MBA 8473 - Data Mining & Knowledge Discovery

MBA 8473

Learning Objectives

55. Explain what is data mining?

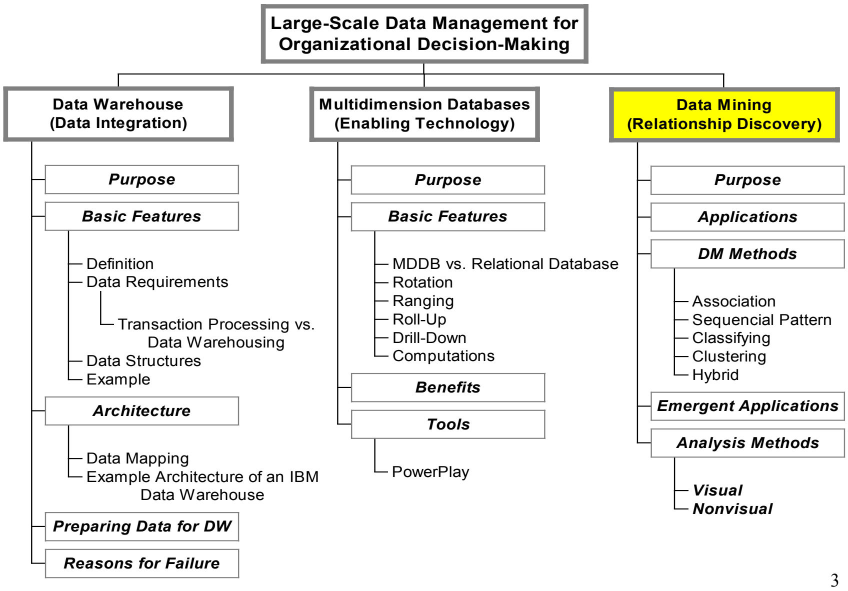
56. Explain two basic types of applications of data mining.

    55.1. Compare and contrast various types of rules.

57. Explain Four Data mining methods and describe how each can use both Visual and Non-visual techniques)
    - 57.1 Association
    - 57.2 Sequence
    - 57.3 Classification
    - 57.4 Clustering

58. Demonstration only- Use of Excel, SPSS (dropped), Backpack a Neural Network technology (dropped).

Organization of Concepts



- **Large-Scale Data Management for Organizational Decision-Making**
  - **Data Warehouse (Data Integration)**
    - *Purpose*
    - *Basic Features*
      - Definition
      - Data Requirements
        - Transaction Processing vs. Data Warehousing
      - Data Structures
      - Example
    - *Architecture*
      - Data Mapping
      - Example Architecture of an IBM Data Warehouse
    - *Preparing Data for DW*
    - *Reasons for Failure*
  - **Multidimension Databases (Enabling Technology)**
    - *Purpose*
    - *Basic Features*
      - MDDB vs. Relational Database
      - Rotation
      - Ranging
      - Roll-Up
      - Drill-Down
      - Computations
    - *Benefits*
    - *Tools*
      - PowerPlay
  - **Data Mining (Relationship Discovery)**
    - *Purpose*
    - *Applications*
    - *DM Methods*
      - Association
      - Sequencial Pattern
      - Classifying
      - Clustering
      - Hybrid
    - *Emergent Applications*
    - *Analysis Methods*
      - *Visual*
      - *Nonvisual*

What is Data Mining and its purpose?

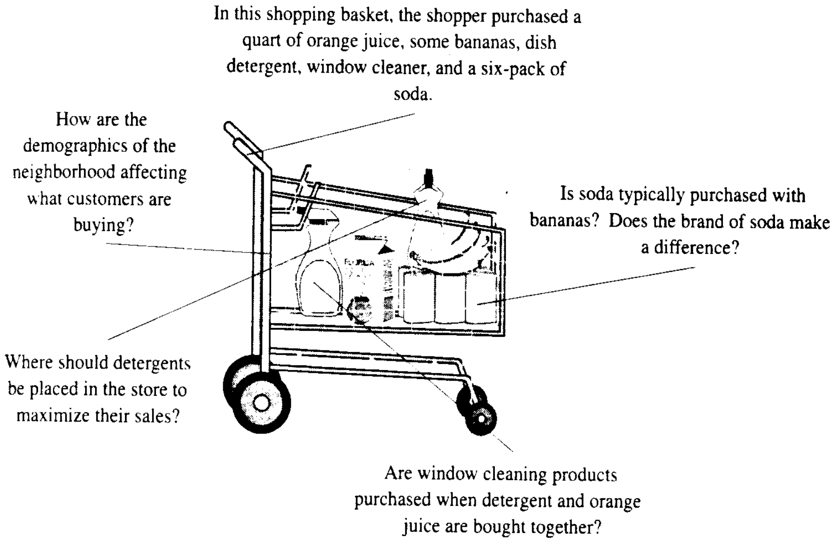
(L.O. 55)

- Search for relationships and global patterns that exist in large databases but are hidden in the vast amounts of data.
- Analyst combines knowledge of data and machine learning technologies to discover nuggets of knowledge hidden in the data.
- Serendipity to science.
- Easier and more effective when the organization has accumulated as much data as possible, such as with a data warehouse
- A data warehouse is *not* a prerequisite to data mining

APPLICATIONS - Market Basket Analysis (MBA) (L.O. 56)

- **MBA is form of clustering used for finding groups that tend to occur together in a transaction (or market basket).**
  The models are built to find the likelihood of different products being purchased together and can be expressed as a rule.

