What is Data Mining?

DAMA-NCR

- Tuesday, November 13, 2001

- Laura Squier
- Technical Consultant
- lsquier@spss.com

Agenda

- What Data Mining IS and IS NOT
- Steps in the Data Mining Process
  - CRISP-DM
  - Explanation of Models
  - Examples of Data Mining Applications
- Questions

SPSS BI®

The Evolution of Data Analysis

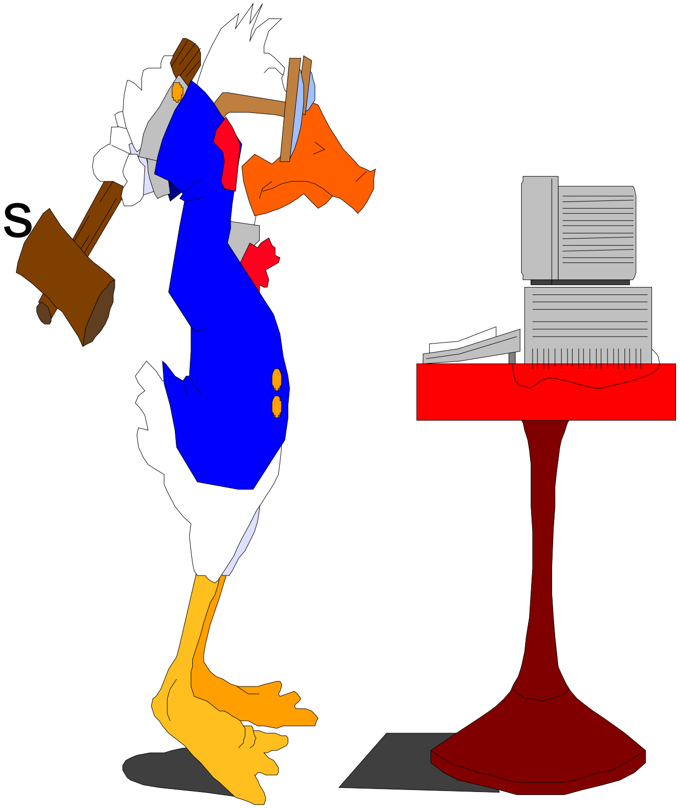
Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	SPSS, Comshare, Arbor, Cognos, Microstrategy,NCR	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	SPSS/Clementine, Lockheed, IBM, SGI, SAS, NCR, Oracle, numerous startups	Prospective, proactive information delivery

Results of Data Mining Include:

- Forecasting what may happen in the future
- Classifying people or things into groups by recognizing patterns
- Clustering people or things into groups based on their attributes
- Associating what events are likely to occur together
- Sequencing what events are likely to lead to later events

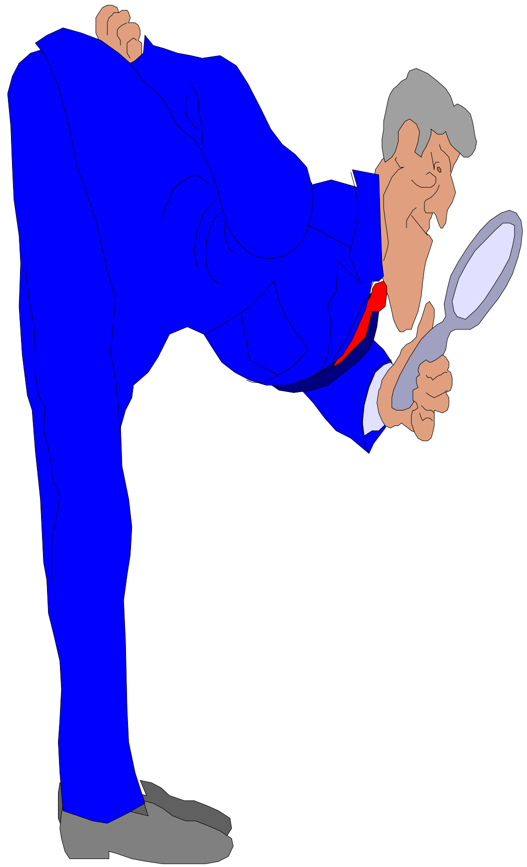
Data mining is <u>not</u>

- Brute-force crunching of bulk data
- “Blind” application of algorithms
- Going to find relationships where none exist
- Presenting data in different ways
- A database intensive task
- A difficult to understand technology requiring an advanced degree in computer science



