Chapter 8
Application and Trends in Data Warehousing and Data Mining

Data Mining Systems Products and Research Prototypes

As a young discipline, data mining has a relatively short history and are constantly evolving-new data mining systems appear on the market every year; new functions, features, and visualization tools are added to existing systems on a constant basis; and efforts toward the standardization of data mining language have only just begun.

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?

- ❖ 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 Database 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

- Scalability
  - Row (or database size) scalability
  - Column (or dimension) scalability
  - Curse of dimensionality: it is much more challenging to make a system column scalable that 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

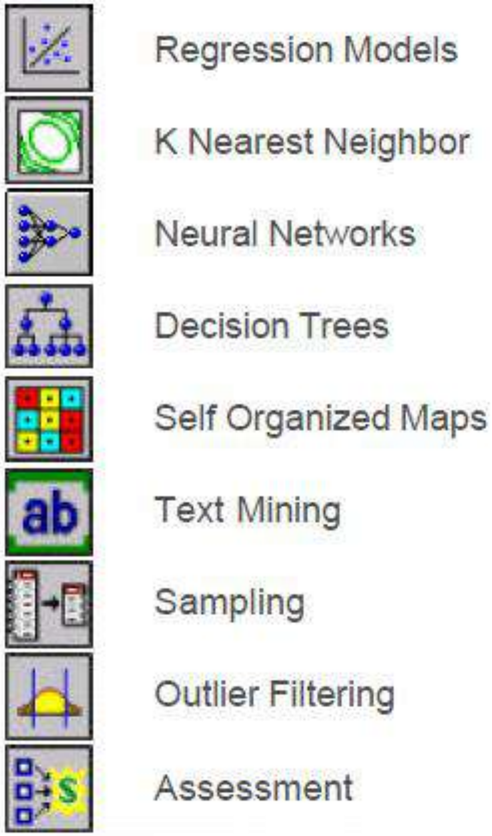
Examples of Data Mining Systems

- ❖ Microsoft SQL Server 2005
  - – Integrate DB and OLAP with mining
  - – Support OLEDB for DM standard

- ❖ IBM Intelligent Miner
  - – Intelligent Miner is an IBM data-mining product
  - – 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
  - – SAS Institute Inc. developed Enterprise Miner
  - – A variety of statistical analysis tools
  - – Data warehouse tools and multiple data mining algorithms

Enterprise Miner Capabilities



Regression Models

K Nearest Neighbor

Neural Networks

Decision Trees

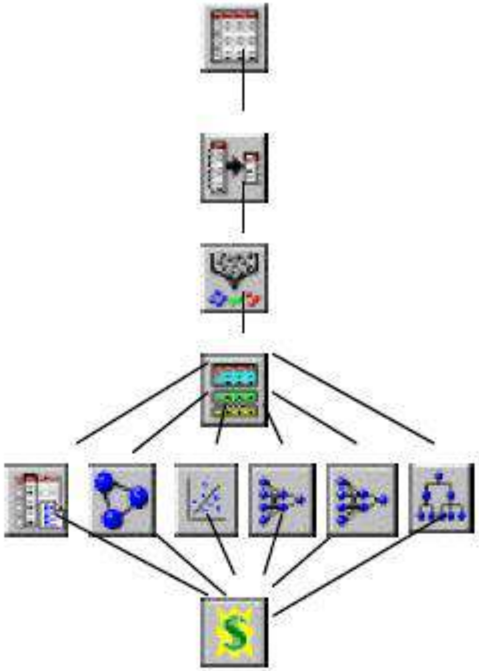
Self Organized Maps

Text Mining

Sampling

Outlier Filtering

Assessment



❖ SGI MineSet

- Silicon Graphics Inc. (SGI) developed MineSet
- Multiple data mining algorithms and advanced statistics
- Advanced visualization tools

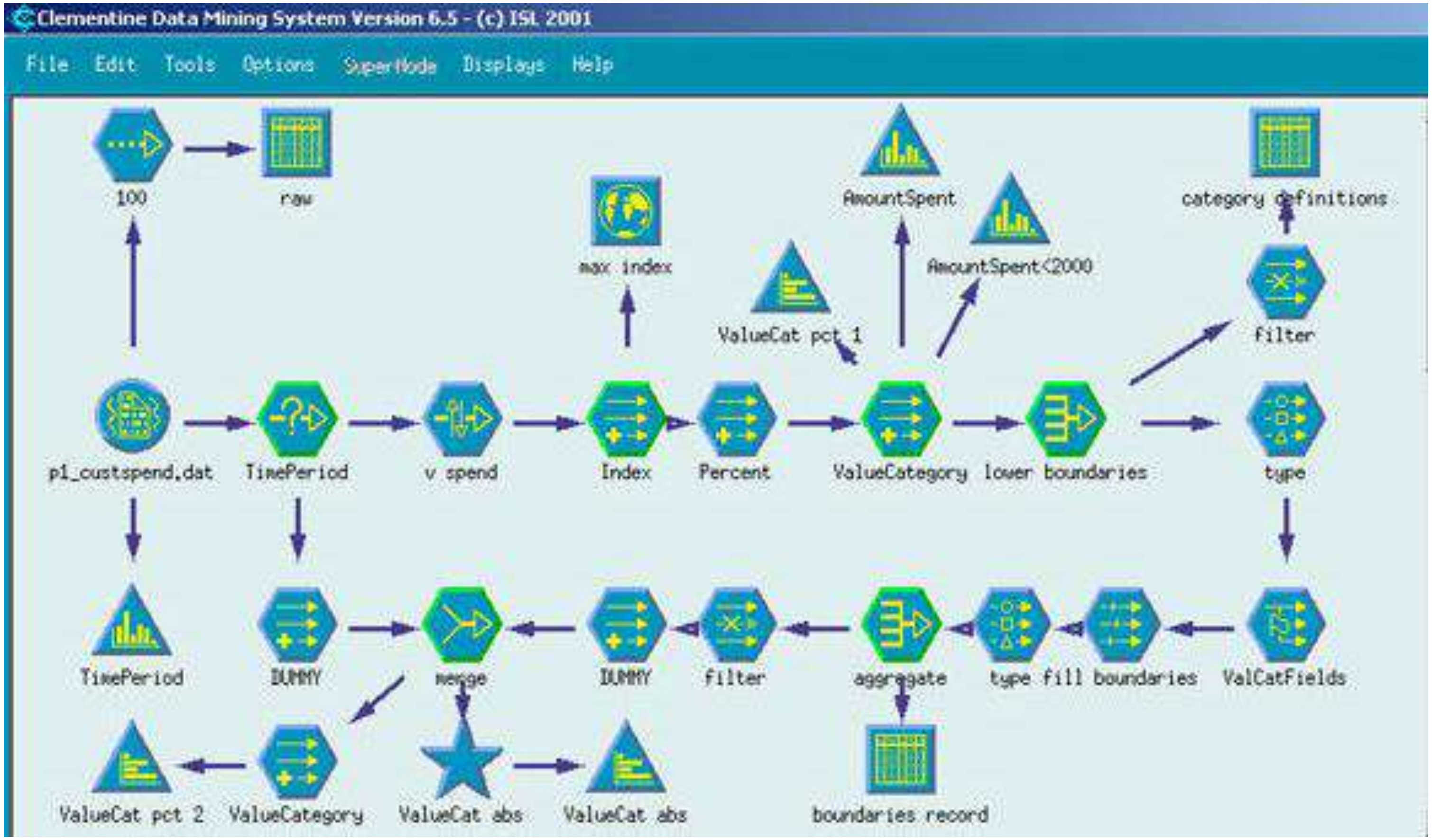
❖ DBMiner

- DBMiner Technology Inc developed DBMiner.
- It provides multiple data mining algorithms including discovery-driven OLAP analysis, association, classification, and clustering