- **Example rules found from real data:**
  - On Thursdays, grocery store consumers often purchase diapers and beer together.
  - Customers who purchase maintenance agreements are very likely to purchase large appliances.
  - When a new hardware store opens, one of the most commonly sold items is toilet rings.



In this shopping basket, the shopper purchased a quart of orange juice, some bananas, dish detergent, window cleaner, and a six-pack of soda.

How are the demographics of the neighborhood affecting what customers are buying?

Is soda typically purchased with bananas? Does the brand of soda make a difference?

Where should detergents be placed in the store to maximize their sales?

Are window cleaning products purchased when detergent and orange juice are bought together?

Figure 8.1 Market basket analysis helps us understand what items are likely to be purchased together.

Taxonomies of items can help decide which items to focus MBA on (O.2).

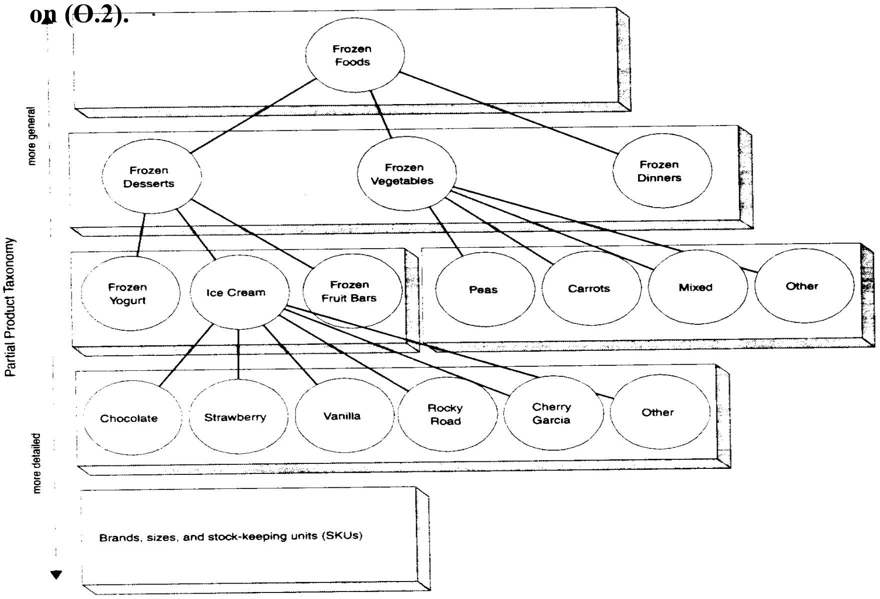


Figure 8.4 Taxonomies start with the most general and move to increasing detail.

All rules are not *useful*

(L.O. 56.1)

- Three common types of rules that can be produced by by MBA:
  - (1) *Useful* rule - have some cause and provides actionable information
  - (2) *Trivial* rule - is one that is already known by anyone at all familiar with the business
  - (3) *Inexplicable* rule - seems to have no explanation and do not suggest a course of action.
- Using the above three types, try to rate the rules from previous slide.

A special case of ‘trivial’ rule..

(L.O.56.1)

- Consider a seemingly interesting result - the people who buy the three-way calling option on their local telephone service almost always buy call waiting
  - A subtle problem could be that this may be the result of marketing promotions and product bundling.
- Results may simply be measuring the success of previous marketing campaigns.

Useful rules lead to action...

(L.O. 56.1)

- How can we incent users to *put* other items that they are likely to purchase into their carts? - Relocate items on the ‘isle’, etc.

Other Data Mining Applications

(L.O.56)

- **Memory Based Reasoning (MBR)**
  - Based on past data (i.e., memory), *identify* similar cases from experience, then *apply* the information to the problem at hand.
- **Example**
  - Fraud detection - new cases of fraud are likely to be similar to known cases.
  - Customer response prediction - the next customers likely to respond to an offer are probably similar to previous customers that have responded.
  - Medical treatments.
  - MCI mines data from 140 million households, each with as many as 10,000 attributes, including life-style and calling habits. Have identified 22 profiles (secret!)

