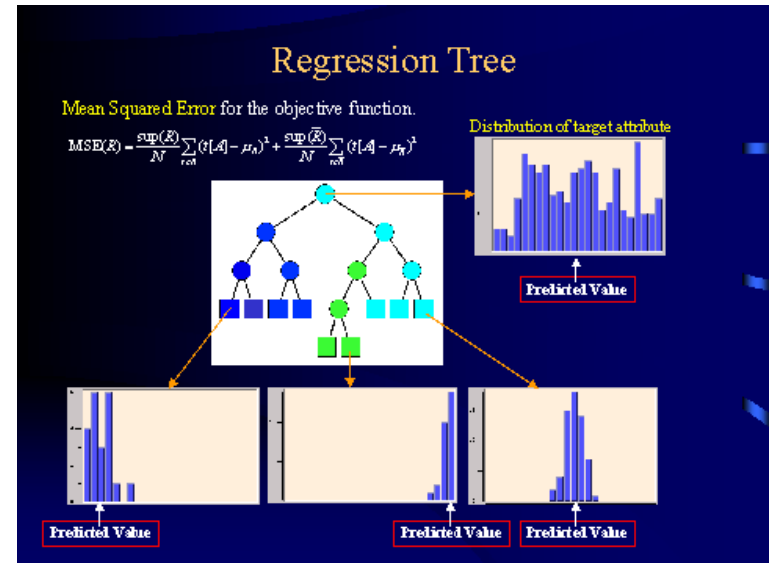
Scientific and Statistical Data Mining (3)

■ Regression trees

- Binary trees used for classification and prediction
- Similar to decision trees: Tests are performed at the internal nodes
- In a regression tree the mean of the objective attribute is computed and used as the predicted value

■ Analysis of variance

- Analyze experimental data for two or more populations described by a numeric response variable and one or more categorical variables (factors)



Scientific and Statistical Data Mining (4)

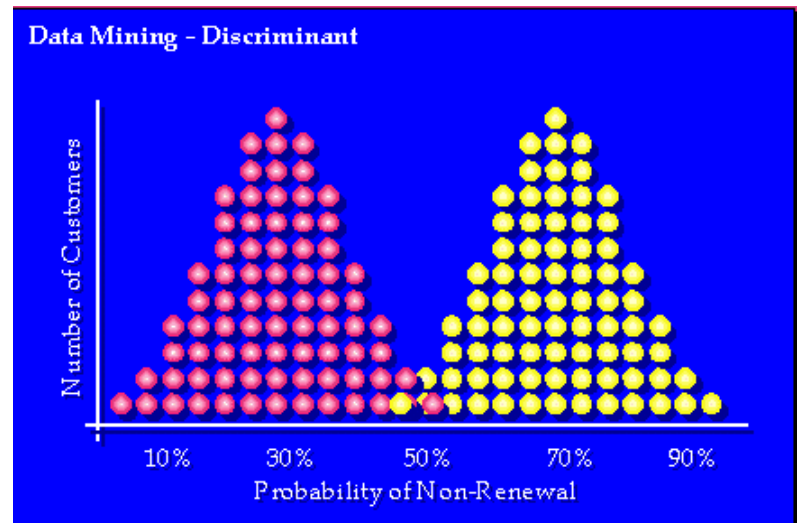
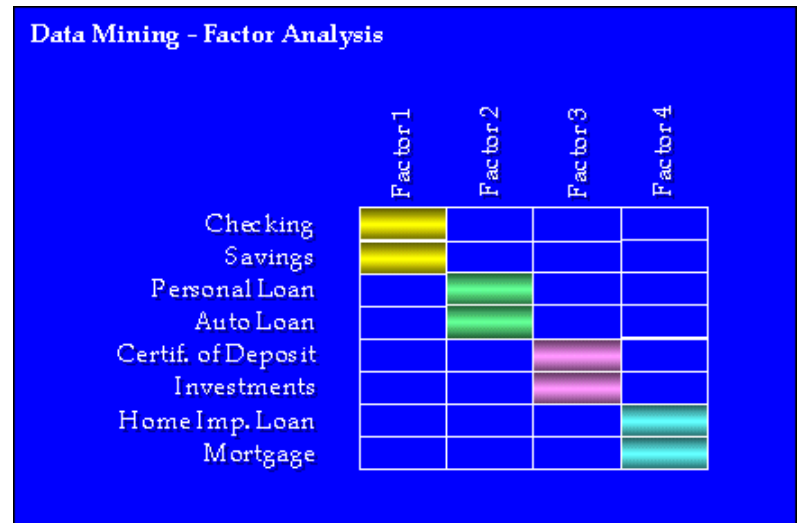
<http://www.spss.com/datamine/factor.htm>

