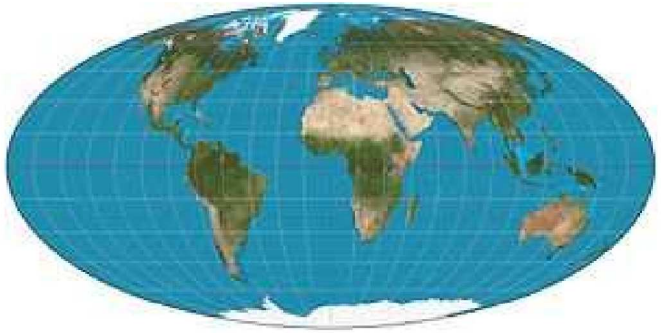
Week 13
Mining Complex Types of Data

Seokho Chi
Professor | Ph.D.
SNU Construction Innovation Lab

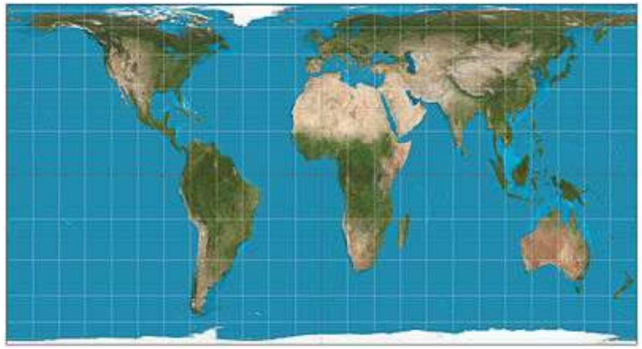
Source: Tan, Kumar, Steinback (2006)

Mining Complex Types of Data

- Mining spatial databases
- Mining multimedia databases
- Mining time-series and sequence data
- Mining the World-Wide Web



Spatial Data



- Spatial data integration: a big issue
  - Structure-specific formats (raster- vs. vector-based, OO vs. relational models, different storage and indexing, etc.)
  - Vendor-specific formats (ESRI, MapInfo, Integraph, IDRISI, etc.)
  - Geo-specific formats (geographic vs. equal area projection(구체의 지구를 평면에 투영하는 방법), etc.)

Raster-based: composed of pixels
Vector-based: composed of paths (points where the paths start and end, straight or curved, border and fill, etc.)
ESRI: GIS mapping software

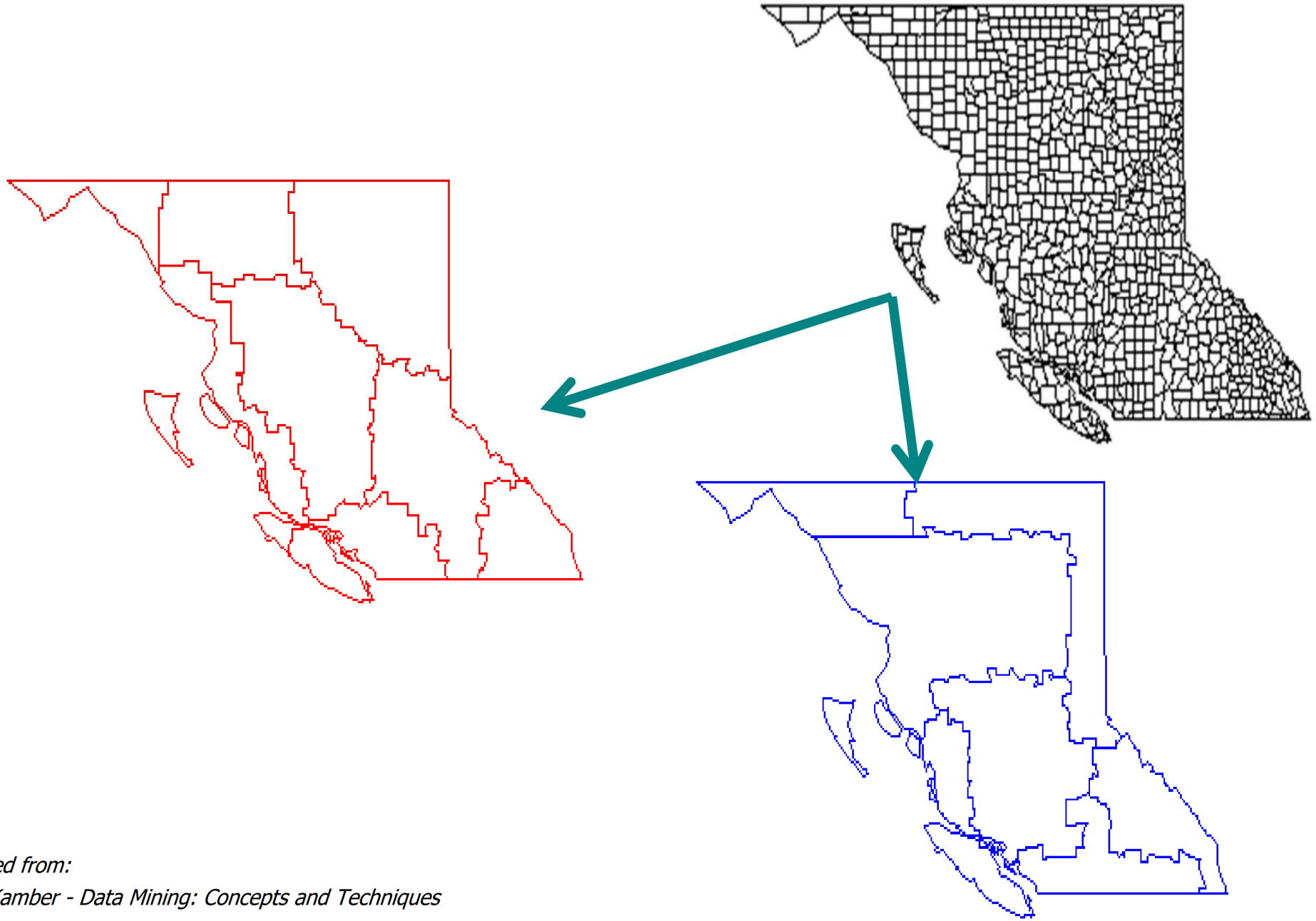
Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

Example: British Columbia Weather Pattern Analysis

- Input
  - A map with about 3,000 weather probes scattered in B.C.
  - Daily data for temperature, precipitation, wind velocity, etc.
- Output
  - A map that reveals patterns: merged (similar) regions
- Goals
  - Interactive analysis
  - Fast response time
  - Minimizing storage space used
- Challenge
  - A merged region may contain hundreds of "primitive" regions (polygons)

Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

Dynamic Merging of Spatial Objects



Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

Spatial Association Analysis

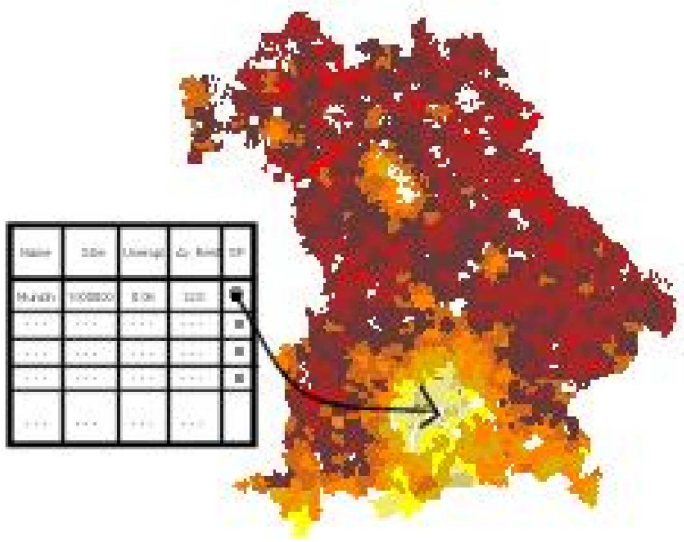
- Spatial association rule: $A \Rightarrow B$ [*s%, c%*]
  - A and B are sets of spatial or non-spatial predicates
    - Topological relations: *intersects, overlaps, disjoint,* etc.
    - Spatial orientations: *left_of, west_of, under,* etc.
    - Distance information: *close_to, within_distance,* etc.
  - *s%* is the support and *c%* is the confidence of the rule
- Examples

1) *is_a(x, large_town) ^ intersect(x, highway)* $\rightarrow$ *adjacent_to(x, water)*
[7%, 85%]

2) What kinds of objects are typically located close to golf courses?

Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

Spatial Trend Analysis

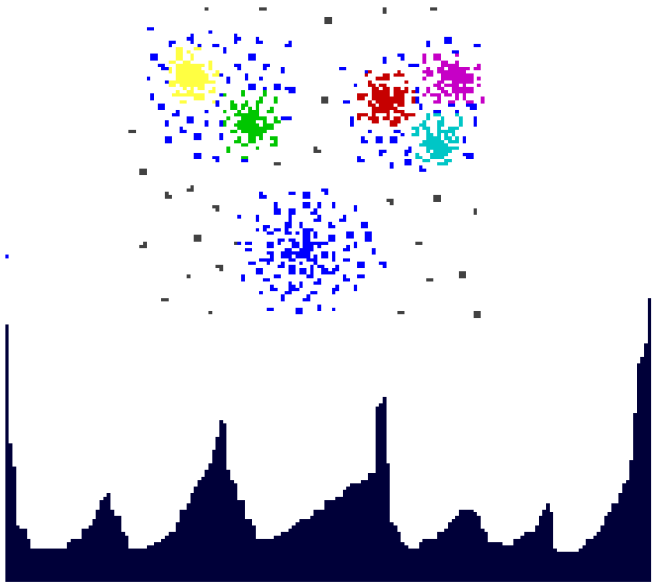


- Function
  - Detect changes and trends along a spatial dimension
  - Study the trend of non-spatial or spatial data changing with space
- Application examples
  - Observe the trend of changes of the climate or vegetation with increasing distance from an ocean
  - Crime rate or unemployment rate change with regard to city geo-distribution
  - Farm Insurance Frauds

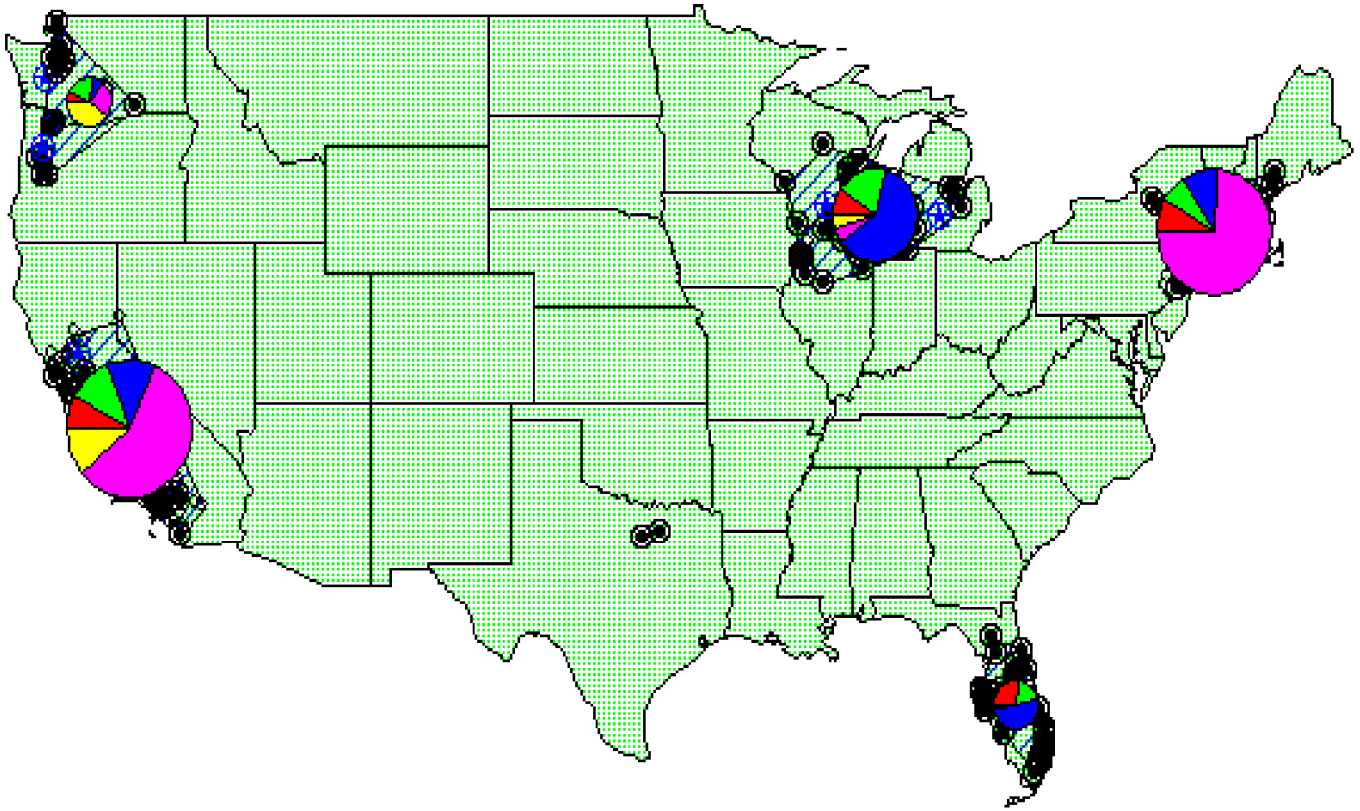
"Perpetrators falsely claim weather or insects destroyed their crops and cash in on a government-backed insurance program. Some don't bother planting at all. Others sell their harvests in secret."

Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

Spatial Cluster Analysis



- Mining clusters—k-means, hierarchical, density-based, etc.
- Analysis of distinct features of the clusters



Mining Complex Types of Data

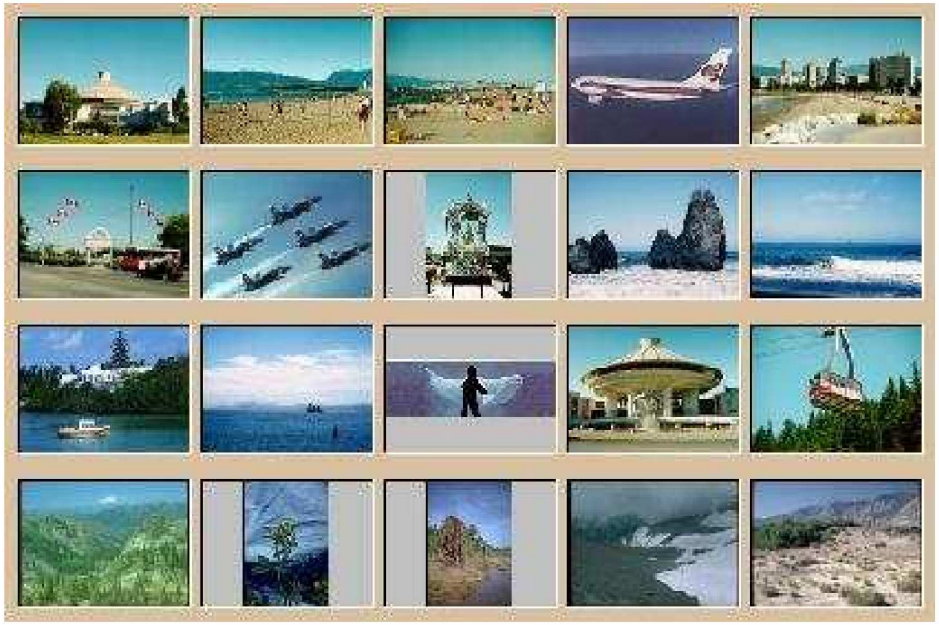
- Mining spatial databases
- Mining multimedia databases
- Mining time-series and sequence data
- Mining the World-Wide Web

Similarity Search in Multimedia Data

- Description-based retrieval systems
  - Build indices and perform object retrieval based on image descriptions, such as keywords, captions, size, and time of creation
  - Labor-intensive if performed manually
  - Results are typically poor quality if automated
- Content-based retrieval systems
  - Support retrieval based on the image content, such as color histogram, texture, shape, objects, and wavelet transforms

Mining Multimedia Databases

Refining or combining searches



Search for "blue sky"
(top layout grid is blue)

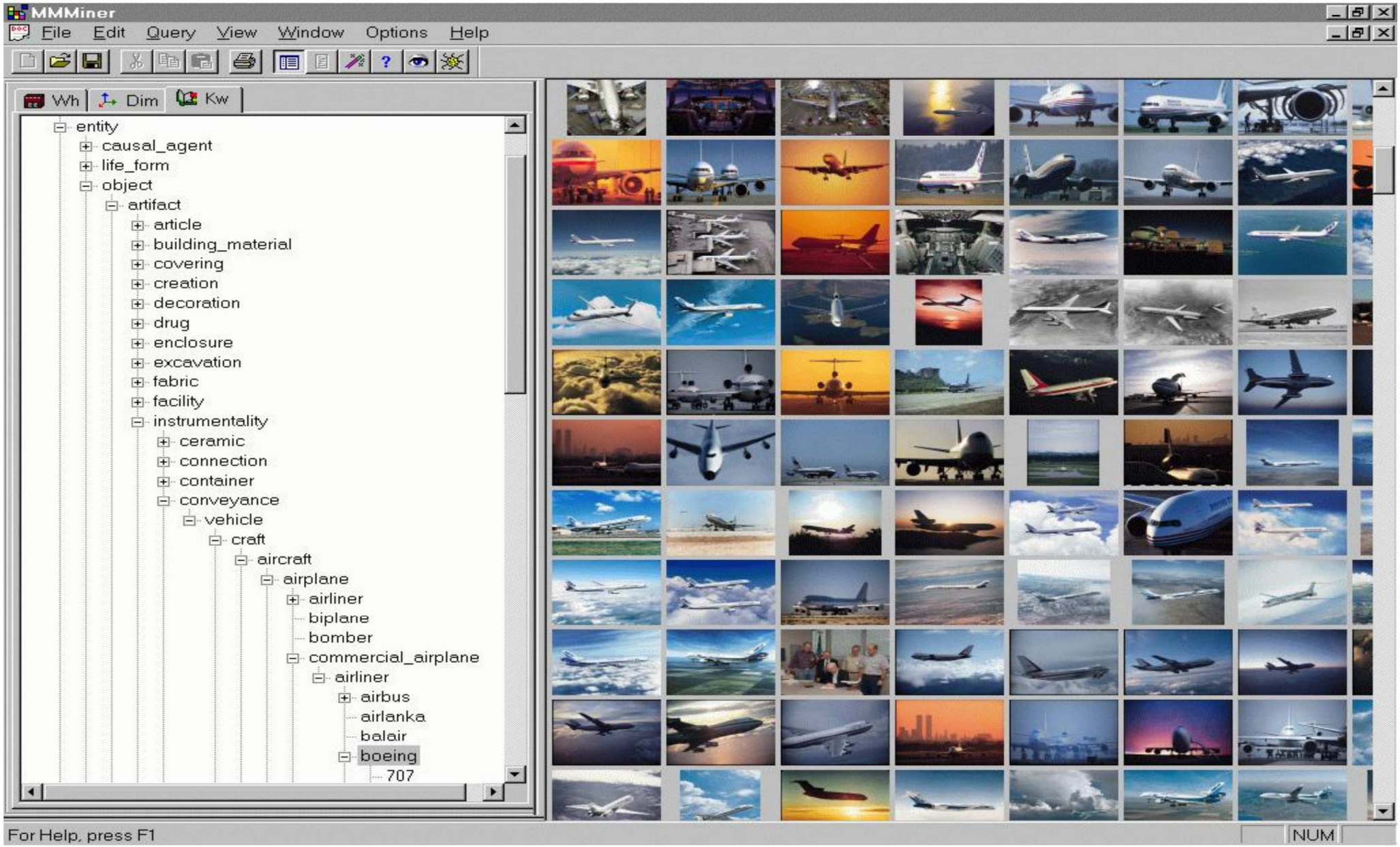
Search for "airplane in blue sky"
(top layout grid is blue and keyword = "airplane")