Some popular use of data mining: Customer Relationship Marketing

- Business-to-Consumer Management
  - Build customer profiles using data collected from web visits
  - Focus on one-to-one marketing
  - Customizing products and services for each consumer
- Profile warehousing business
  - Track what customers do during each site visit
  - Record time between clicks, links between clicks
  - AOL purchasing profile warehouses (e.g., *Junglee*)
  - Oracle developing product line for profile warehousing
  - Mine the data for relationships

Four data Mining Methods

(L.O.57)

1. Looking for *association* or co-existence, co-occurrence of events (suitable for MBA)
2. Looking for *sequence* or temporal patterns (MBA, MBR)
3. Looking for *classification* of data (MBA, MBR) - target groups are known in the beginning.
4. Looking for *clustering* of data (MBA, MBR) - target groups are NOT known in the beginning

Data Mining Method #1

(L.O.57.1)

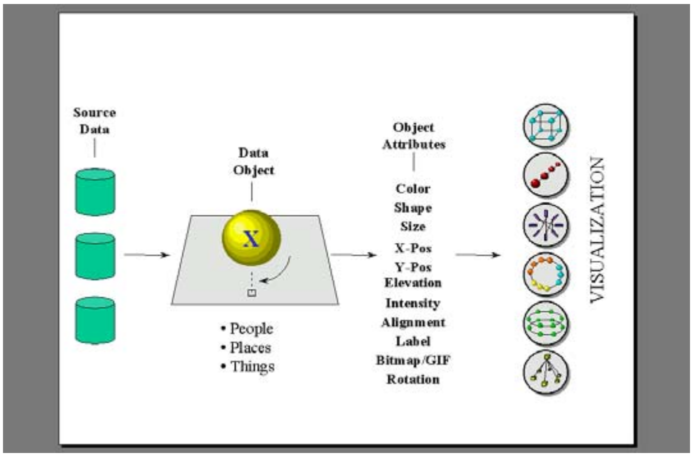
1. **Find Association** (can be converted into rules)
   - Identifies affinities existing among the collection of items in a given set of records
   - 80 percent of all records that contain A, B and C also contain D and E; I.e., if A, B and C Then D and E.
   - 85 percent of customers who buy a certain wine brand also buy a certain type of pasta; If buys Wine X then buys Pasta C.
   - On Thursdays, many customers buy a six-pack when they purchase diapers. If Thursday and buys six-pack, then buys diapers.
   - How good is the rule? (We will use grocery data example to clarify the issue of 'confidence')

Analysis Methods for Discovering Association

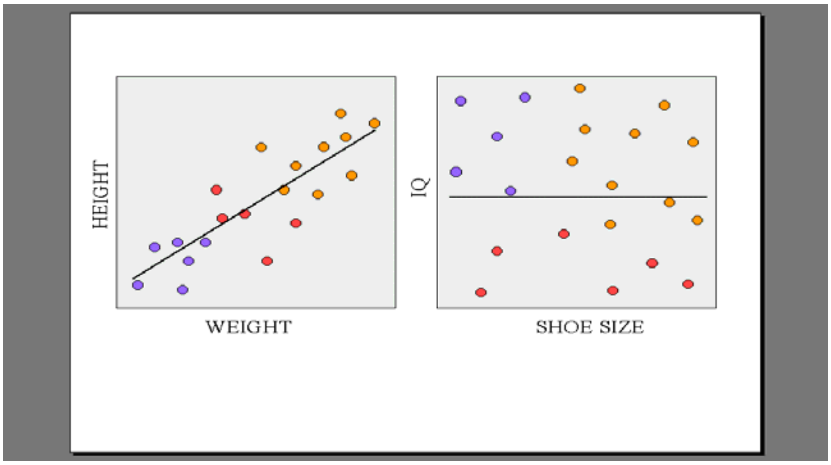
(L.O.57.1)

- Visual methods
  - Strategy for visualizing associations
  - Specific association detection
    - Scatter plot
    - Segmented scatter plot
    - Link analysis
      - builds up networks of interconnected objects.
    - Landscape visualization
      - the relative positioning of data elements within the geometric terrain represents information important for analysis

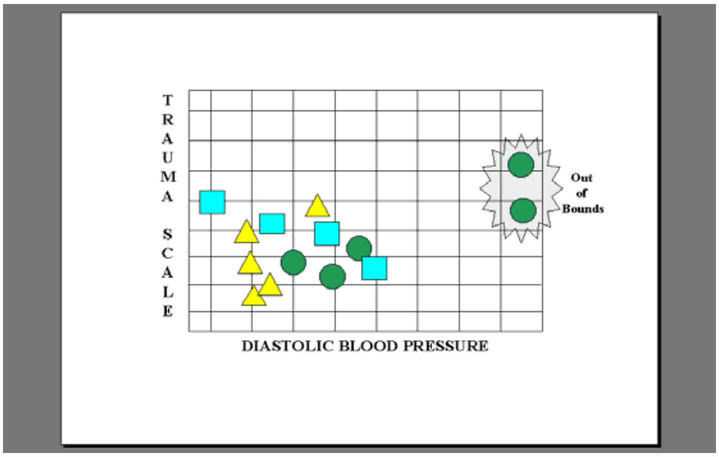
Strategy for Visualizing Objects and Their Associations