❖ SPSS Clementine

- Integral Solutions Ltd. (ISL) developed Clementine
- Clementine has been acquired by SPSS Inc.
- An integrated data mining development environment for end-users and developers
- Multiple data mining algorithms and visualization tools including rule induction, neural nets, classification, and visualization tools

SPSS Clementine



Theoretical Foundations of Data Mining

- 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.

❖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

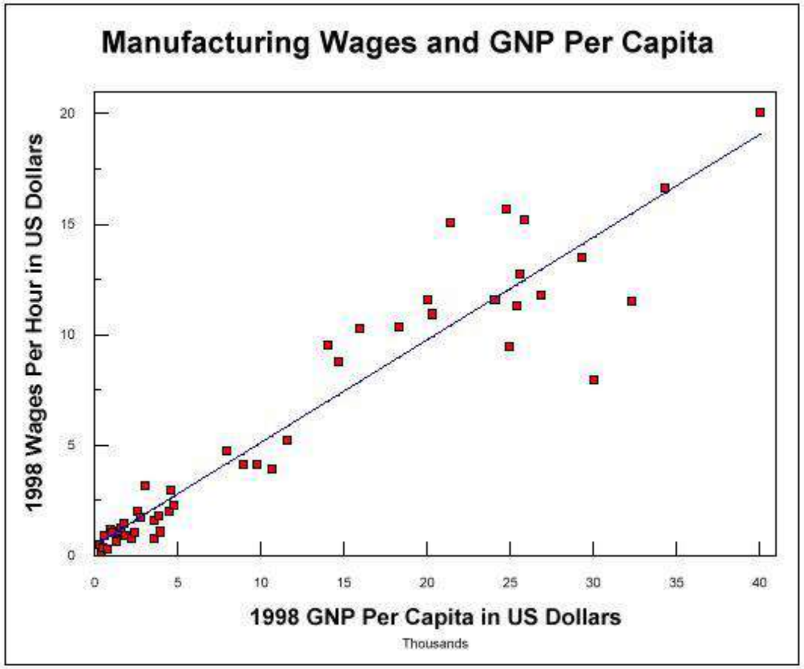
– Popular among many researchers in database systems

Statistical Data Mining

- 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

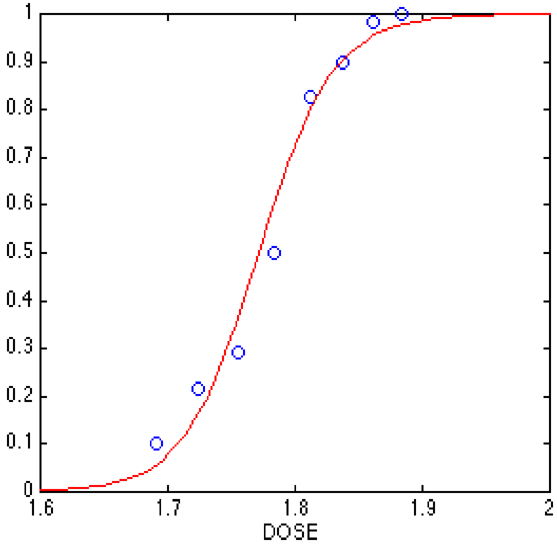
Regression

- 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



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

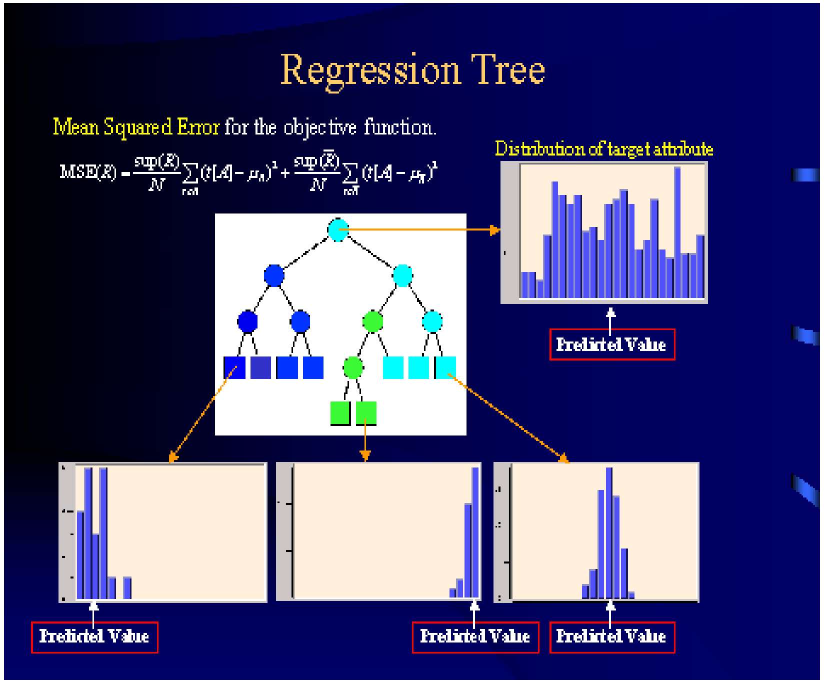


Mixed-effect models

- 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

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



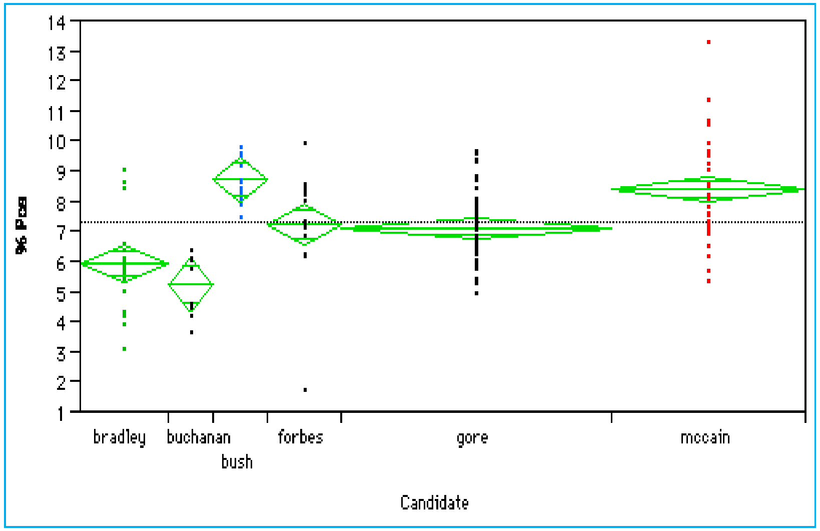
Regression Tree

Mean Squared Error for the objective function.