Search for "blue sky and green meadows"
(top layout grid is blue and bottom is green)

Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

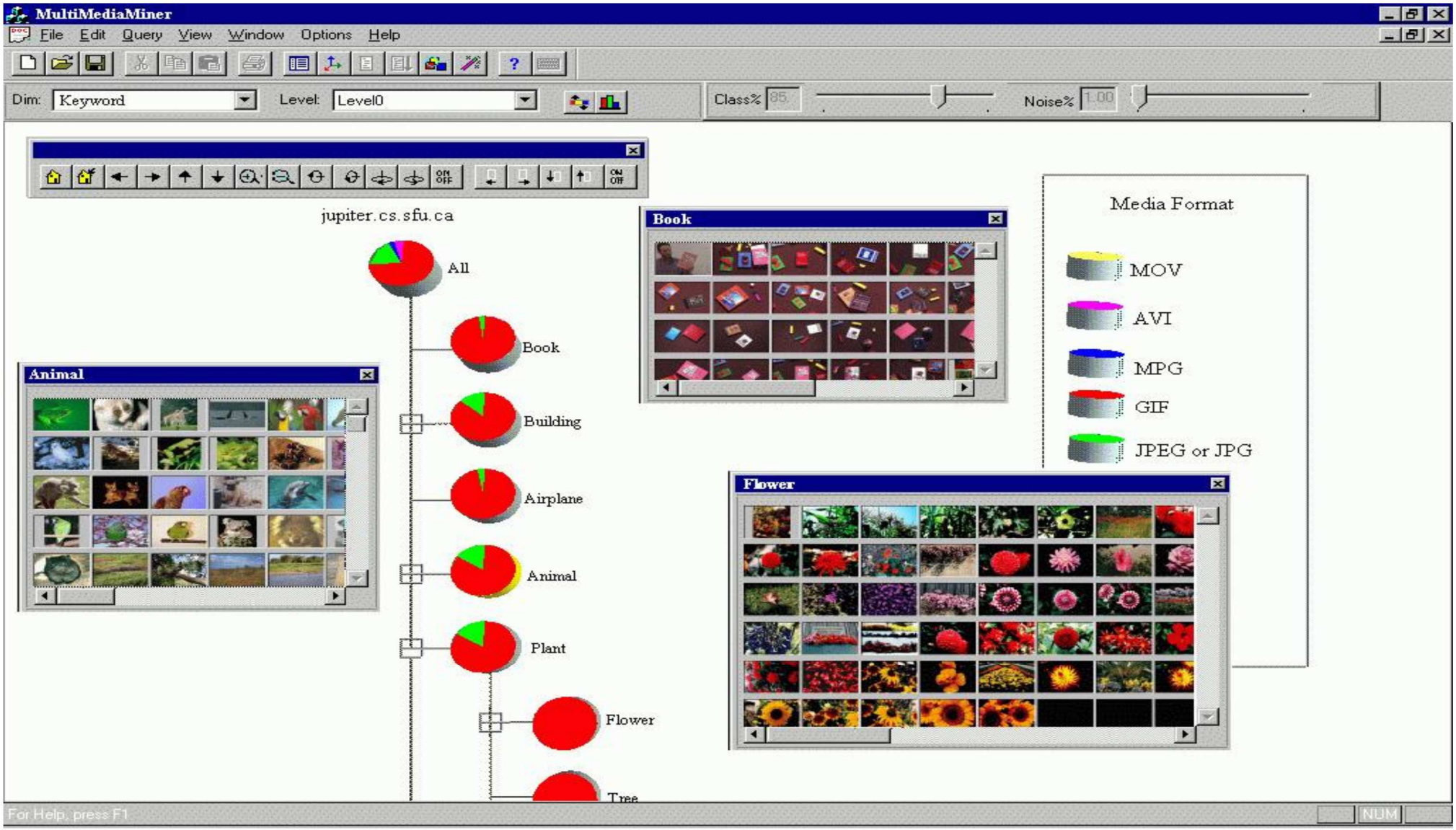
Mining Multimedia Databases in MultiMediaMiner

Thumbnails of images and video frames in the database can be browsed with MultiMediaMiner user interface.



Classification in MultiMediaMiner

MM-Characterizer, MM-Comparator, MM-Associator, MM-Classifier

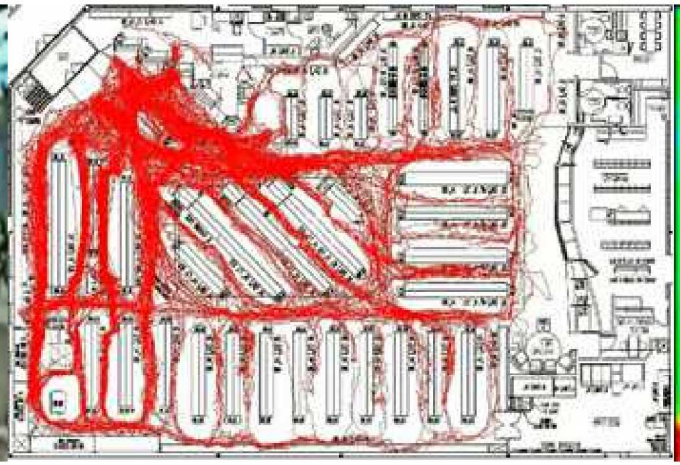


Adapted from:

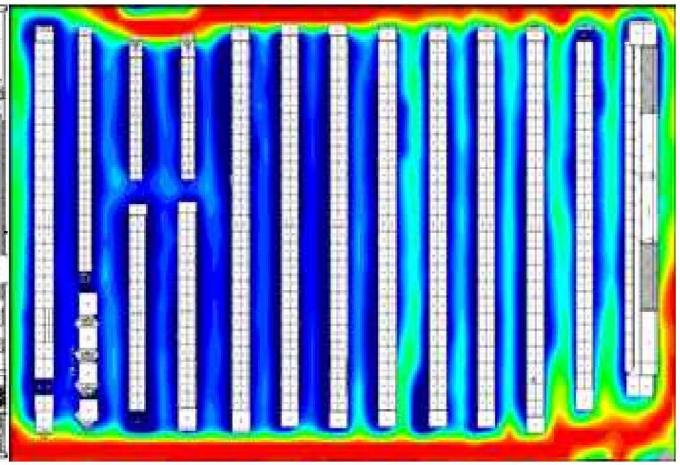
Han, Kamber - Data Mining: Concepts and Techniques

Classification in VideoMining (www.videomining.com)

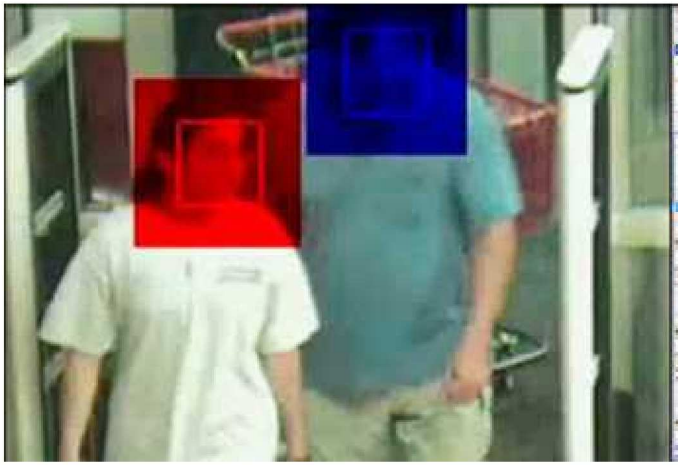
Tracking the Shopper Path



Multiple Shopping Trips



Heat Maps



Demographics Analysis

Market Analysis

Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

Mining Complex Types of Data

- Mining spatial databases
- Mining multimedia databases
- Mining time-series and sequence data
- Mining the World-Wide Web

Mining Time-Series and Sequence Data

- Time-series database
  - Consists of sequences of values or events changing with time
  - Data is recorded at regular intervals
  - Characteristic time-series components
    - Trend, cycle, seasonal, irregular
- Applications
  - Financial: stock price, inflation
  - Biomedical: blood pressure
  - Meteorological: precipitation

Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

Mining Time-Series and Sequence Data

Time-series plot

Adapted from:

Han, Kamber - Data Mining: Concepts a

Mining Time-Series and Sequence Data: Trend analysis

- A time series can be illustrated as a time-series graph which describes a point moving with the passage of time
- Categories of Time-Series Movements
  - Long-term or trend movements (trend curve)
  - Cyclic movements or cycle variations, e.g., business cycles
  - Seasonal movements or seasonal variations
    - i.e, almost identical patterns that a time series appears to follow during corresponding months of successive years.
  - Irregular or random movements

Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

Mining Complex Types of Data

- Mining spatial databases
- Mining multimedia databases
- Mining time-series and sequence data
- Mining the World-Wide Web