Scatter Plot

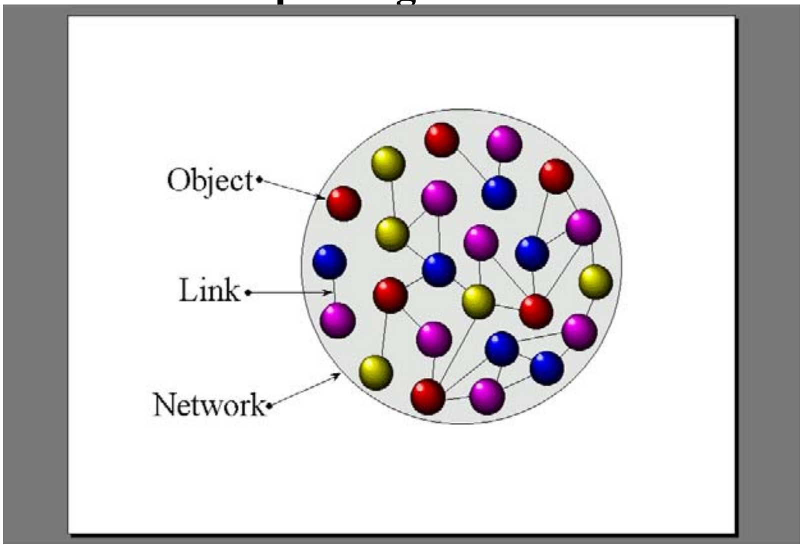


Scatter Plot

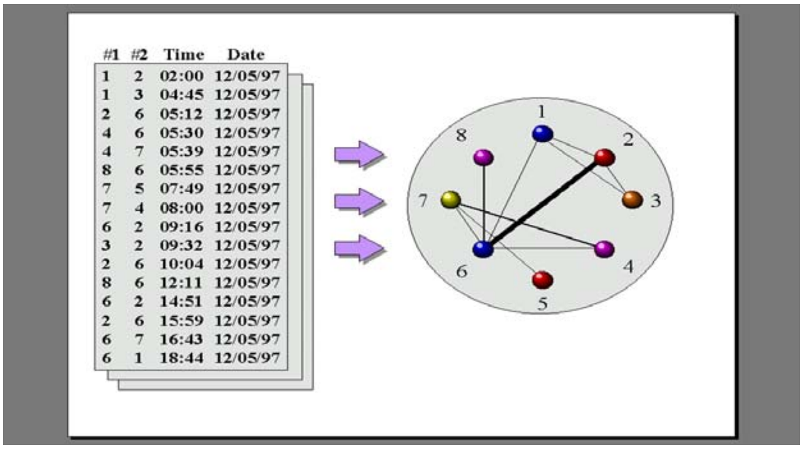


Shows out-of-bounds data signifying 'new' findings or corrupt data

Network is one popular visualization paradigm



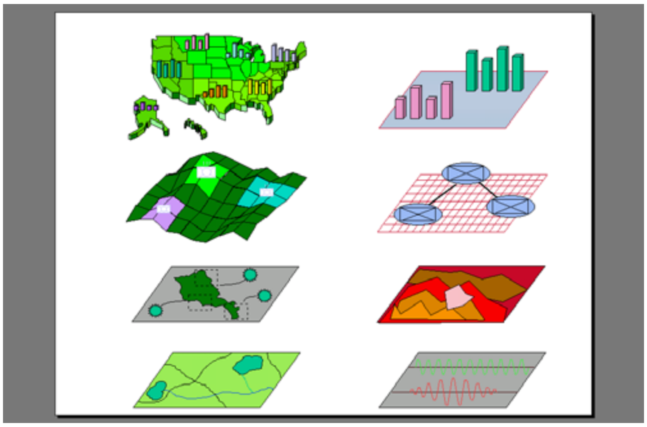
Link Analysis for Association



#1	#2	Time	Date
1	2	02:00	12/05/97
1	3	04:45	12/05/97
2	6	05:12	12/05/97
4	6	05:30	12/05/97
4	7	05:39	12/05/97
8	6	05:55	12/05/97
7	5	07:49	12/05/97
7	4	08:00	12/05/97
6	2	09:16	12/05/97
3	2	09:32	12/05/97
2	6	10:04	12/05/97
8	6	12:11	12/05/97
6	2	14:51	12/05/97
2	6	15:59	12/05/97
6	7	16:43	12/05/97
6	1	18:44	12/05/97