Factor analysis

- determine which variables are combined to generate a given factor
- e.g., for many psychiatric data, one can indirectly measure other quantities (such as test scores) that reflect the factor of interest

Discriminant analysis

- predict a categorical response variable, commonly used in social science
- Attempts to determine several discriminant functions (linear combinations of the independent variables) that discriminate among the groups defined by the response variable

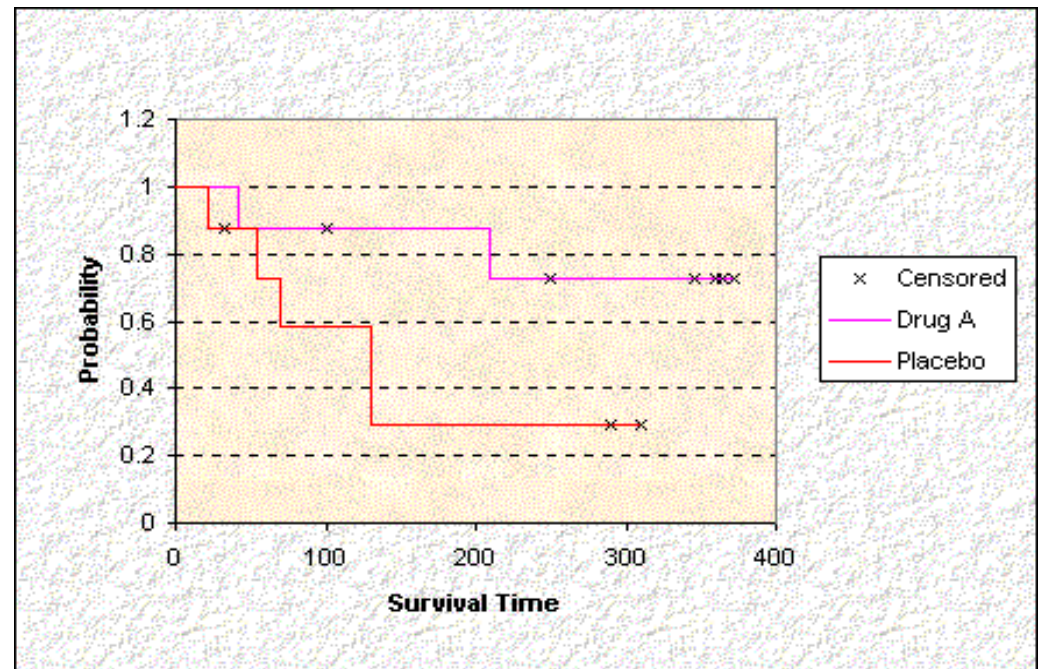


Scientific and Statistical Data Mining (5)

- **Time series:** many methods such as autoregression, ARIMA (Autoregressive integrated moving-average modeling), long memory time-series modeling
- **Quality control:** displays group summary charts

- **Survival analysis**

- predicts the probability that a patient undergoing a medical treatment would survive at least to time t (life span prediction)



Theoretical Foundations of Data Mining (1)

- **Data reduction**
 - The basis of data mining is to reduce the data representation
 - Trades accuracy for speed in response
- **Data compression**
 - The basis of data mining is to compress the given data by encoding in terms of bits, association rules, decision trees, clusters, etc.
- **Pattern discovery**
 - The basis of data mining is to discover patterns occurring in the database, such as associations, classification models, sequential patterns, etc.