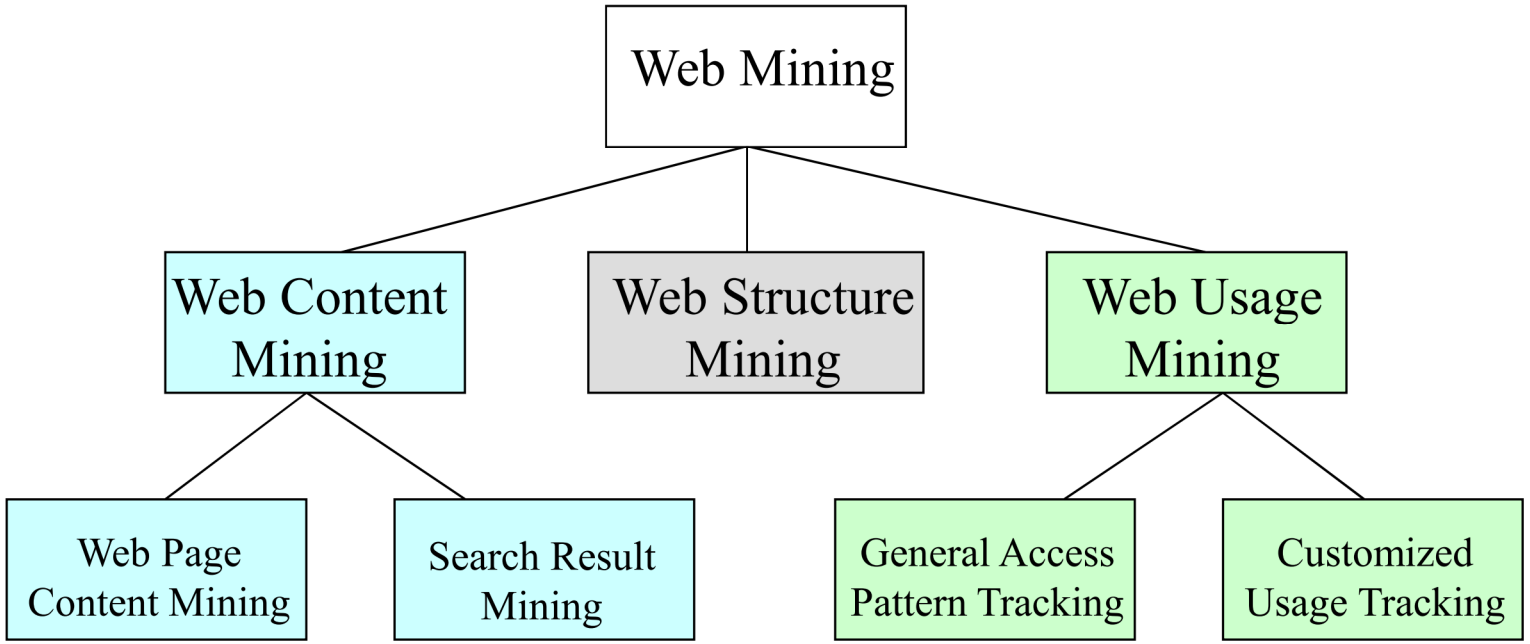
Mining the World-Wide Web

- The WWW is huge, widely distributed, global information service center for:
  - Information services: news, advertisements, consumer information, financial management, education, government, e-commerce, etc.
  - Hyper-link information
  - Access and usage information
- WWW provides rich sources for data mining
- Challenges
  - Too huge for effective data warehousing and data mining
  - Too complex and heterogeneous: no standards and structure

99% of the Web information is useless to 99% of Web users
How can we find high-quality Web pages on a specified topic?

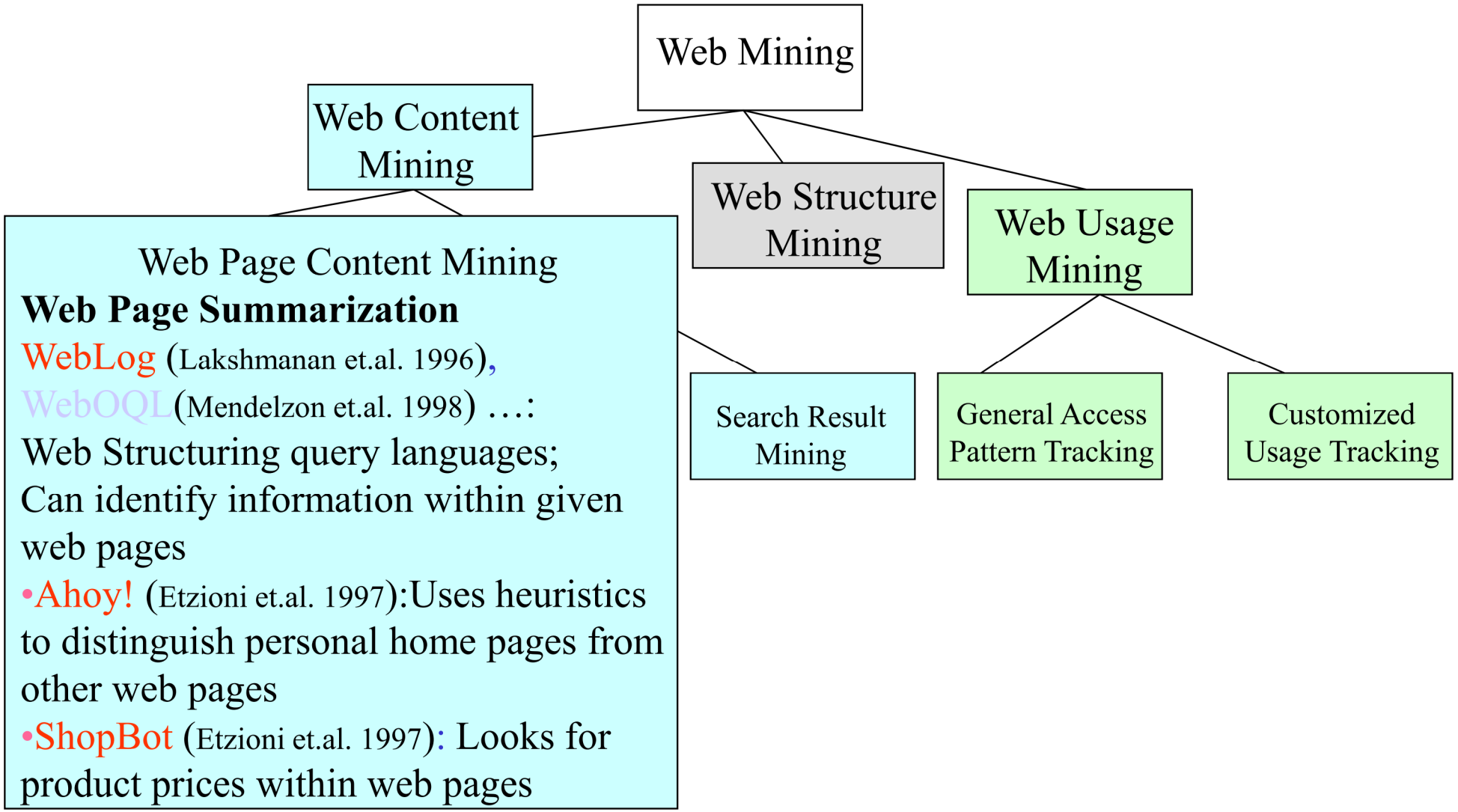
Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