Visual Networks for Phone Call Data

Landscape Visualization for Association

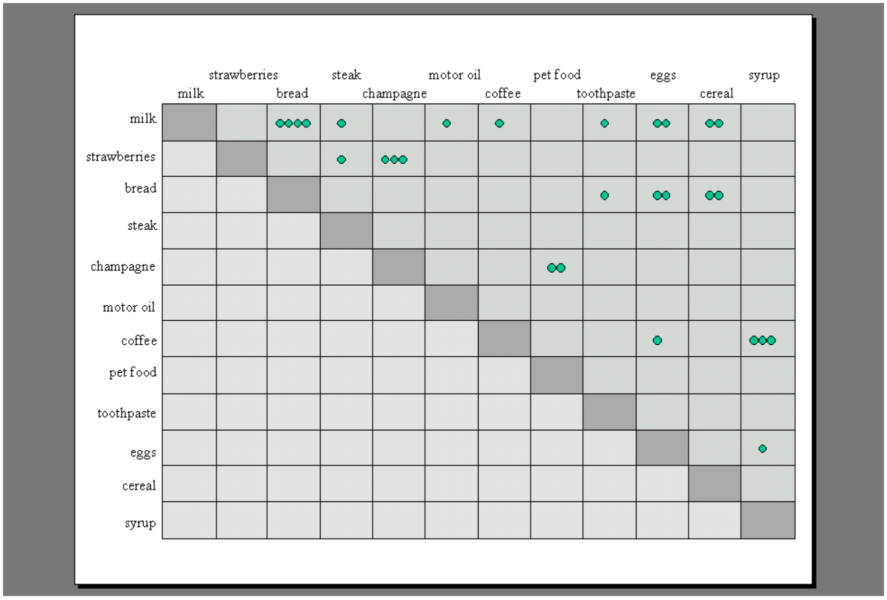


Exploring association between interest variables and their relative Cartesian positioning, such as geography

Analysis Methods for Discovering Association

- **Non-visual techniques**
  - Correlation analysis (can be done in Excel)
    - Are the variables nominal, ordinal, or continuous?
    - Interpret the strength of the correlation coefficient
  - Contingency tables
    - Cross-tabulate nominal variables (can be done by Pivot-table in Excel)
    - Examine the proportion of cases in each cell of the table
    - Use chi-square tests to assess significance

Association - Market Basket Analysis



	milk	strawberries	bread	steak	champagne	motor oil	coffee	pet food	toothpaste	eggs	cereal	syrup
milk			●●●●	●		●	●		●	●●	●●	
strawberries				●	●●●							
bread									●	●●	●●	
steak												
champagne								●●				
motor oil												
coffee										●		●●●
pet food												
toothpaste												
eggs												●
cereal												
syrup												

Two in-class examples by using Excel

- Grocery Point-of-sale data (very small set, calculation by hand)
  - – Discussion on how to know the "confidence" of the rule.
- Coffee store data (in coffee.xls)

Discovering Association

(L.O.57.1 finishes here)

- **Non-visual techniques continues …**
  - **Analysis of variance (ANOVA)**
    - Assess if there are mean differences in the dependent variable across *two or more predefined groups*

Data Mining Method #2

(L.O.57.2)

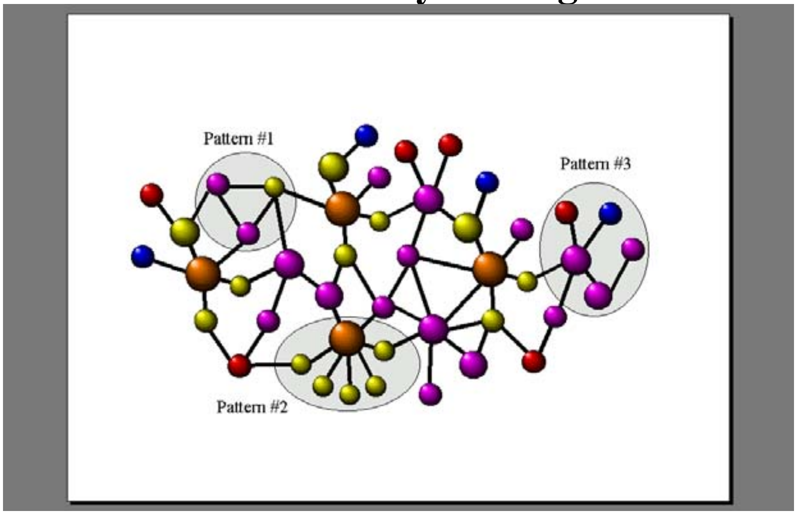
2. Discovering Sequential Pattern
   - Identify frequently occurring sequences from given records
     - 40 percent of female customers buy a gray skirt six months after buying a red jacket

Analysis Methods for Discovering Sequential Patterns

- Visual Methods
  - Link analysis
  - Temporal Patterns (Time based plots)

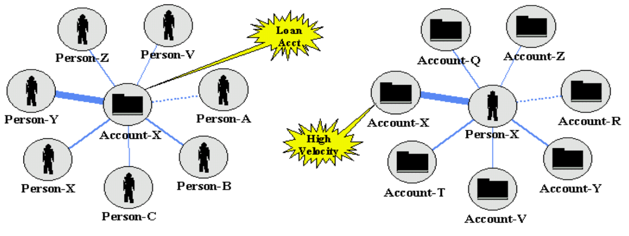
- Non-visual methods
  - Time-series analysis