Theoretical Foundations of Data Mining (2)

- Probability theory
 - The basis of data mining is to discover joint probability distributions of random variables
- Microeconomic view
 - A view of utility: the task of data mining is finding patterns that are interesting only to the extent in that they can be used in the decision-making process of some enterprise
- Inductive databases
 - Data mining is the problem of performing inductive logic on databases,
 - The task is to query the data and the theory (i.e., patterns) of the database
 - Popular among many researchers in database systems

Data Mining and Intelligent Query Answering

- A general framework for the integration of data mining and intelligent query answering
 - **Data query:** finds concrete data stored in a database; returns exactly what is being asked
 - **Knowledge query:** finds rules, patterns, and other kinds of knowledge in a database
 - Intelligent (or cooperative) query answering: analyzes the intent of the query and provides generalized, neighborhood or associated information relevant to the query

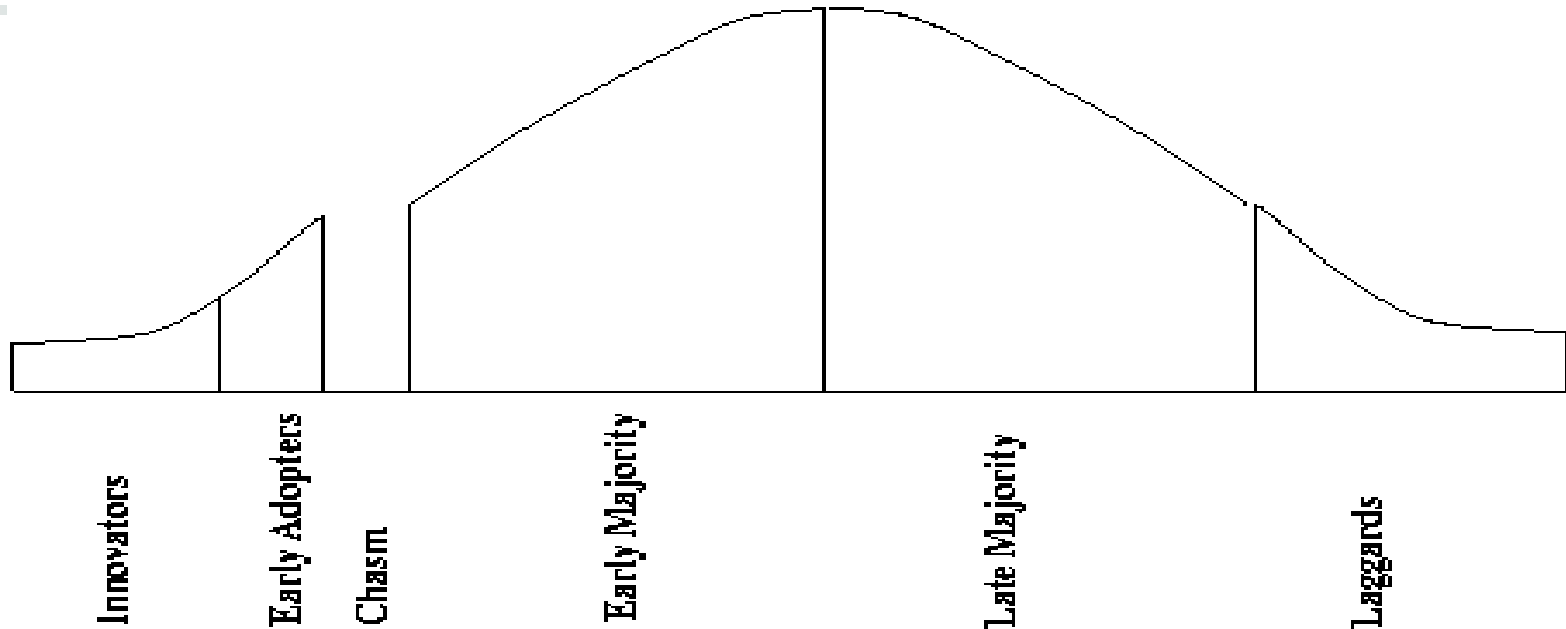
Chapter 10: Applications and Trends in Data Mining