Web Mining Taxonomy



Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

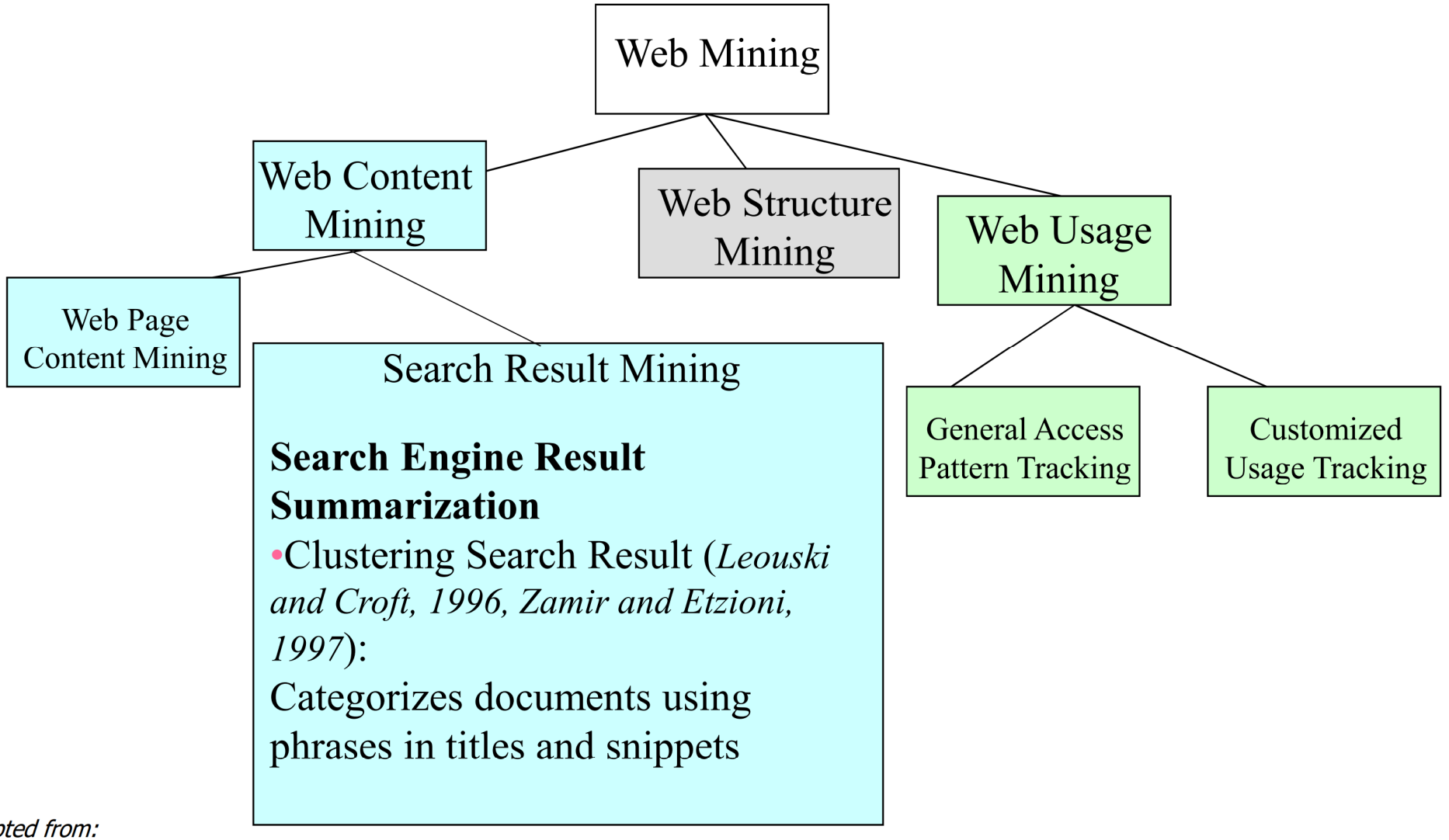
Mining the World-Wide Web



Adapted from:

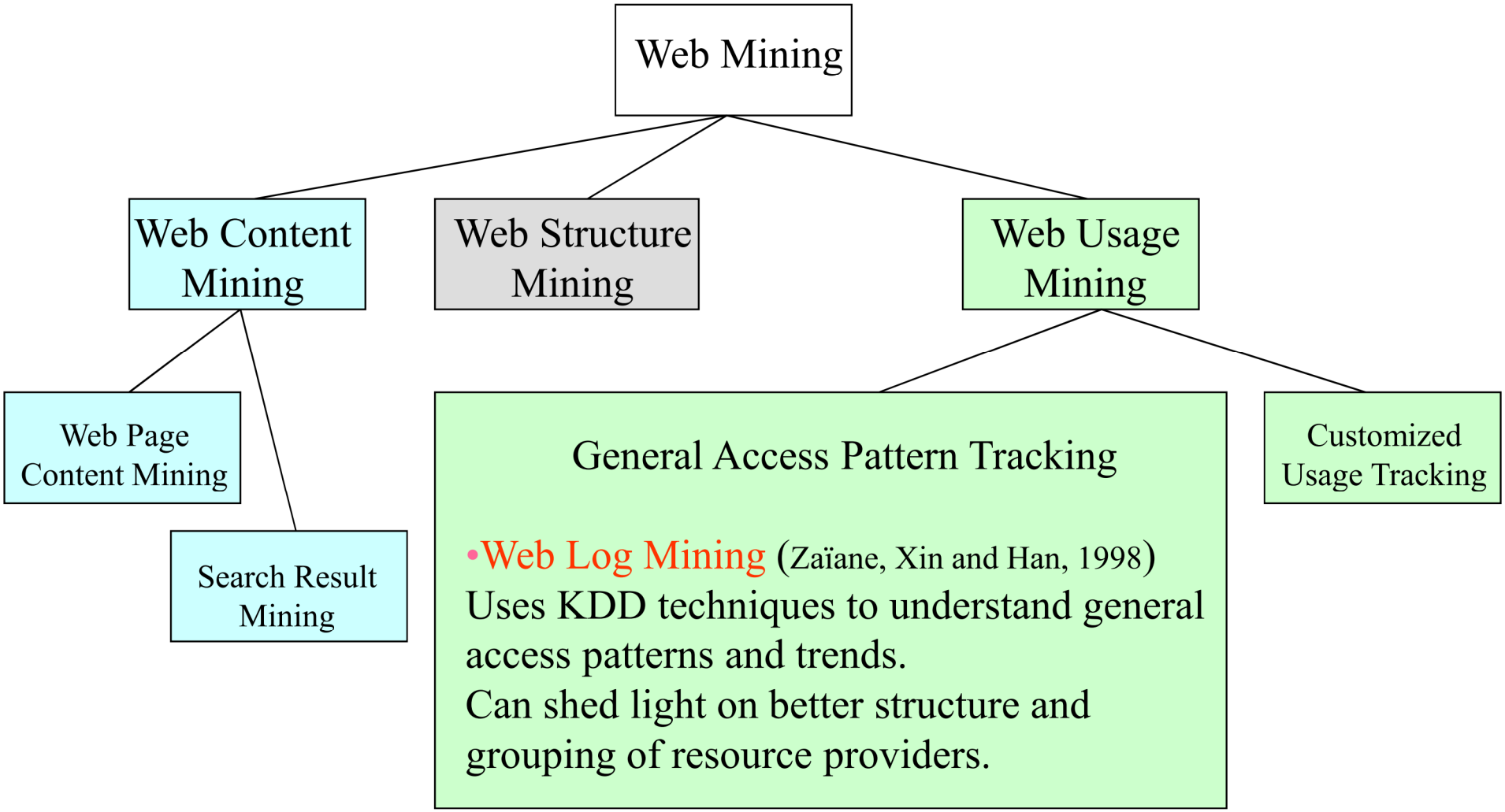
Han, Kamber - Data Mining: Concepts and Techniques

Mining the World-Wide Web



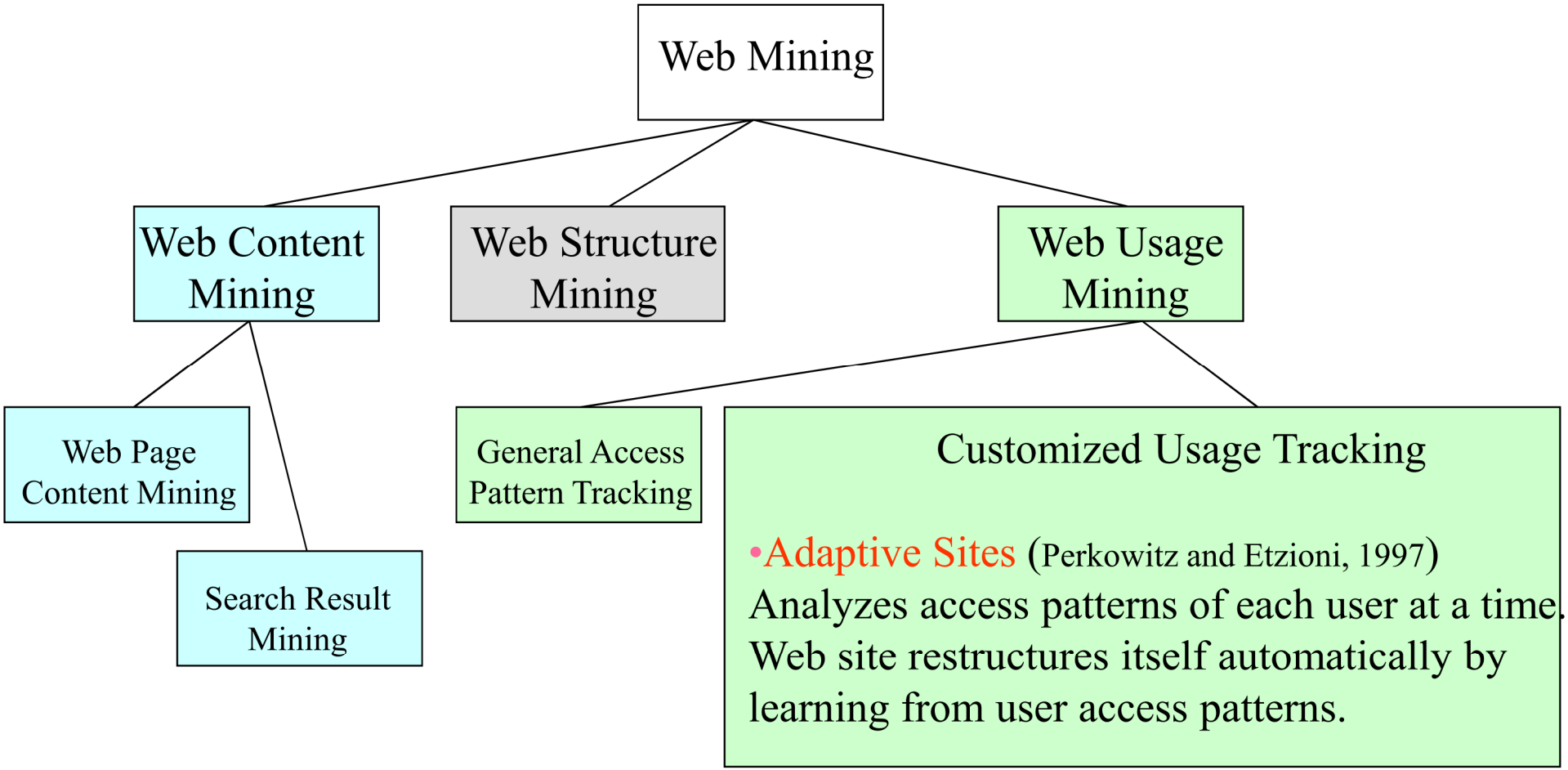
Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