$$MSE(R) = \frac{sup(R)}{N}\sum_{t \in R}(t[A] - \mu_R)^2 + \frac{sup(\bar{R})}{N}\sum_{t \in \bar{R}}(t[A] - \mu_{\bar{R}})^2$$

Analysis of variance

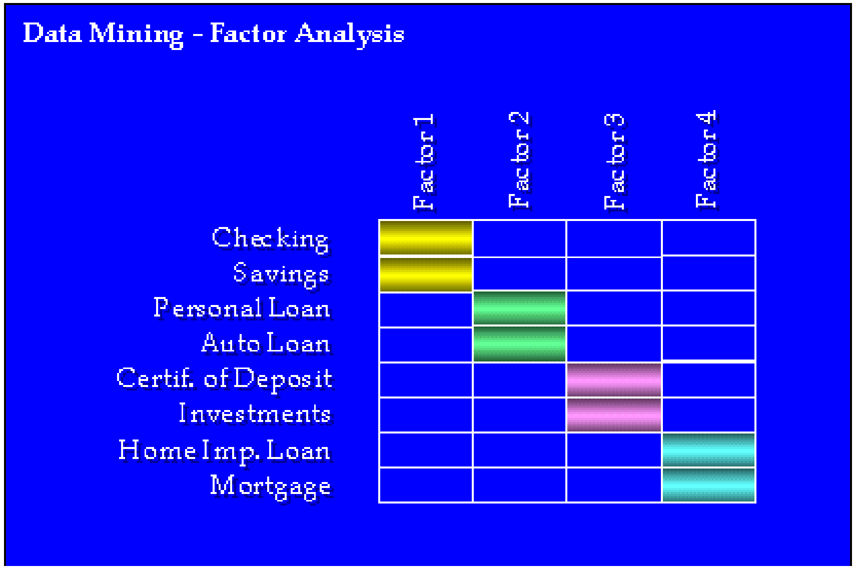
- Analyze experimental data for two or more populations described by a numeric response variable and one or more categorical variables (factors)



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

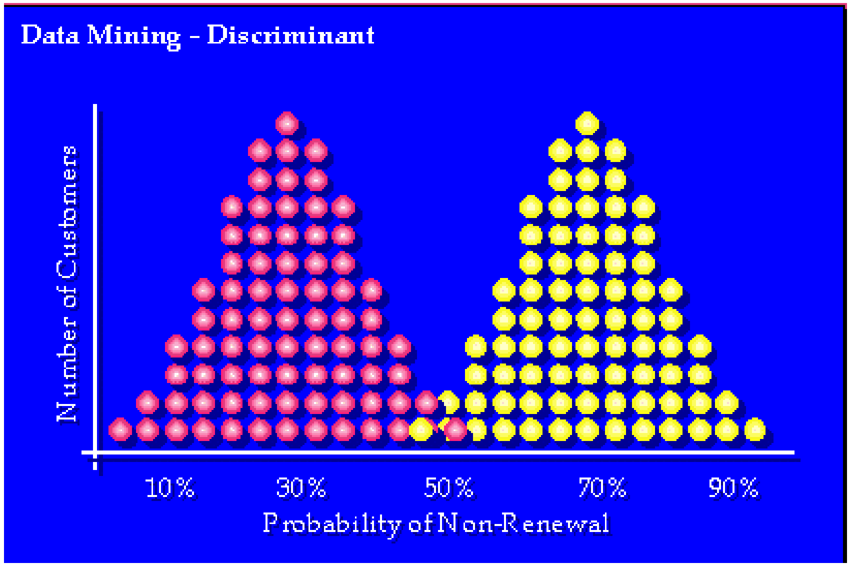


Data Mining - Factor Analysis

	Factor 1	Factor 2	Factor 3	Factor 4
Checking	🟨			
Savings	🟨			
Personal Loan		🟩		
Auto Loan		🟩		
Certif. of Deposit			🟪	
Investments			🟪	
Home Imp. Loan				🟦
Mortgage				🟦

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



Data Mining - Discriminant

Time series:

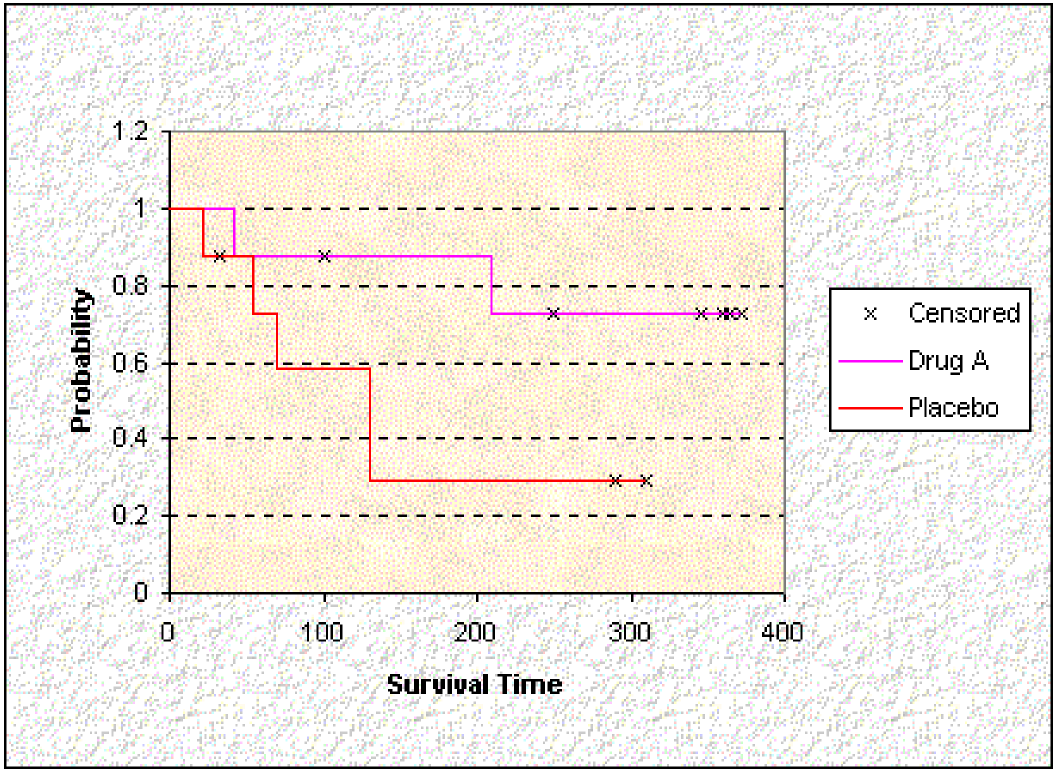
Many methods such as auto regression, ARIMA (Autoregressive integrated moving-average modeling), long memory time-series modeling

Quality control:

Displays group summary charts

Survival analysis

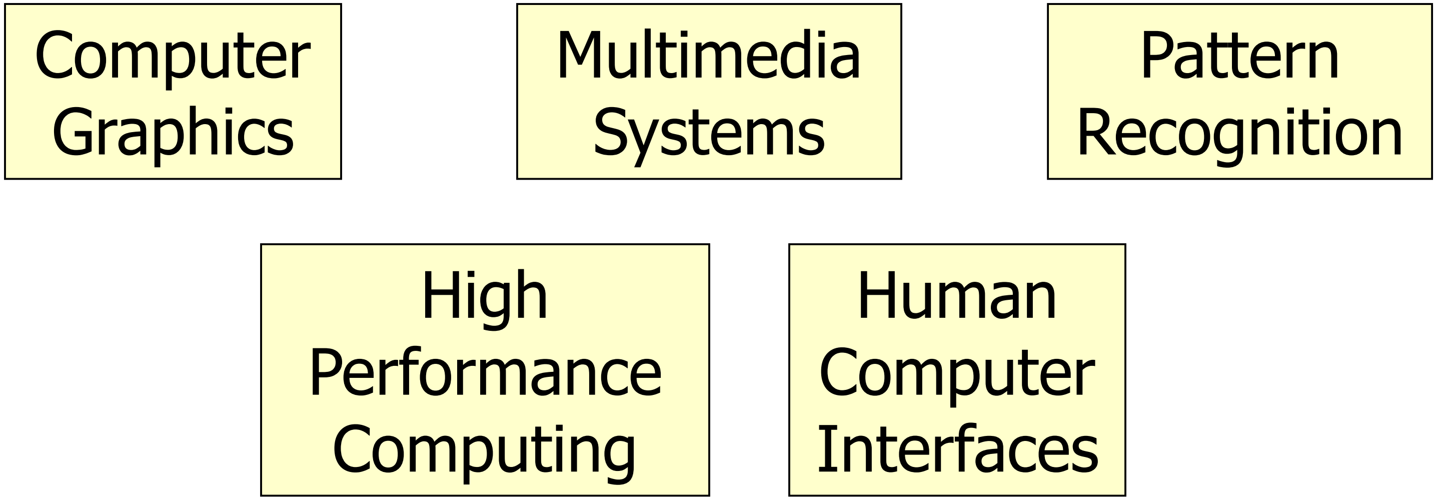
Predicts the probability that a patient undergoing a medical treatment would survive at least to time *t* (life span prediction)



Visual and Audio 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

Pattern Recognition

High Performance Computing

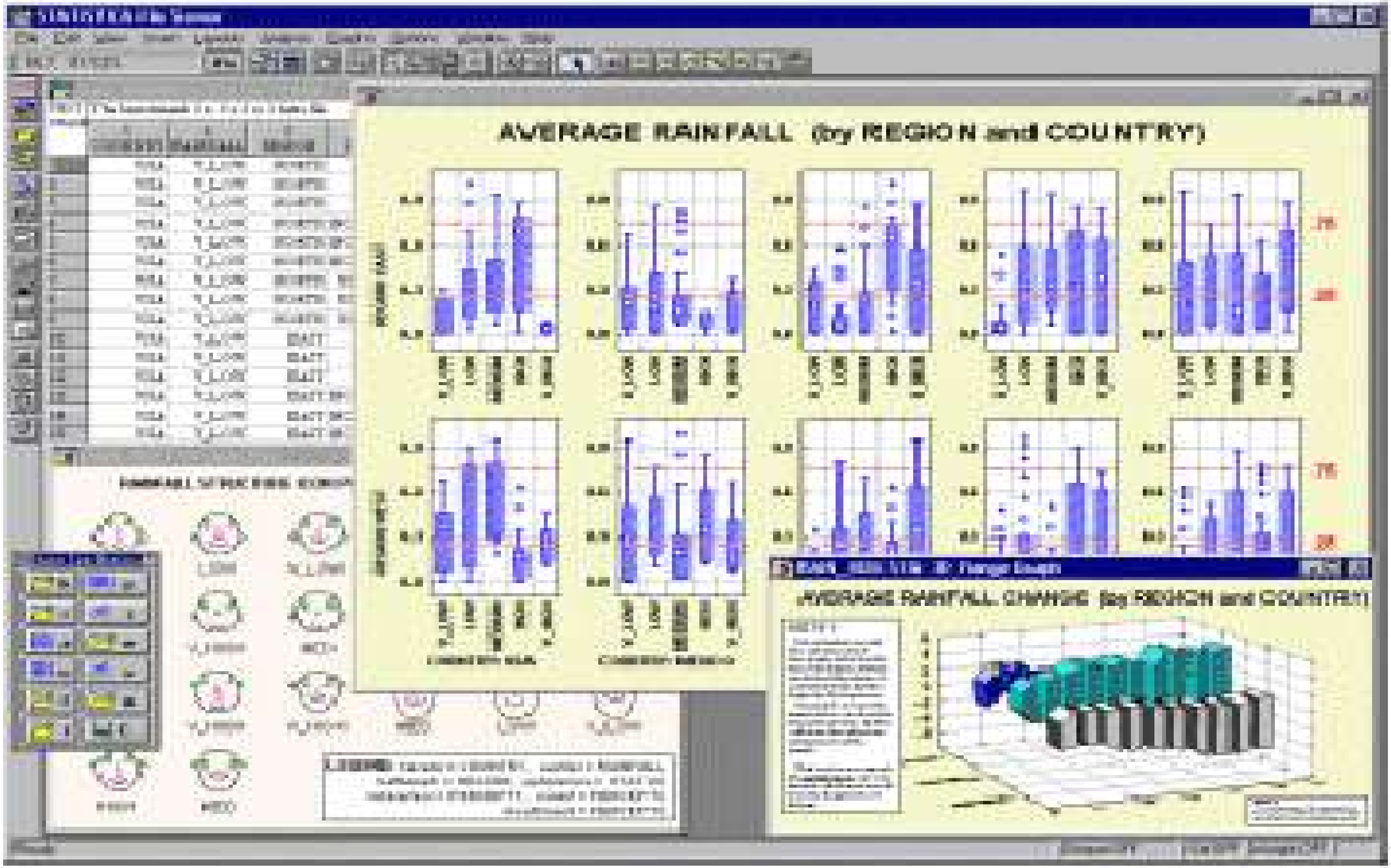
Human Computer Interfaces

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

- 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 granularity or abstraction
    - as different combinations of attributes or dimensions
  - Data can be presented in various visual forms