SPSS BI

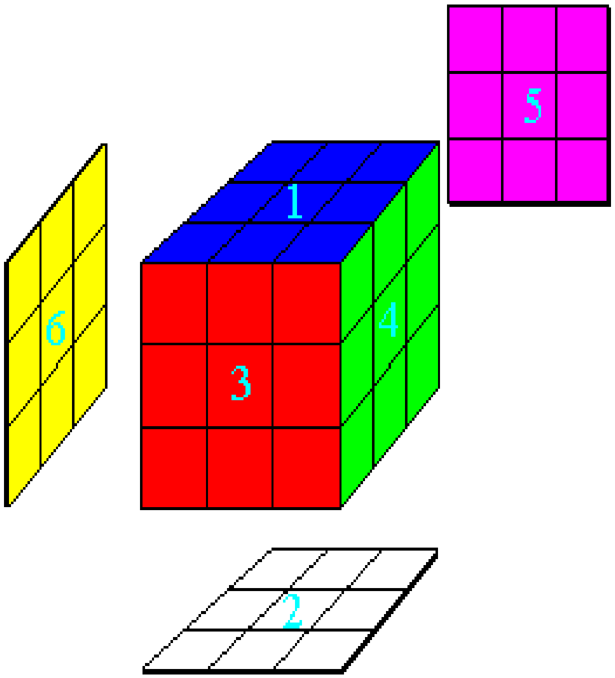
Data Mining Is



- A hot buzzword for a class of techniques that find patterns in data
- A user-centric, interactive process which leverages analysis technologies and computing power
- A group of techniques that find relationships that have not previously been discovered
- Not reliant on an existing database
- A relatively easy task that requires knowledge of the business problem/subject matter expertise

Data Mining versus OLAP

- OLAP - On-line Analytical Processing
  - Provides you with a very good view of what is happening, but can not predict what will happen in the future or why it is happening



Data Mining Versus Statistical Analysis

- Data Mining
  - Originally developed to act as expert systems to solve problems
  - Less interested in the mechanics of the technique
  - If it makes sense then let's use it
  - Does not require assumptions to be made about data
  - Can find patterns in very large amounts of data
  - Requires understanding of data and business problem

- Data Analysis
  - Tests for statistical correctness of models
    - Are statistical assumptions of models correct?
      - Eg Is the R-Square good?
  - Hypothesis testing
    - Is the relationship significant?
      - Use a t-test to validate significance
  - Tends to rely on sampling
  - Techniques are not optimised for large amounts of data
  - Requires strong statistical skills

Examples of What People are Doing with Data Mining:

- Fraud/Non-Compliance Anomaly detection
  - Isolate the factors that lead to fraud, waste and abuse
  - Target auditing and investigative efforts more effectively
- Credit/Risk Scoring
- Intrusion detection
- Parts failure prediction
- Recruiting/Attracting customers
- Maximizing profitability (cross selling, identifying profitable customers)
- Service Delivery and Customer Retention
  - Build profiles of customers likely to use which services
- Web Mining

How Can We Do Data Mining?

By Utilizing the CRISP-DM Methodology

- a standard process
- existing data
- software technologies
- situational expertise

Why Should There be a Standard Process?

The data mining process must be reliable and repeatable by people with little data mining background.