Mining the World-Wide Web



Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

Mining the World-Wide Web



Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

Web Usage Mining

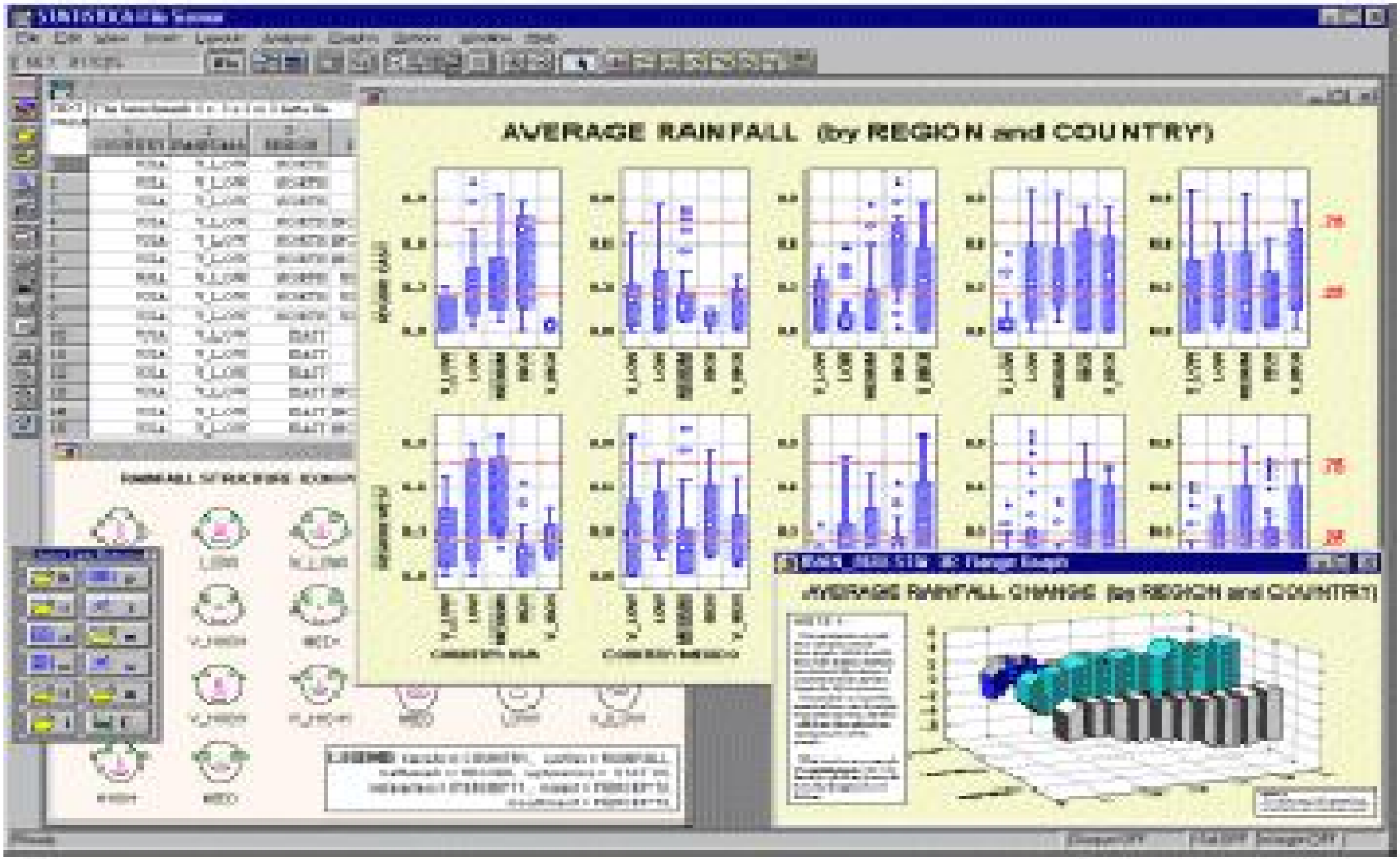
- Mining Web log records to discover user access patterns of Web pages
- Applications
  - Target potential customers for electronic commerce
  - Enhance the quality and delivery of Internet information services to the end user
  - Improve Web server system performance
  - Identify potential prime advertisement locations
- Web logs provide rich information about Web dynamics
  - Typical Web log entry includes the URL requested, the IP address from which the request originated, and a timestamp

Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

Others

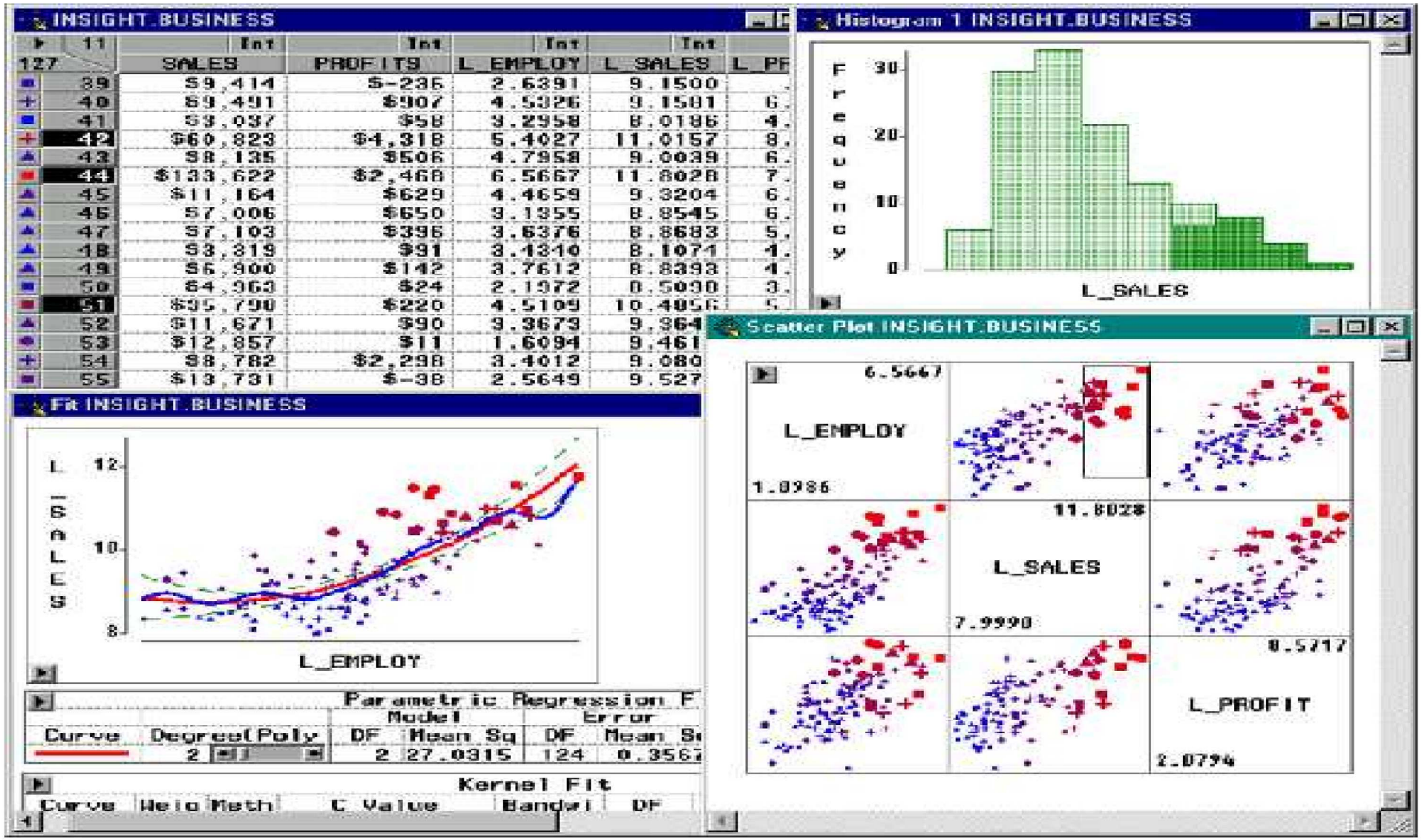
Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