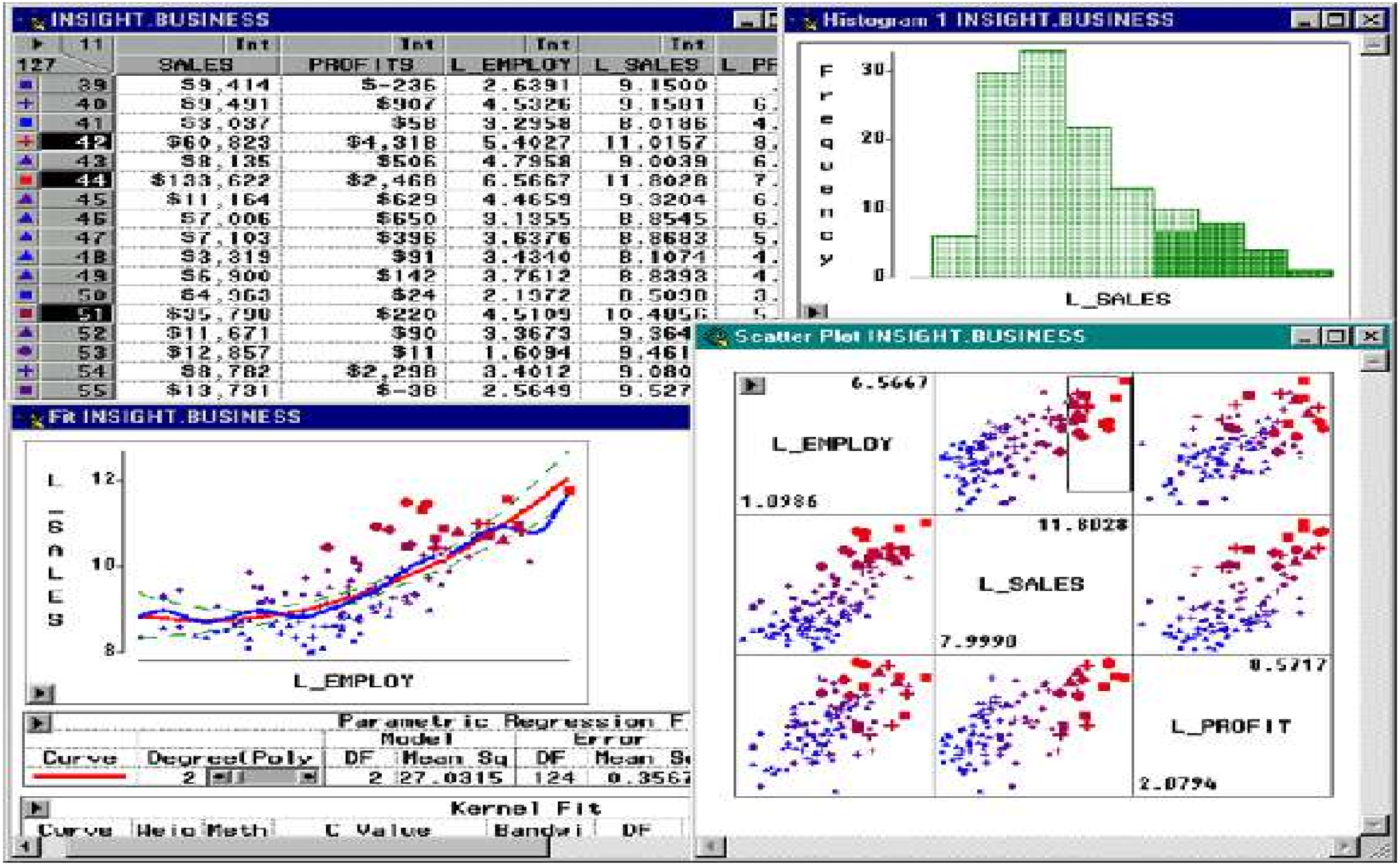
Boxplots from Statsoft: Multiple Variable Combinations


AVERAGE RAINFALL (by REGION and COUNTRY)
AVERAGE RAINFALL CHANGE (by REGION and COUNTRY)

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



INSIGHT.BUSINESS

▶ 11 127	Int SALES	Int PROFITS	Int L_EMPLOY	Int L_SALES	 L_PF
■ 39	$9,414	$-236	2.6391	9.1500	.
+ 40	$9,491	$907	4.5326	9.1581	6.
■ 41	$3,037	$58	3.2958	8.0186	4.
+ 42	$60,823	$4,318	5.4027	11.0157	8.
▲ 43	$8,135	$506	4.7958	9.0039	6.
■ 44	$133,622	$2,468	6.5667	11.8028	7.
▲ 45	$11,164	$629	4.4659	9.3204	6.
▲ 46	$7,006	$650	3.1355	8.8545	6.
▲ 47	$7,103	$396	3.6376	8.8683	5.
▲ 48	$3,319	$91	3.4340	8.1074	4.
▲ 49	$6,900	$142	3.7612	8.8393	4.
■ 50	$4,963	$24	2.1972	8.5090	3.
■ 51	$35,790	$220	4.5109	10.4856	5.
▲ 52	$11,671	$90	3.3673	9.364	
● 53	$12,857	$11	1.6094	9.461	
+ 54	$8,782	$2,298	3.4012	9.080	
■ 55	$13,731	$-38	2.5649	9.527	