Patterns from Link Analysis Diagram -



U.S. Government's secret data analyzed to find unusual patterns in the network structure (Kicker: data labels not known)

Patterns from Link Analysis Diagram -

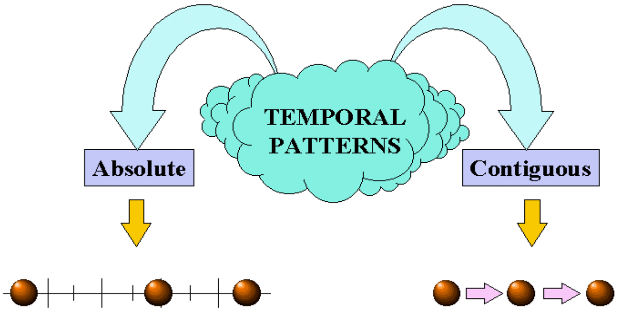


Loan Back Scheme

High Velocity Account

Intersection of account type and transaction velocity detects money laundering.

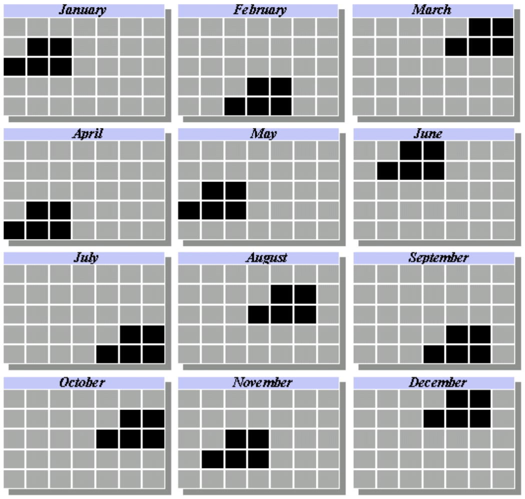
Discovering Temporal Patterns



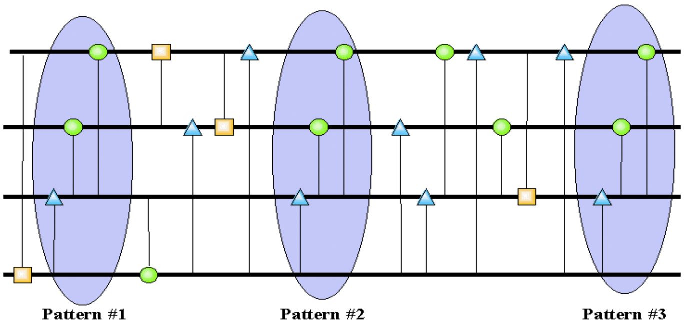
Actual Time Differences

Order of Occurrence

Absolute Time Cycle Events

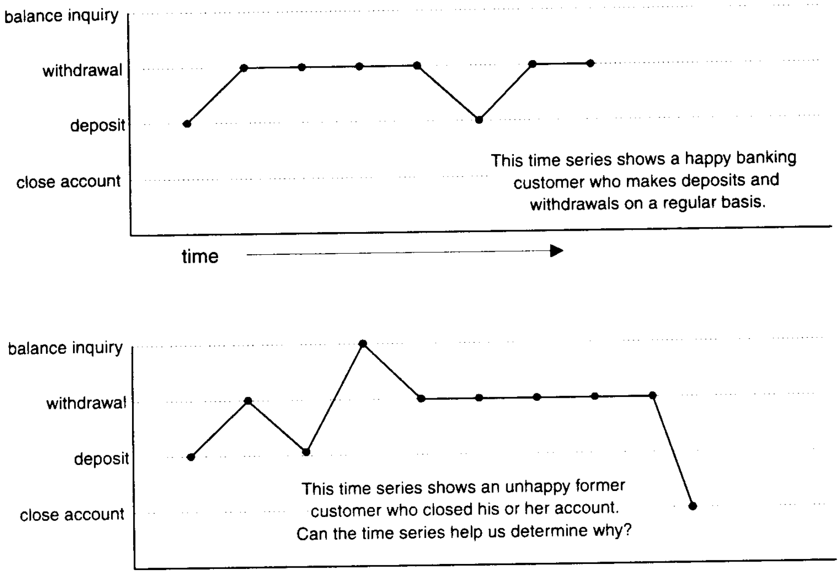


Contiguous Time Cycle Events



Finds *co-occurrence* of two or more events within a *non-standard* time interval

L.O.57.2 Finishes here.



This time series shows a happy banking customer who makes deposits and withdrawals on a regular basis.

This time series shows an unhappy former customer who closed his or her account. Can the time series help us determine why?

Figure 8.8 Time series provide snapshots of customer behavior through time.