Boxplots from Statsoft: Multiple Variable Combinations



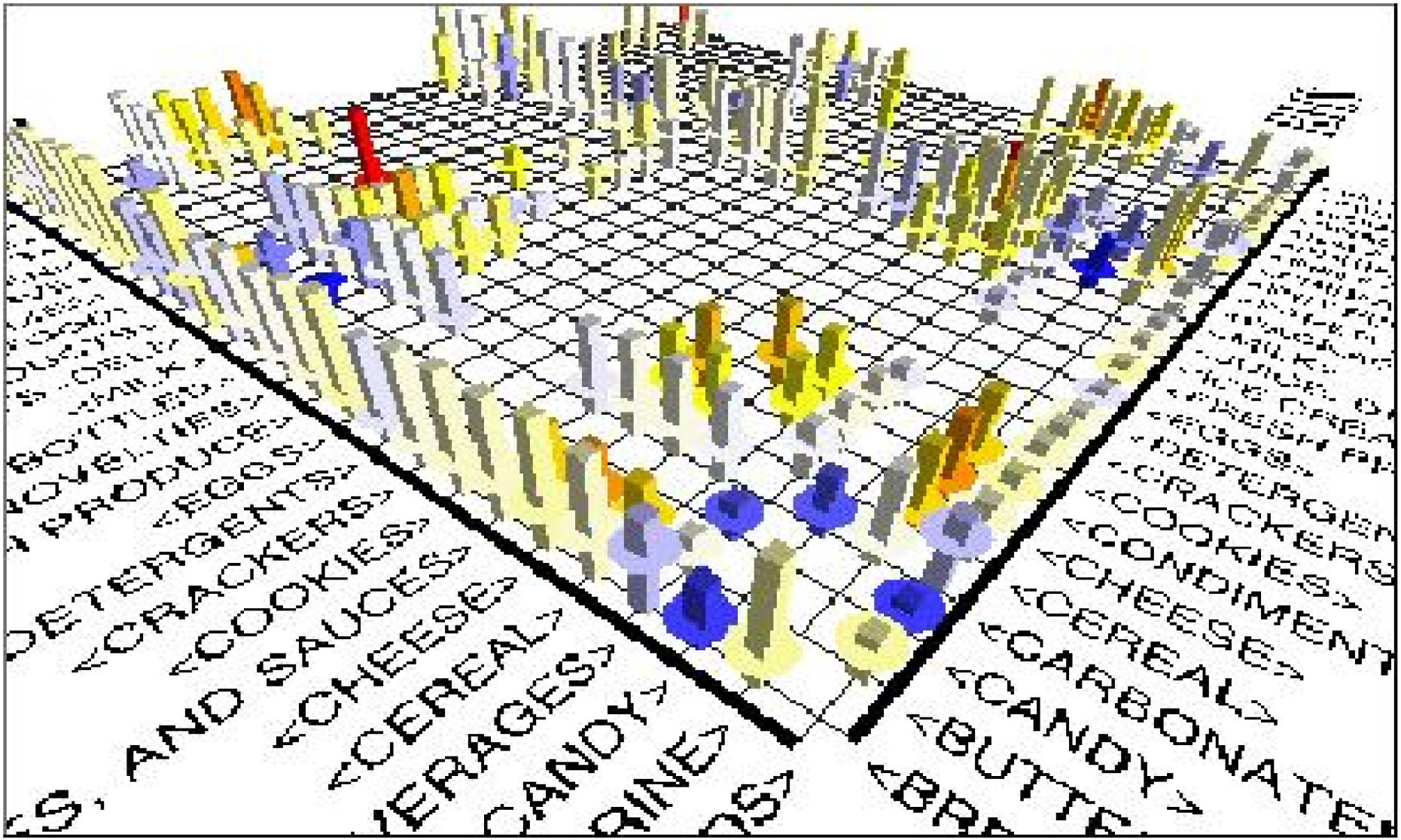
Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

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



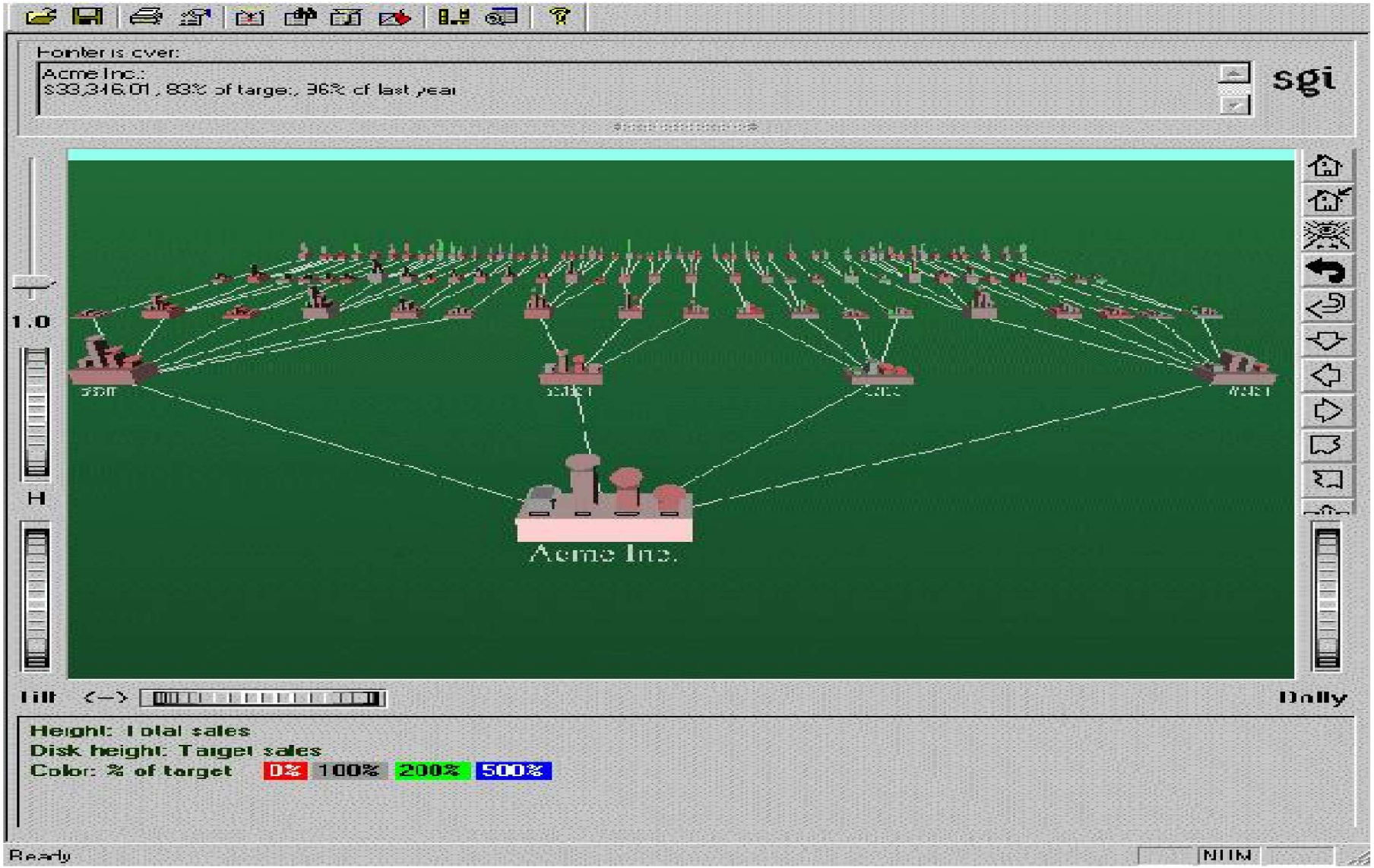
Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

Visualization of Association Rules in SGI/MineSet 3.0



Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

Visualization of a Decision Tree in SGI/MineSet 3.0



Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

Visualization of Cluster Grouping in IBM Intelligent Miner

Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

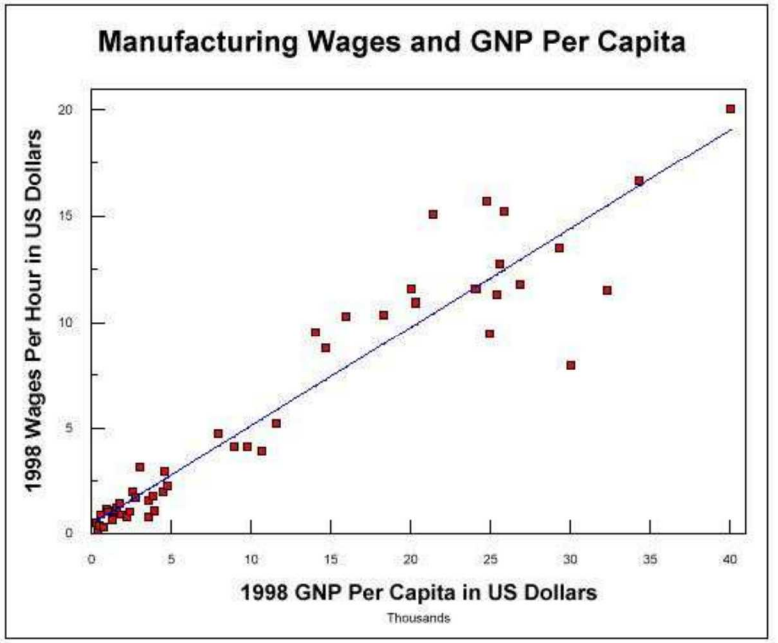
Scientific and Statistical Data Mining

- 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

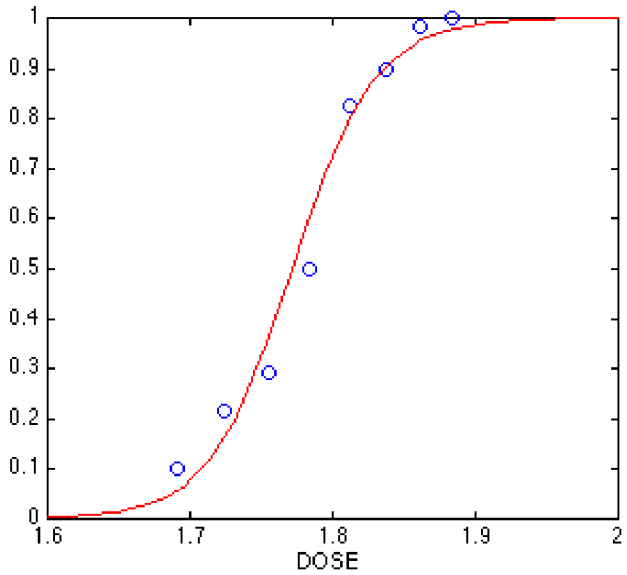
  - forms of regression: linear, multiple, weighted, polynomial, etc.



Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

Scientific and Statistical Data Mining

- **Generalized linear models**
  - Unifying various other statistical models including linear regression, logistic regression, and Poisson regression
  - Iteratively reweighted least squares methods for maximum likelihood estimation of parameters
  - Similar to the modeling of a numeric response variable using linear regression

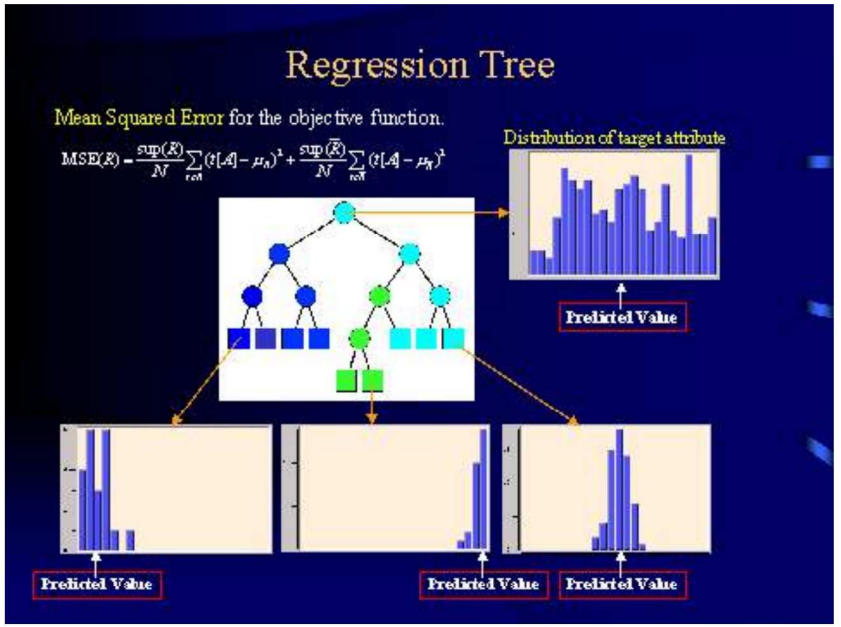


- **Mixed-effect models**
  - For analyzing grouped data, i.e. data that can be classified according to one or more grouping variables
  - Typically describe relationships between a response variable and some covariates in data grouped according to one or more factors

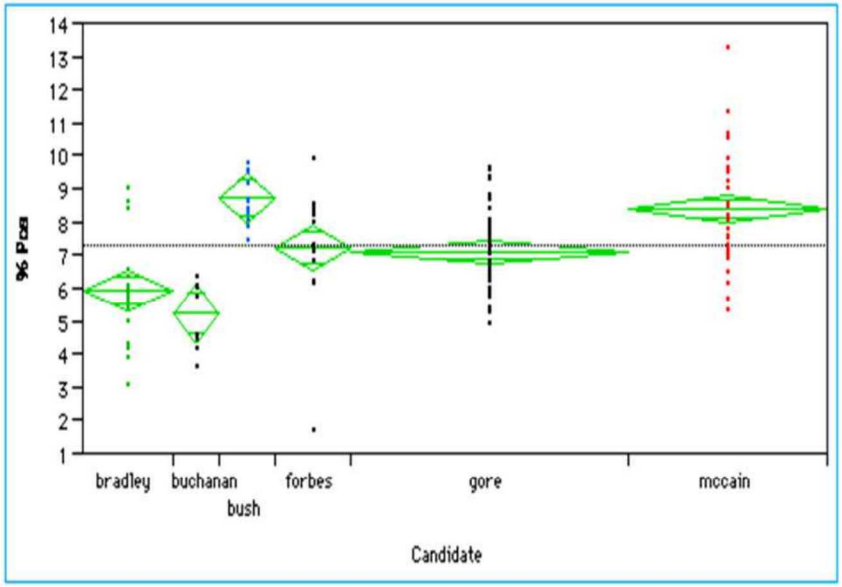
Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

Scientific and Statistical Data Mining

- **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



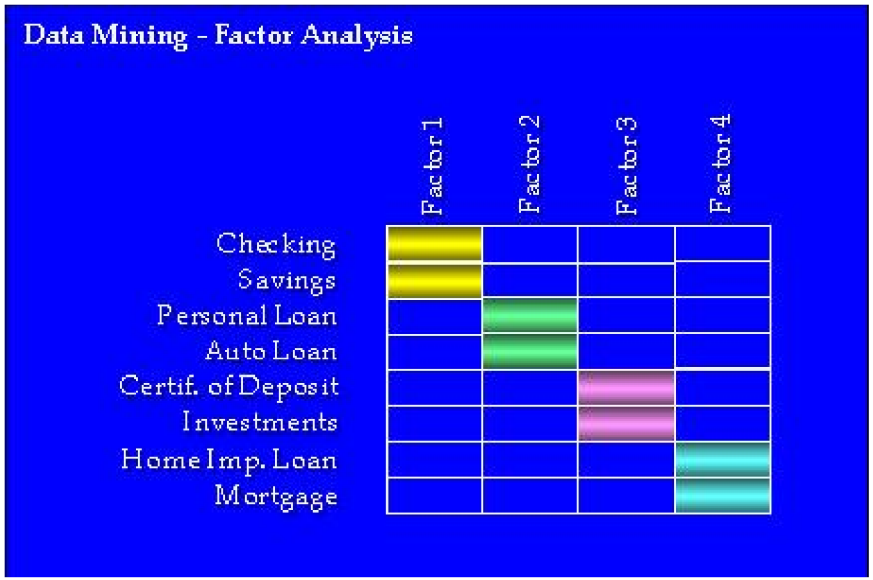
- **Analysis of variance**
  - Analyze experimental data for two or more populations described by a numeric response variable and one or more categorical variables (factors)



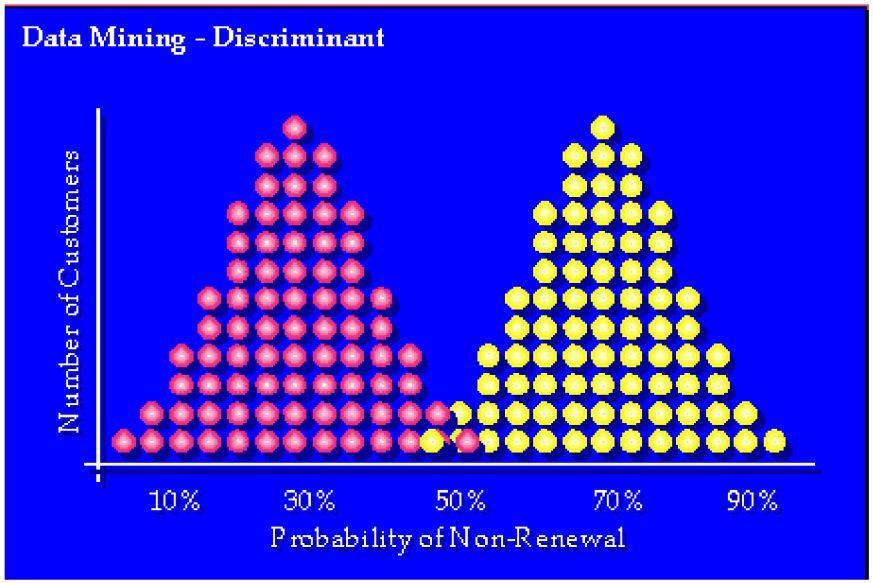
Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

Scientific and Statistical Data Mining

- **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



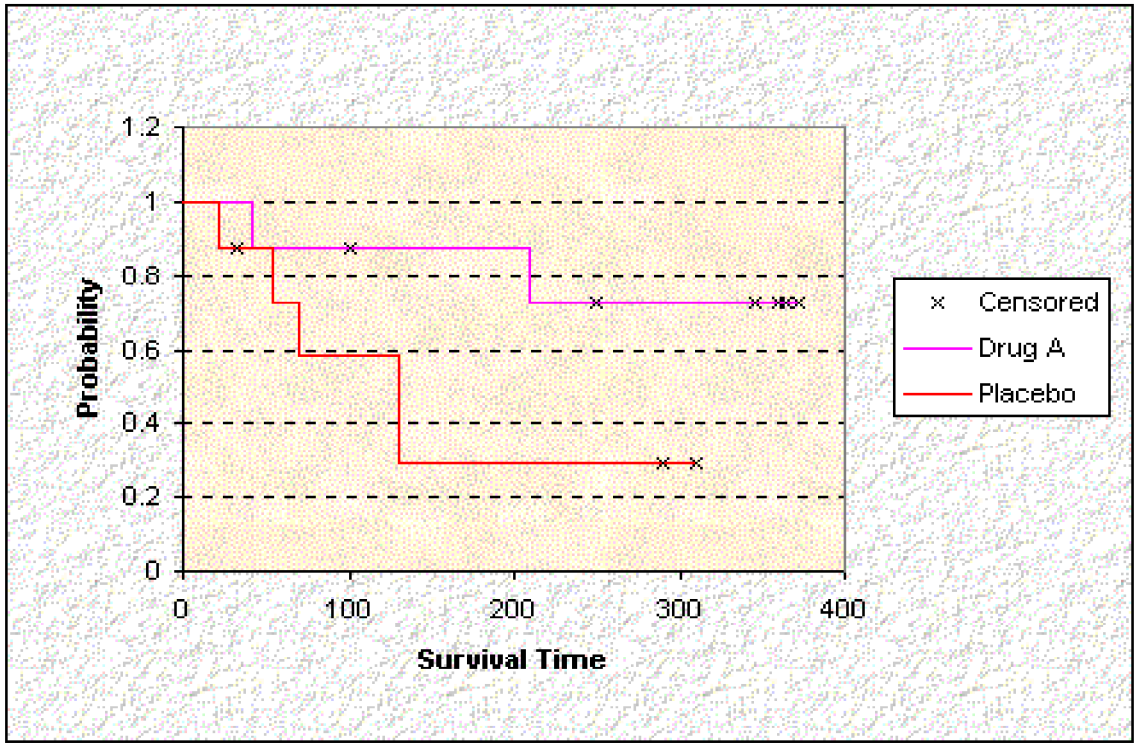
Adapted from:

Han, Kamber - Data Mining: Concepts and Techniques

Scientific and Statistical Data Mining

- **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)



Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

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

Adapted from:
Han, Kamber - Data Mining: Concepts and Techniques

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.