- **Framework for recording experience**
  - **Allows projects to be replicated**
- **Aid to project planning and management**
- **"Comfort factor" for new adopters**
  - **Demonstrates maturity of Data Mining**
  - **Reduces dependency on "stars"**

Process Standardization

CRISP-DM:

- CRoss Industry Standard Process for Data Mining
- Initiative launched Sept.1996
- SPSS/ISL, NCR, Daimler-Benz, OHRA
- Funding from European commission
- Over 200 members of the CRISP-DM SIG worldwide
  - DM Vendors - SPSS, NCR, IBM, SAS, SGI, Data Distilleries, Syllogic, Magnify, ..
  - System Suppliers / consultants - Cap Gemini, ICL Retail, Deloitte & Touche, …
  - End Users - BT, ABB, Lloyds Bank, AirTouch, Experian, ...

SPSS BI

CRISP-DM

- Non-proprietary
- Application/Industry neutral
- Tool neutral
- Focus on business issues
  - As well as technical analysis
- Framework for guidance
- Experience base
  - Templates for Analysis

CRISP-DM

CRoss Industry Standard Process for Data Mining

SPSS BI®

The CRISP-DM Process Model

CRISP-DM

CRoss Industry Standard Process for Data Mining

SPSS BI

Why CRISP-DM?

- The data mining process must be reliable and repeatable by people with little data mining skills

- CRISP-DM provides a uniform framework for
  - guidelines
  - experience documentation

- CRISP-DM is flexible to account for differences
  - Different business/agency problems
  - Different data

Phases and Tasks

Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Situation Assessment
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goal
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Asessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Data Set
Data Set Description

Select Data
Rationale for Inclusion / Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Modeling

Select Modeling Technique
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Description