- Data mining applications
- Data mining system products and research prototypes
- Additional themes on data mining
- **Social impacts of data mining**
- Trends in data mining
- Summary

Is Data Mining a Hype or Will It Be Persistent?

- Data mining is a technology
- Technological life cycle
 - Innovators
 - Early adopters
 - Chasm
 - Early majority
 - Late majority
 - Laggards

Life Cycle of Technology Adoption



- Data mining is at Chasm!?
 - Existing data mining systems are too **generic**
 - Need **business-specific** data mining solutions and smooth **integration** of business logic with data mining functions

Data Mining: Merely Managers' Business or Everyone's?

- Data mining will surely be an important tool for managers' decision making
 - Bill Gates: "Business @ the speed of thought"
- The amount of the available data is increasing, and data mining systems will be more affordable
- Multiple personal uses
 - Mine your family's medical history to identify genetically-related medical conditions
 - Mine the records of the companies you deal with
 - Mine data on stocks and company performance, etc.
- Invisible data mining
 - Build data mining functions into many intelligent tools

Social Impacts: Threat to Privacy and Data Security?

- Is data mining a threat to privacy and data security?
 - “Big Brother”, “Big Banker”, and “Big Business” are carefully watching you
 - Profiling information is collected every time
 - credit card, debit card, supermarket loyalty card, or frequent flyer card, or apply for any of the above
 - You surf the Web, rent a video, fill out a contest entry form,
 - You pay for prescription drugs, or present you medical care number when visiting the doctor
 - Collection of personal data may be beneficial for companies and consumers, there is also **potential for misuse**
 - **Medical Records, Employee Evaluations, Etc.**

Protect Privacy and Data Security

- Fair information practices
 - International guidelines for data privacy protection
 - Cover aspects relating to data collection, purpose, use, quality, openness, individual participation, and accountability
 - **Purpose specification and use limitation**
 - **Openness**: Individuals have the right to know what information is collected about them, who has access to the data, and how the data are being used
- Develop and use data security-enhancing techniques
 - Blind signatures
 - Biometric encryption
 - Anonymous databases

Chapter 10: Applications and Trends in Data Mining

- Data mining applications
- Data mining system products and research prototypes
- Additional themes on data mining
- Social impact of data mining
- Trends in data mining
- Summary

Trends in Data Mining (1)

- Application exploration
 - development of application-specific data mining system
 - Invisible data mining (mining as built-in function)
- Scalable data mining methods
 - Constraint-based mining: use of constraints to guide data mining systems in their search for interesting patterns
- Integration of data mining with database systems, data warehouse systems, and Web database systems
- Invisible data mining

Trends in Data Mining (2)

- Standardization of data mining language
 - A standard will facilitate systematic development, improve interoperability, and promote the education and use of data mining systems in industry and society
- Visual data mining
- New methods for mining complex types of data
 - More research is required towards the integration of data mining methods with existing data analysis techniques for the complex types of data
- Web mining
- Privacy protection and information security in data mining

Chapter 10: Applications and Trends in Data Mining

- Data mining applications
- Data mining system products and research prototypes
- Additional themes on data mining
- Social impact of data mining
- Trends in data mining
- **Summary**

Summary

- **Domain-specific applications** include biomedicine (DNA), finance, retail and telecommunication data mining
- There exist some **data mining systems** and it is important to know their power and limitations
- **Visual data mining** include data visualization, mining result visualization, mining process visualization and interactive visual mining
- There are many other **scientific and statistical data mining methods** developed but not covered in this book
- Also, it is important to study **theoretical foundations** of data mining
- **Intelligent query answering** can be integrated with mining
- It is important to watch **privacy and security** issues in data mining