Histogram 1 INSIGHT.BUSINESS

Scatter Plot INSIGHT.BUSINESS

Fit INSIGHT.BUSINESS

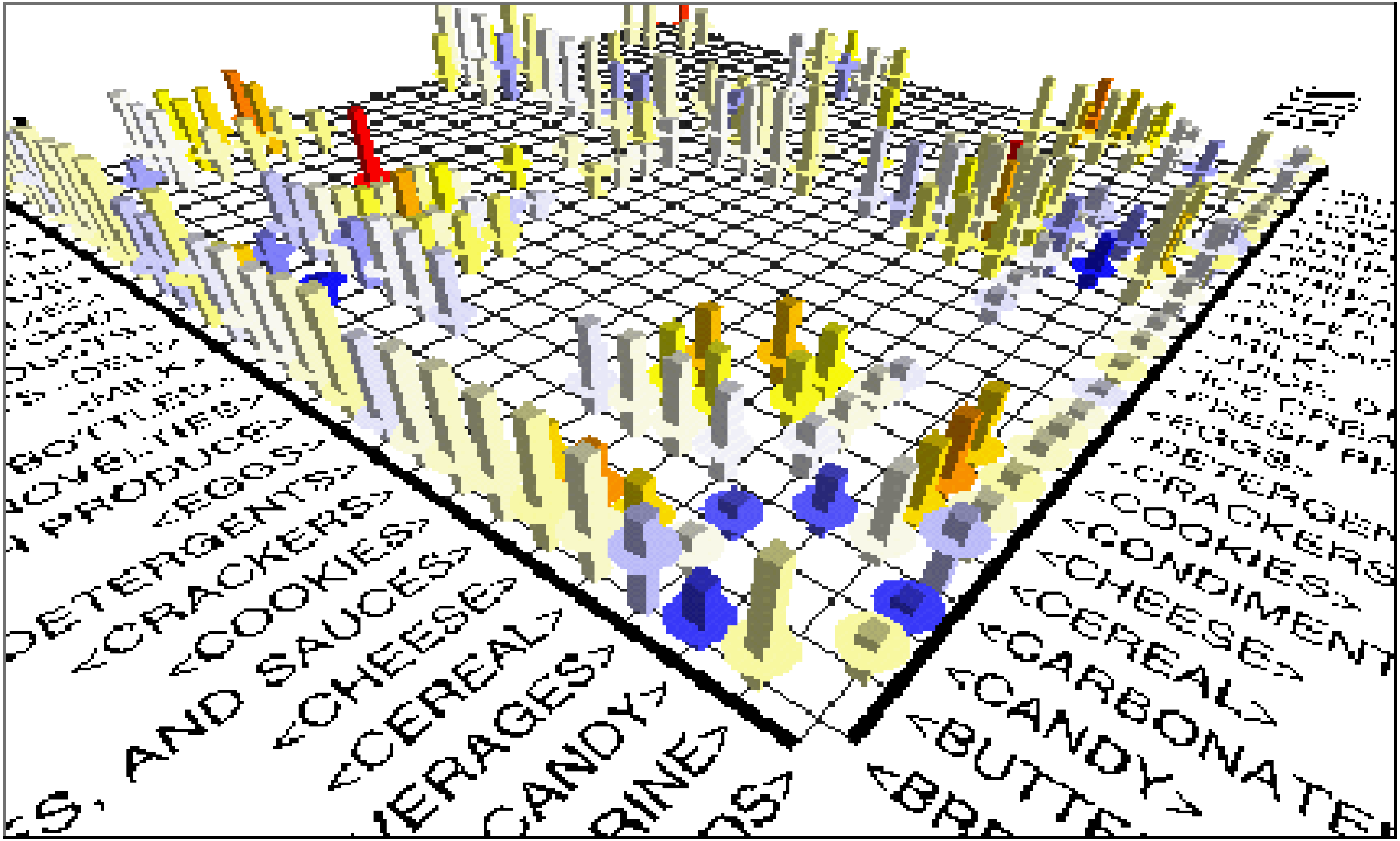
Parametric Regression F

		Model		Error	
Curve	Degree(Poly)	DF	Mean Sq	DF	Mean Sq
🟥	2	2	27.0315	124	0.3567

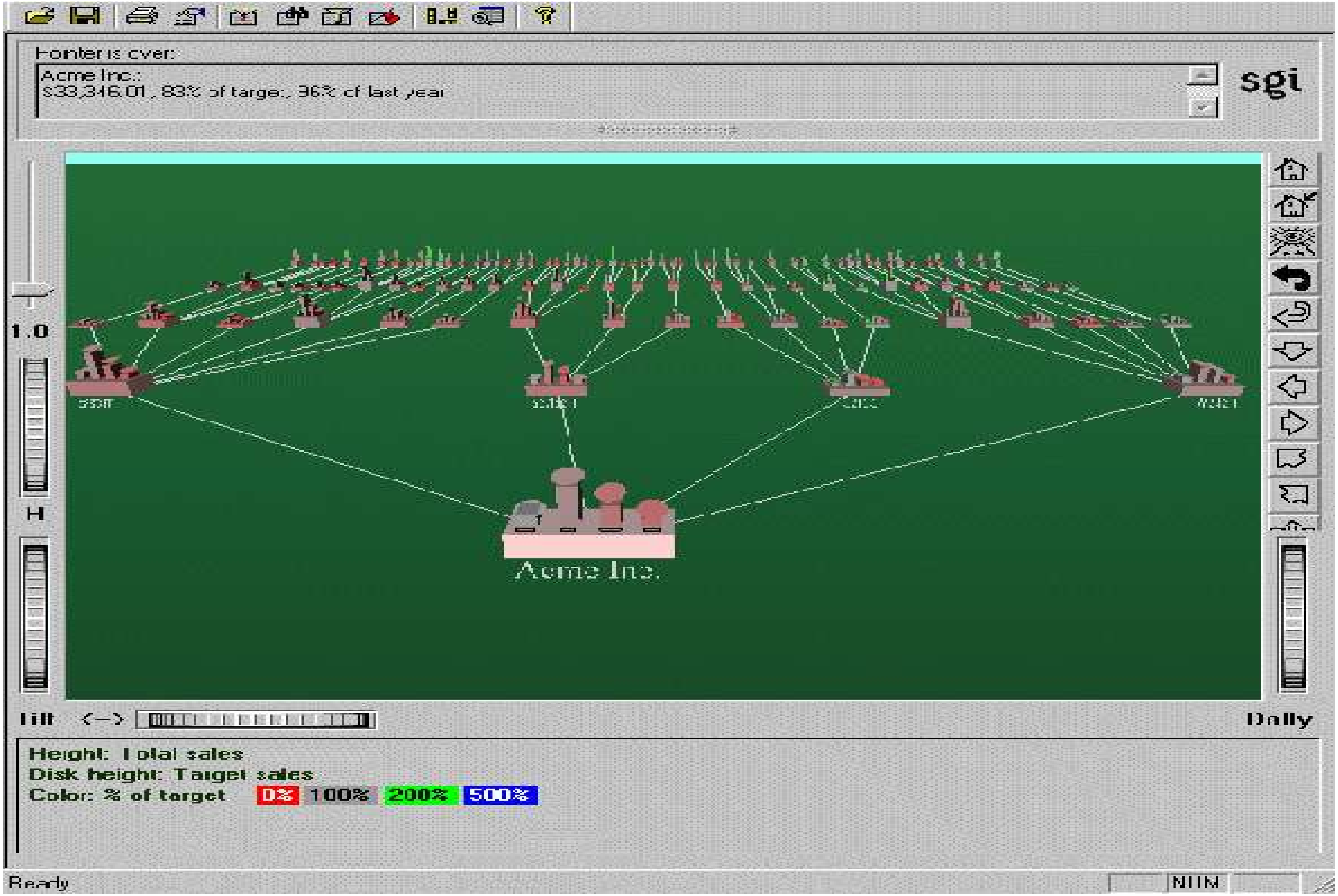
Kernel Fit

Curve	Weig Meth	C Value	Bandwi	DF

Visualization of Association Rules in SGI/MineSet 3.0



Visualization of a Decision Tree in SGI/MineSet 3.0



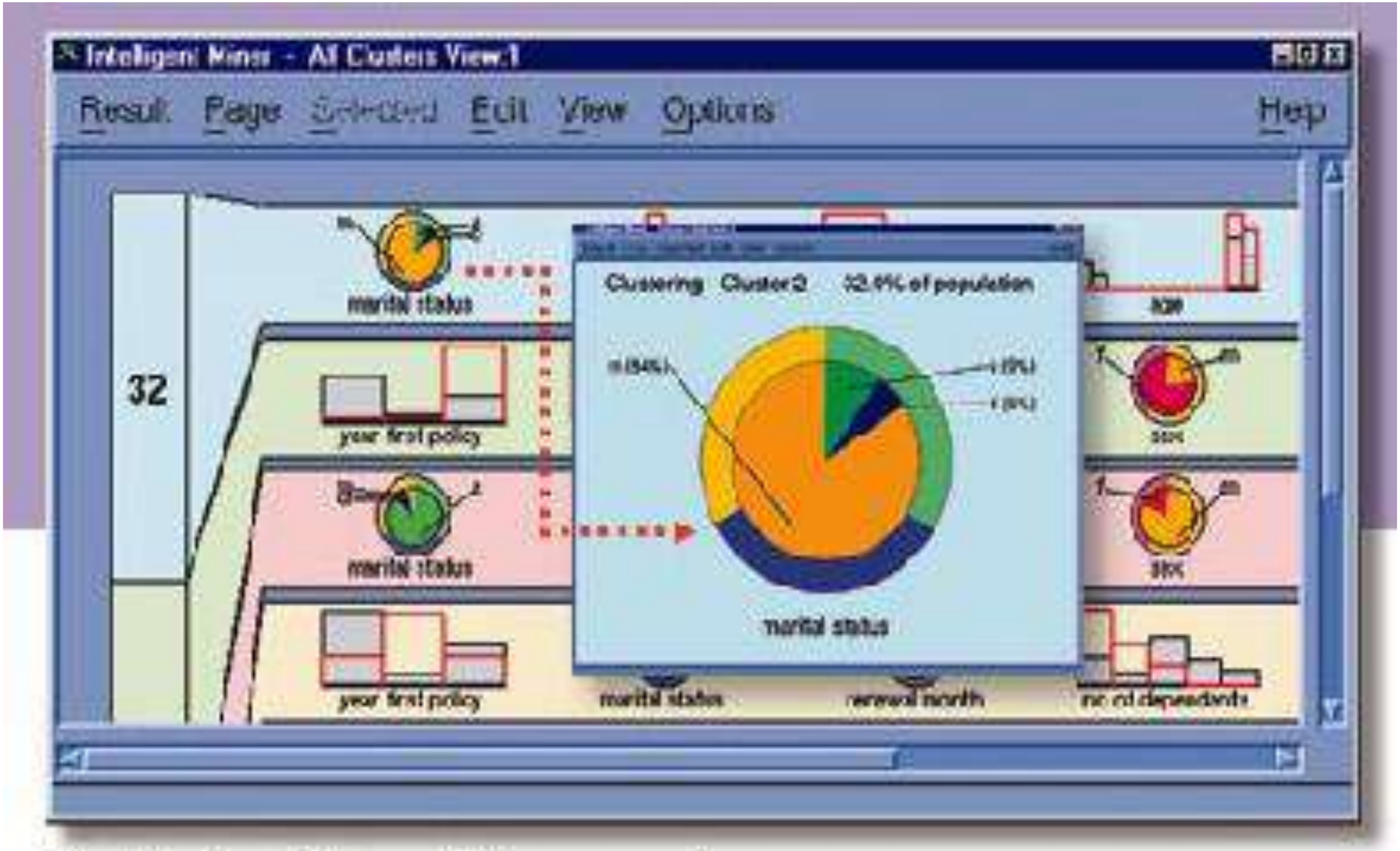
Pointer is over:

Acme Inc.:
$33,346.01, 83% of target, 96% of last year

Height: Total sales
Disk height: Target sales