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

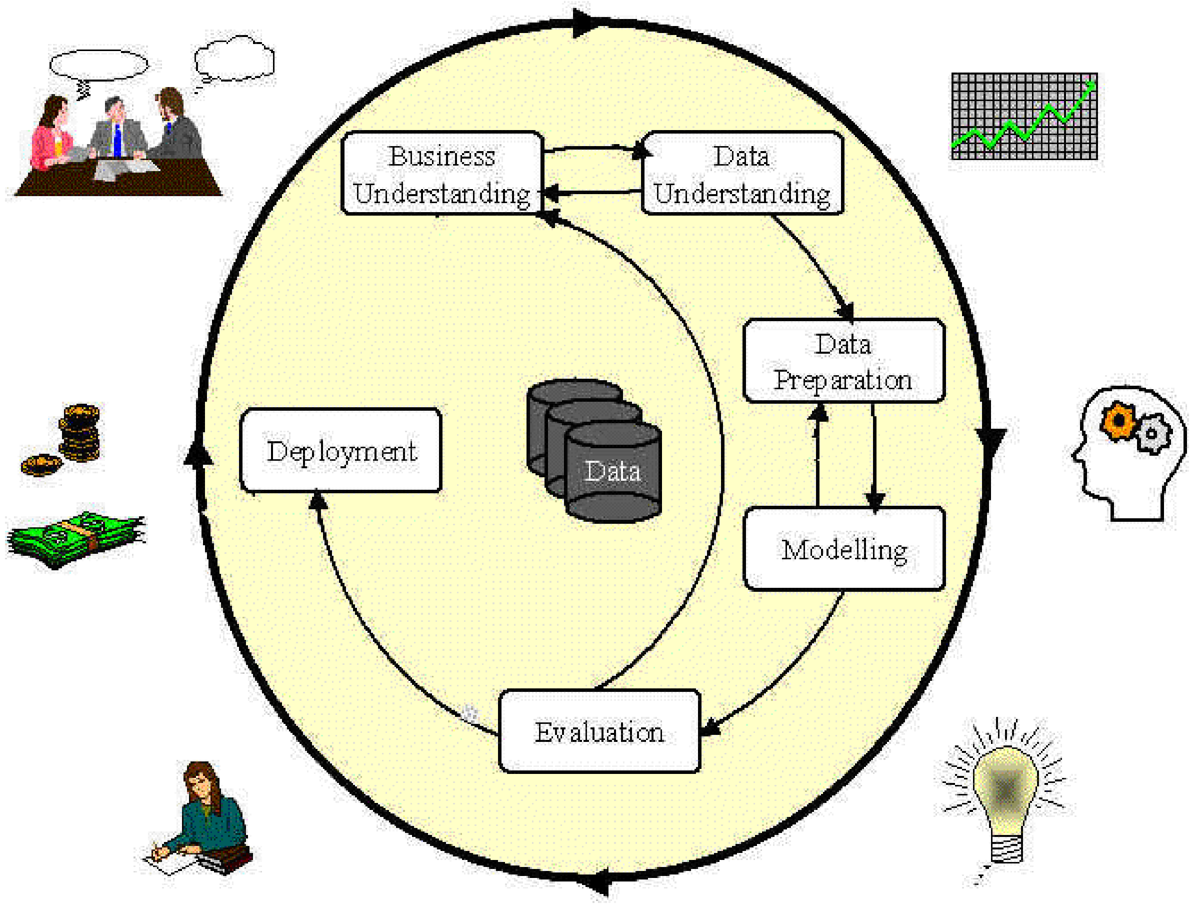
Plan Deployment
Deployment Plan

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

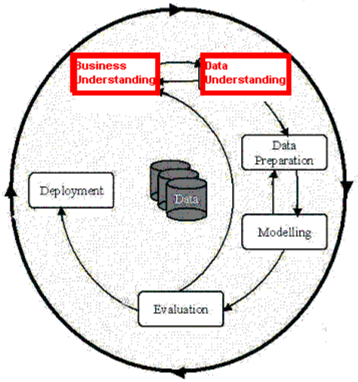
Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

Phases in the DM Process: CRISP-DM



Phases in the DM Process (1 & 2)

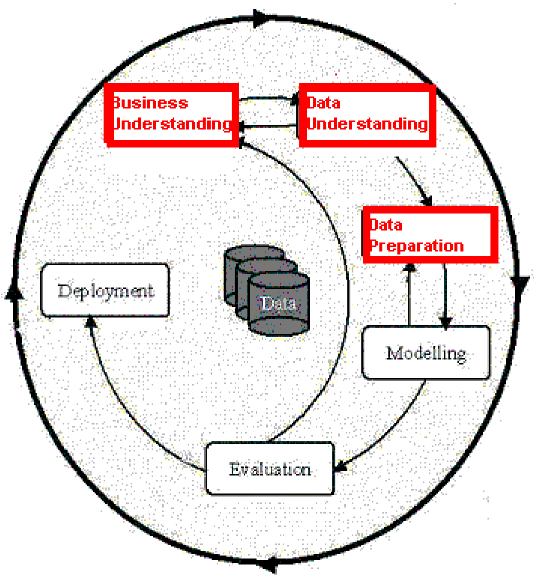


- Business Understanding:
  - Statement of Business Objective
  - Statement of Data Mining objective
  - Statement of Success Criteria

- Data Understanding
  - Explore the data and verify the quality
  - Find outliers

Phases in the DM Process (3)

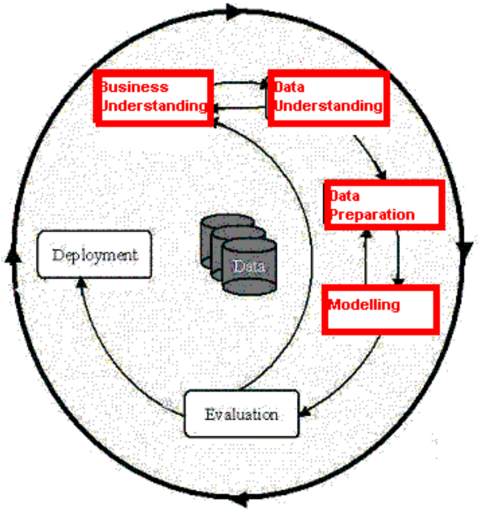
- **Data preparation:**
  - Takes usually over 90% of our time
    - Collection
    - Assessment
    - Consolidation and Cleaning
      - table links, aggregation level, missing values, etc
    - Data selection
      - active role in ignoring non-contributory data?
      - outliers?
      - *Use of samples*
      - visualization tools
    - Transformations - create new variables