Data Mining Method #3 (L.O.57.3)

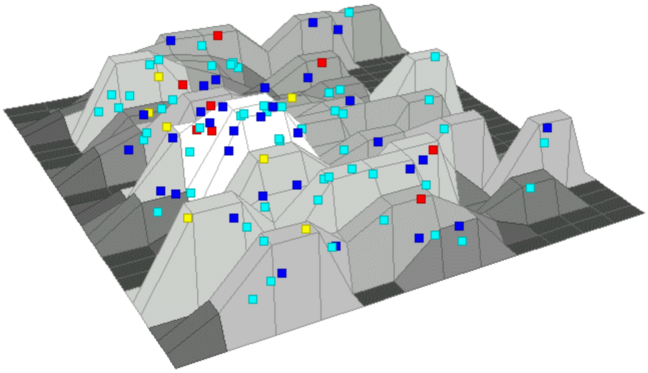
3. Classification
   - Identify ***a priori*** certain mutually exclusive classes
   - Identify a set of *meaningful* attributes that discriminate among the classes
   - Illustrations
     - Using a *meaningful* set of attributes, can we differentiate between frequent, moderate and infrequent customers?
     - Using a *meaningful* set of attributes, can we differentiate between repeat purchasers and one-time purchasers?

Analysis Techniques for Classification

- Neural networks
  - develops non-linear functions to associate inputs with outputs
  - no assumptions about distribution of data
  - handles missing data well (graceful degradation)
- Supervised neural networks
  - Estimating and testing the model
    - Construct a training sample and a holdout sample
    - Estimate model parameters using training sample
    - Test the estimated model's classification ability using holdout sample

Topographical Map Produced by an *Unsupervised* Learning Neural Network

(L.O.57.3 finishes here)



Data Mining Method #4

4. Visual Clustering

- Objects are assigned a place on the display based on general descriptive values and clustered around shared values.
- Positioning algorithms for
  - clustering (K Means method - can be done in SPSS)
  - self-organizing network

Analysis Methods for Clustering

(L.O. 57.4 finishes here)

- Non-visual methods
  - Cluster Analysis
  - Define *indicator* variables to define clusters on
    - income, age, education, etc.
  - Examine differences in clusters on key *criterion* variables
    - purchase loyalty, purchase behavior, etc
  - Do values of indicator and criterion variables vary systematically across clusters?

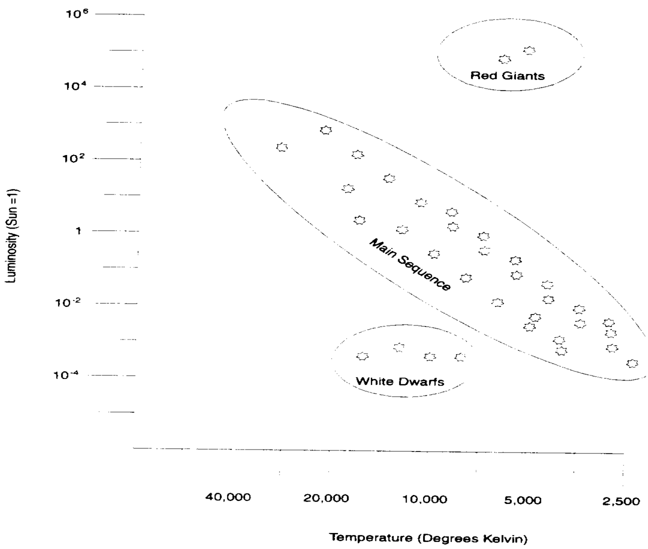


Figure 10.1 The Hertzsprung-Russell diagram clusters stars by temperature and luminosity.

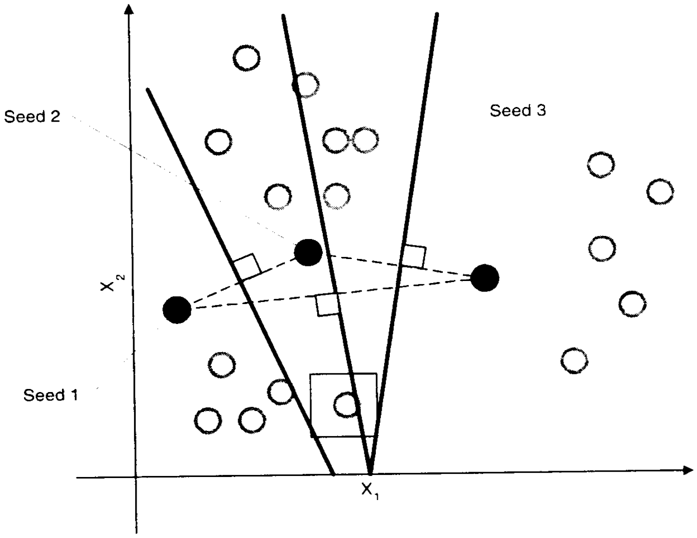


Figure 10.3 The initial seeds determine the initial cluster boundaries.