References (1)

- M. Ankerst, C. Elsen, M. Ester, and H.-P. Kriegel. Visual classification: An interactive approach to decision tree construction. KDD'99, San Diego, CA, Aug. 1999.
- P. Baldi and S. Brunak. Bioinformatics: The Machine Learning Approach. MIT Press, 1998.
- S. Benninga and B. Czaczkes. Financial Modeling. MIT Press, 1997.
- L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth International Group, 1984.
- M. Berthold and D. J. Hand. Intelligent Data Analysis: An Introduction. Springer-Verlag, 1999.
- M. J. A. Berry and G. Linoff. Mastering Data Mining: The Art and Science of Customer Relationship Management. John Wiley & Sons, 1999.
- A. Baxevanis and B. F. F. Ouellette. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. John Wiley & Sons, 1998.
- Q. Chen, M. Hsu, and U. Dayal. A data-warehouse/OLAP framework for scalable telecommunication tandem traffic analysis. ICDE'00, San Diego, CA, Feb. 2000.
- W. Cleveland. Visualizing Data. Hobart Press, Summit NJ, 1993.
- S. Chakrabarti, S. Sarawagi, and B. Dom. Mining surprising patterns using temporal description length. VLDB'98, New York, NY, Aug. 1998.

References (2)

- J. L. Devore. Probability and Statistics for Engineering and the Science, 4th ed. Duxbury Press, 1995.
- A. J. Dobson. An Introduction to Generalized Linear Models. Chapman and Hall, 1990.
- B. Gates. Business @ the Speed of Thought. New York: Warner Books, 1999.
- M. Goebel and L. Gruenwald. A survey of data mining and knowledge discovery software tools. SIGKDD Explorations, 1:20-33, 1999.
- D. Gusfield. Algorithms on Strings, Trees and Sequences, Computer Science and Computation Biology. Cambridge University Press, New York, 1997.
- J. Han, Y. Huang, N. Cercone, and Y. Fu. Intelligent query answering by knowledge discovery techniques. IEEE Trans. Knowledge and Data Engineering, 8:373-390, 1996.
- R. C. Higgins. Analysis for Financial Management. Irwin/McGraw-Hill, 1997.
- C. H. Huberty. Applied Discriminant Analysis. New York: John Wiley & Sons, 1994.
- T. Imielinski and H. Mannila. A database perspective on knowledge discovery. Communications of ACM, 39:58-64, 1996.
- D. A. Keim and H.-P. Kriegel. VisDB: Database exploration using multidimensional visualization. Computer Graphics and Applications, pages 40-49, Sept. 94.

References (3)

- J. M. Kleinberg, C. Papadimitriou, and P. Raghavan. A microeconomic view of data mining. *Data Mining and Knowledge Discovery*, 2:311-324, 1998.
- H. Mannila. Methods and problems in data mining. ICDT'99 Delphi, Greece, Jan. 1997.
- R. Mattison. *Data Warehousing and Data Mining for Telecommunications*. Artech House, 1997.
- R. G. Miller. *Survival Analysis*. New York: Wiley, 1981.
- G. A. Moore. *Crossing the Chasm: Marketing and Selling High-Tech Products to Mainstream Customers*. Harperbusiness, 1999.
- R. H. Shumway. *Applied Statistical Time Series Analysis*. Prentice Hall, 1988.
- E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, 1983.
- E. R. Tufte. *Envisioning Information*. Graphics Press, Cheshire, CT, 1990.
- E. R. Tufte. *Visual Explanations : Images and Quantities, Evidence and Narrative*. Graphics Press, Cheshire, CT, 1997.
- M. S. Waterman. *Introduction to Computational Biology: Maps, Sequences, and Genomes (Interdisciplinary Statistics)*. CRC Press, 1995.

www.cs.uiuc.edu/~hanj



Thank you !!!