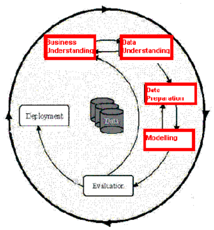
SPSS BI

Phases in the DM Process (4)

- **Model building**
  - **Selection of the modeling techniques is based upon the data mining objective**
  - **Modeling is an iterative process - different for *supervised* and *unsupervised learning***
    - **May model for either description or prediction**



Types of Models

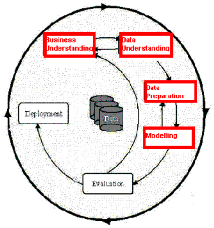


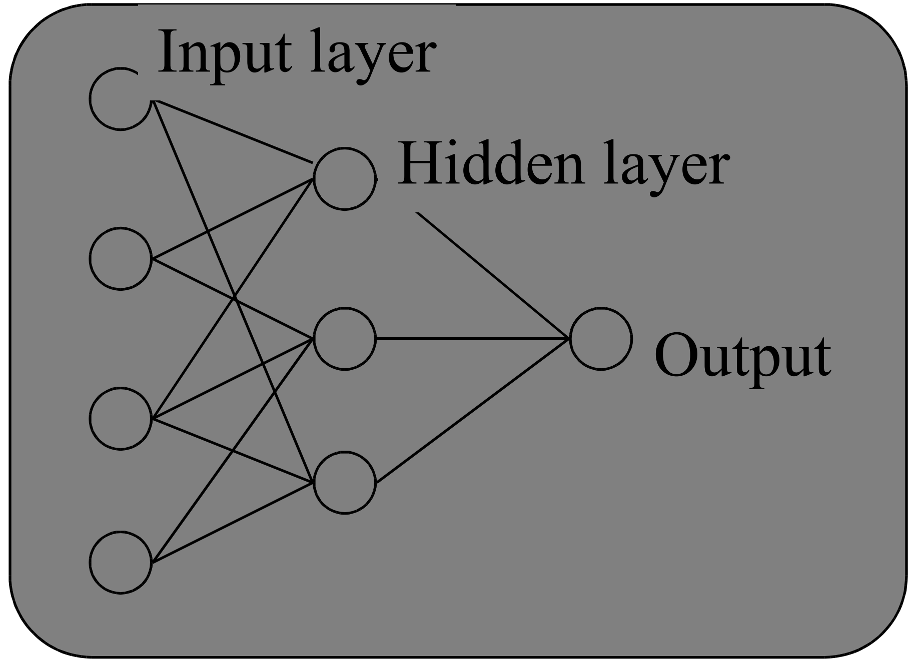
- **Prediction Models for Predicting and Classifying**
  - Regression algorithms (predict numeric outcome): neural networks, rule induction, CART (OLS regression, GLM)
  - Classification algorithm predict symbolic outcome): CHAID, C5.0 (discriminant analysis, logistic regression)

- **Descriptive Models for Grouping and Finding Associations**
  - Clustering/Grouping algorithms: K-means, Kohonen
  - Association algorithms: apriori, GRI