Color: % of target 0% 100% 200% 500%

Ready

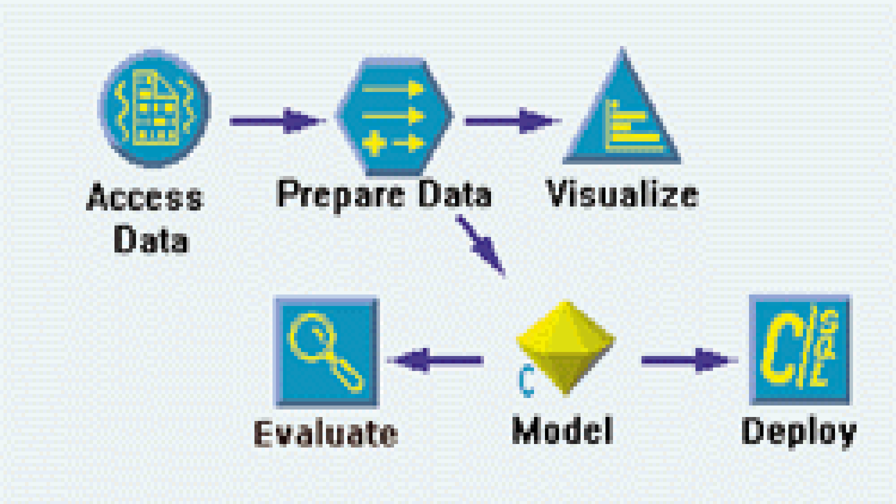
Visualization of Cluster Grouping in IBM Intelligent Miner



Data Mining Process Visualization

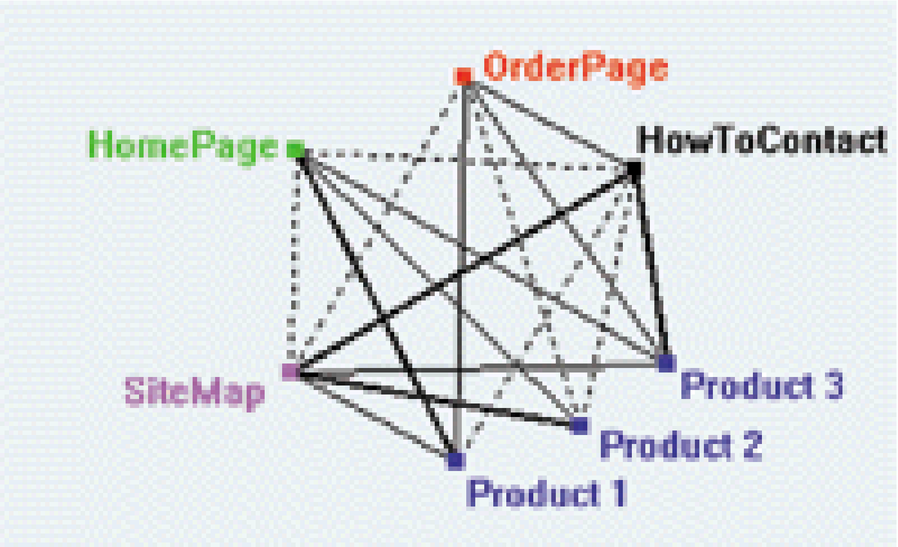
- ❖Presentation of the various processes of data mining in visual forms so that users can see
  - – Data extraction process
  - – Where the data is extracted
  - – How the data is cleaned, integrated, preprocessed, and mined
  - – Method selected for data mining
  - – Where the results are stored
  -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