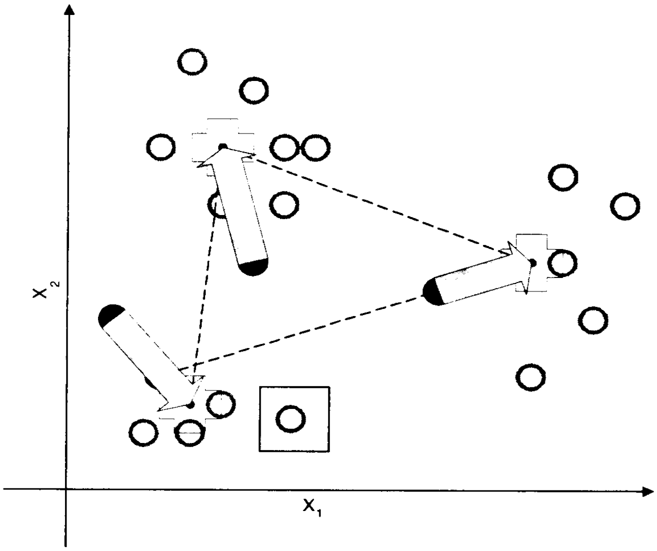
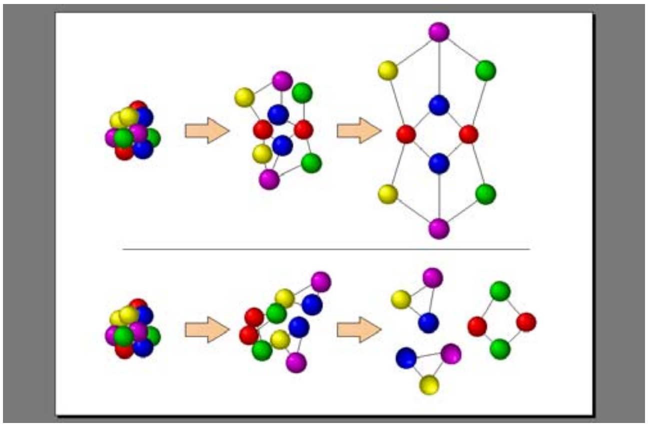
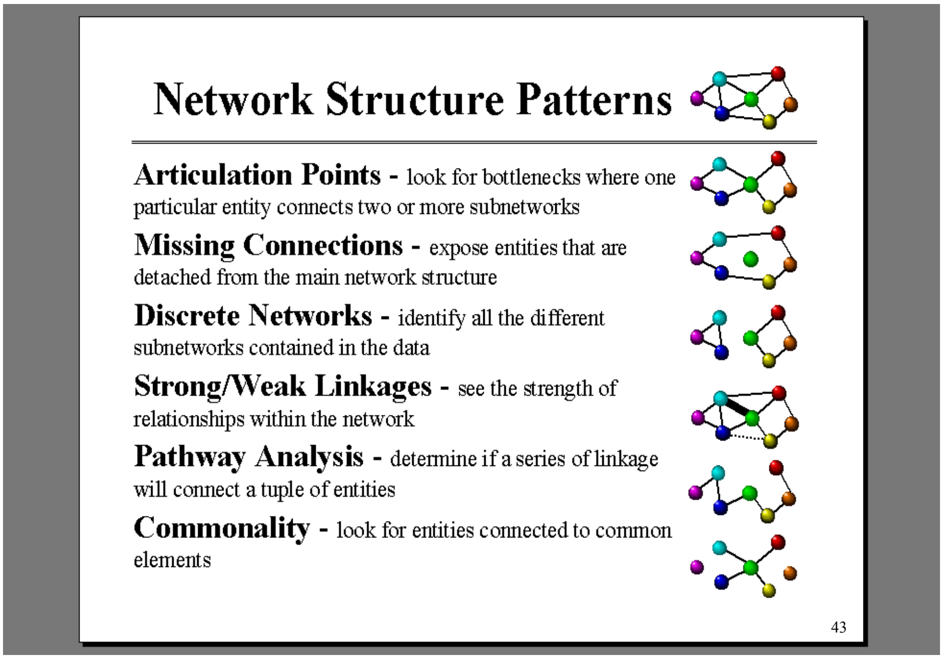


Figure 10.4 Calculating the centroids of the new clusters.

Self-Organizing Network





Network Structure Patterns

Articulation Points - look for bottlenecks where one particular entity connects two or more subnetworks

Missing Connections - expose entities that are detached from the main network structure

Discrete Networks - identify all the different subnetworks contained in the data

Strong/Weak Linkages - see the strength of relationships within the network

Pathway Analysis - determine if a series of linkage will connect a tuple of entities

Commonality - look for entities connected to common elements

Summary and Review

- What is data mining? What are its two main applications?
- Do you know how rules are created by Market Basket Analysis (MBA) ? Can you ‘compute’ a rule from a small set of example data?
- Are all rules useful? If not, why not?
- We have discussed four different data mining methods.
  - Do you know what they are and what kind of situations they are applicable for?