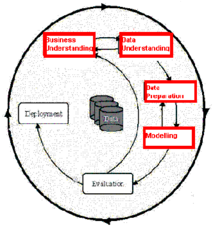
SPSS BI

Neural Network





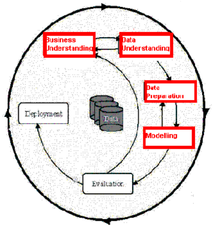
Neural Networks

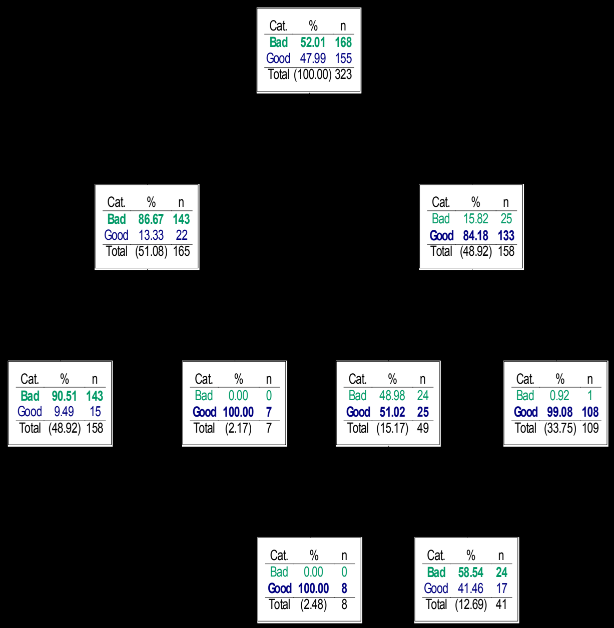


- Description
  - Difficult interpretation
  - Tends to ‘overfit’ the data
  - Extensive amount of training time
  - A lot of data preparation
  - Works with all data types

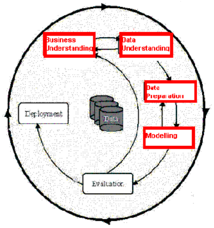
Rule Induction

- **Description**
  - Produces decision trees:
    - income < $40K
      - job > 5 yrs then *good risk*
      - job < 5 yrs then *bad risk*
    - income > $40K
      - high debt then *bad risk*
      - low debt then *good risk*
  - Or Rule Sets:
    - Rule #1 for good risk:
      - if income > $40K
      - if low debt
    - Rule #2 for good risk:
      - if income < $40K
      - if job > 5 years





Rule Induction



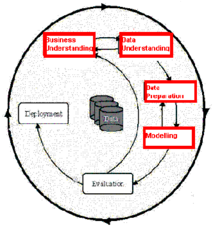
Description

- **Intuitive output**
- **Handles all forms of numeric data, as well as non-numeric (symbolic) data**