Visualization of Data Mining Processes by Clementine



See your solution discovery process clearly

Understand variations with visualized data

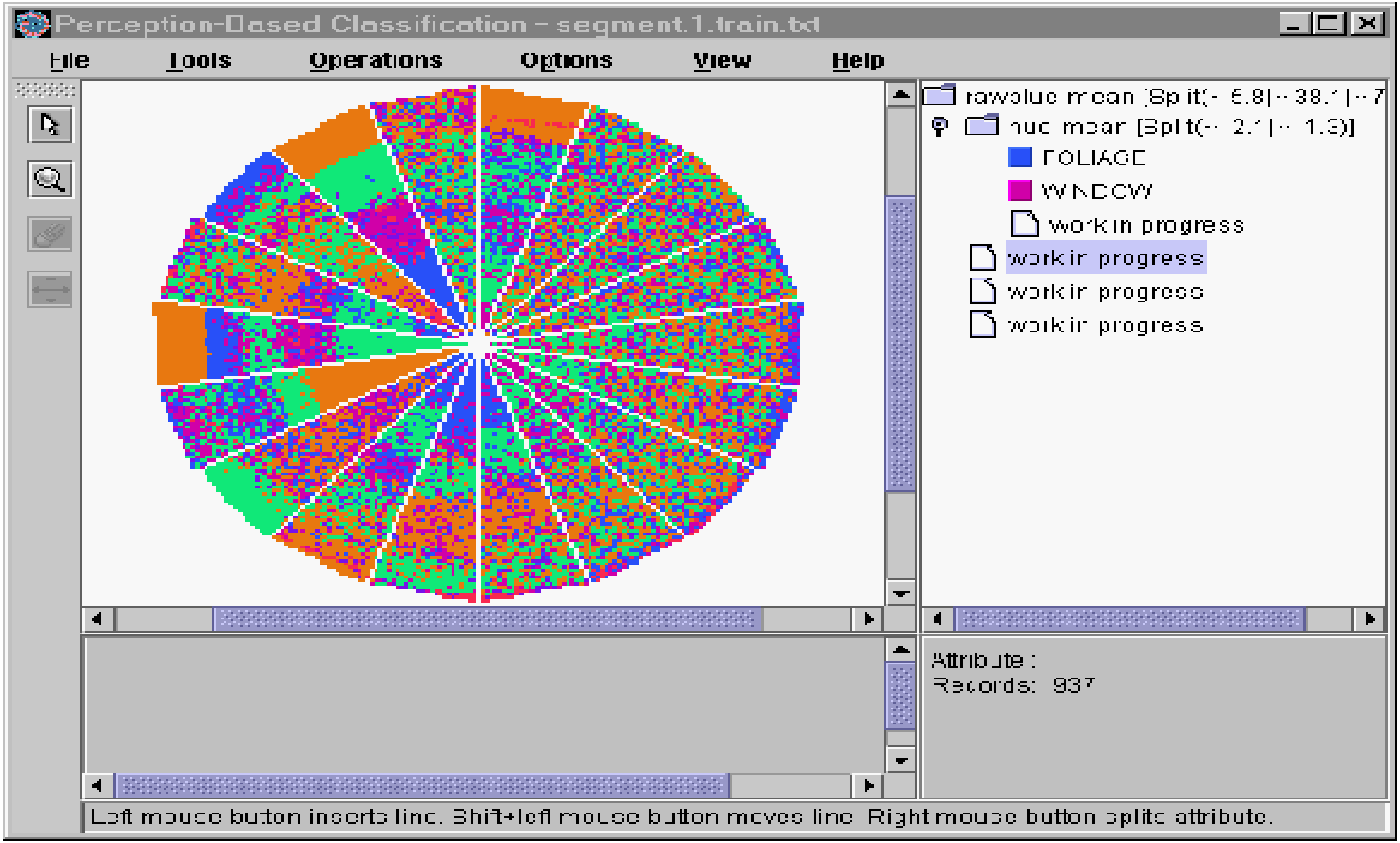


Interactive Visual Data Mining

- 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)



rawvalue mean [Split(- 6.8|.. 38.1|.. 7

hue mean [Split(.. 2.1|.. 1.3)]

🟦 FOLIAGE

🟪 WINDOW

work in progress

work in progress

work in progress

work in progress

Attribute :
Records: 937

Left mouse button inserts line. Shift+left mouse button moves line. Right mouse button splits attribute.

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

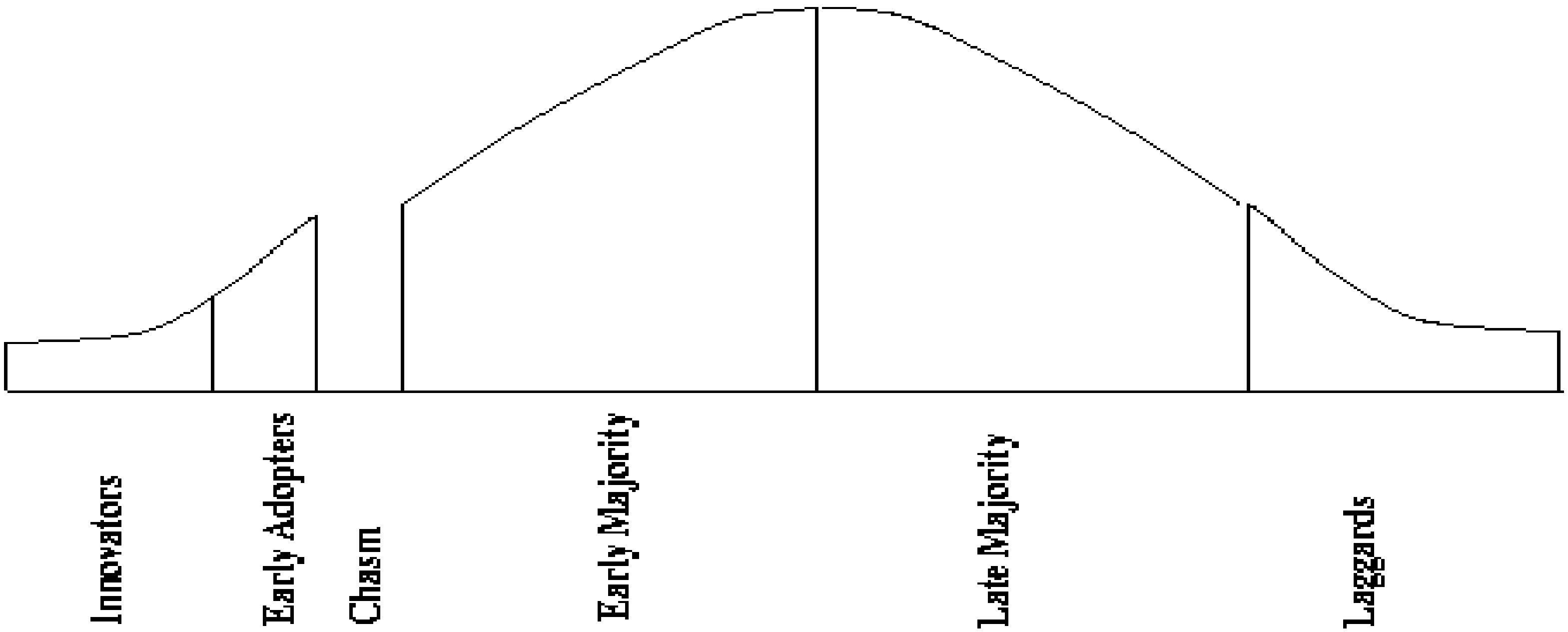
Data Mining and Collaborative Filtering

Social Impact of Data Mining

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

Social Impacts: Threat to Privacy

- ❖ 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
    - You use your credit card, debit card, supermarket loyalty card, or frequent flyer card, or apply for any of the above
    - You surf the Web, reply to an Internet newsgroup, subscribe to a magazine, rent a video, join a club, 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

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

Trends in Data Mining

- 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

- 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