<u>C5 Algorithm</u> a special case of rule induction

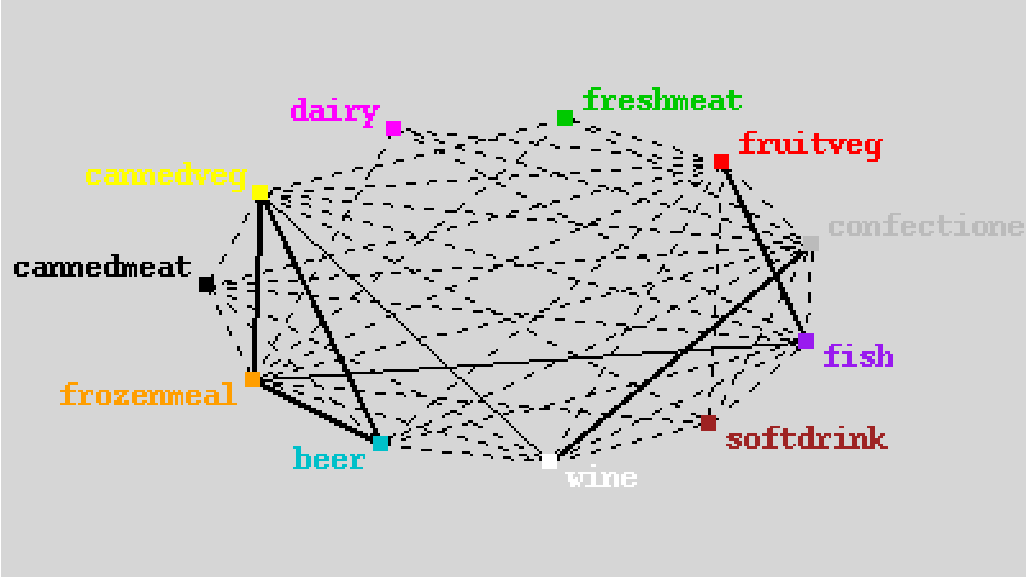
- **Target variable must be symbolic**

Apriori



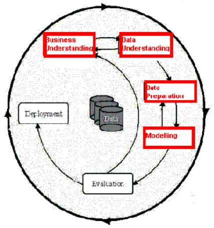
Description

- Seeks *association rules* in dataset
- ‘Market basket’ analysis
- Sequence discovery

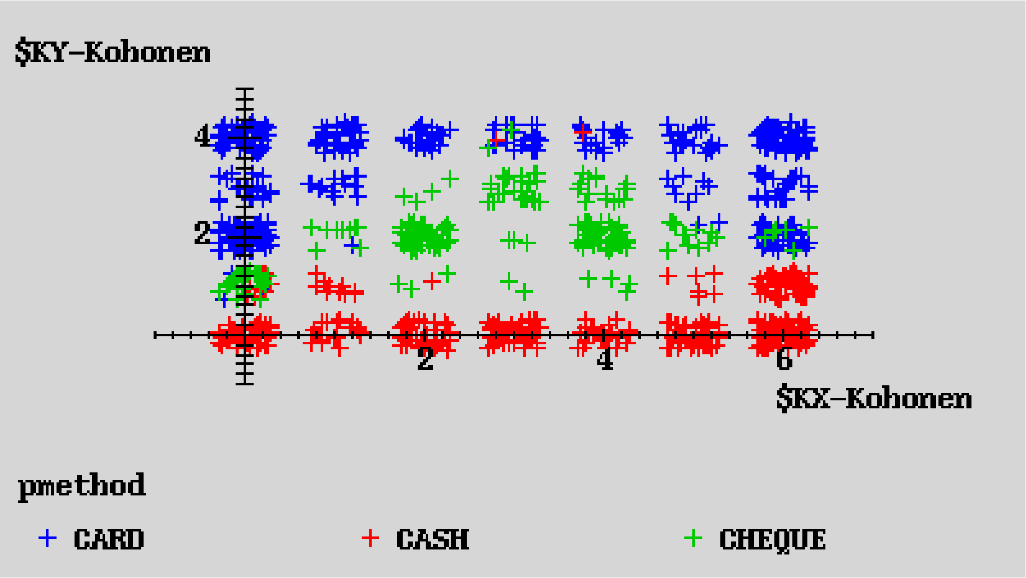


Kohonen Network

Description

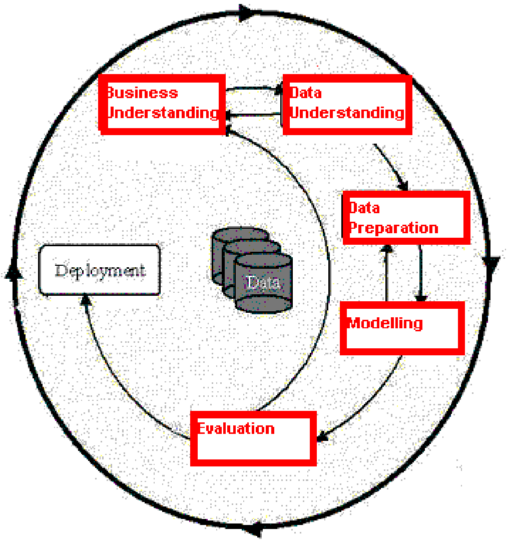


- unsupervised
- seeks to *describe* dataset in terms of natural *clusters* of cases

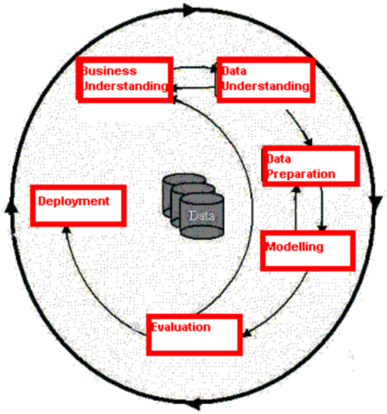


Phases in the DM Process (5)

- **Model Evaluation**
  – Evaluation of model: how well performed on test data
  – Methods and criteria depend on model type:
    - e.g., coincidence matrix with classification models, mean error rate with regression models
  – Interpretation of model: important or not, easy or hard depends on algorithm



Phases in the DM Process (6)



- Deployment
  - Determine how the results need to be utilized
  - Who needs to use them?
  - How often do they need to be used
- Deploy Data Mining results by:
  - Scoring a database
  - Utilizing results as business rules
  - interactive scoring on-line

Specific Data Mining Applications:

What data mining has done for...

IRS

The US Internal Revenue Service needed to improve customer service and...

Scheduled its workforce to provide faster, more accurate answers to questions.

SPSS BI

What data mining has done for...

The US Drug Enforcement Agency needed to be more effective in their drug “busts” and

analyzed suspects’ cell phone usage to focus investigations.

What data mining has done for... HSBC

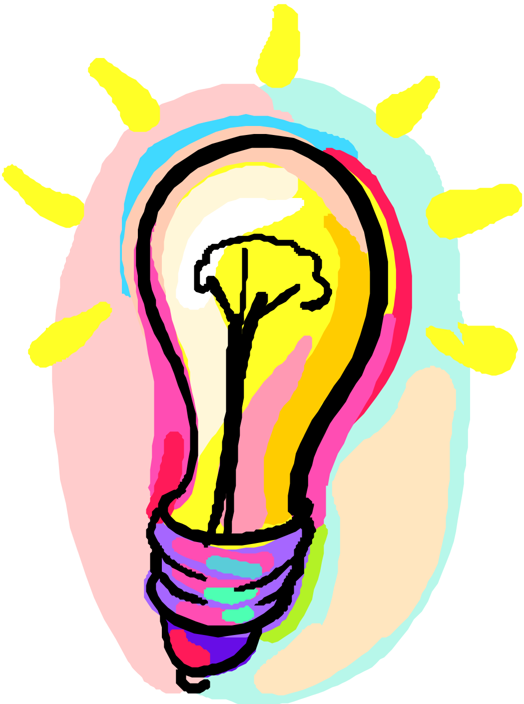
HSBC need to cross-sell more effectively by identifying profiles that would be interested in higher yielding investments and...

Reduced direct mail costs by 30% while garnering 95% of the campaign's revenue.

Final Comments

- Data Mining can be utilized in any organization that needs to find patterns or relationships in their data.
- By using the CRISP-DM methodology, analysts can have a reasonable level of assurance that their Data Mining efforts will render useful, repeatable, and valid results.

Questions?



SPSS@work

Ignatius Hospital

Ogilvy & Mather

FIRST UNION

FinCEN

Financial Crimes Enforcement Network

REUTERS

Deloitte & Touche

AmericanAirlines

BRITISH AIRWAYS

State of Georgia 1776

Constitution

Wisdom Justice Moderation

IRS

Aspen Publishers

Ameritech

In a world of technology, people make the difference.™

SIEMENS

Gus Home Shopping Group

Bell Atlantic

The Seal of the State of Washington 1889

Bell

Department of Defense

Defense Finance and Accounting Service

Unilever

HSBC

Sprint

FOX KIDS

BT

BARCLAYS

VANGUARD CELLULAR SYSTEMS, INC.

Southwestern Bell