Chapter 5: Data Mining

Data Mining Concepts/Definitions

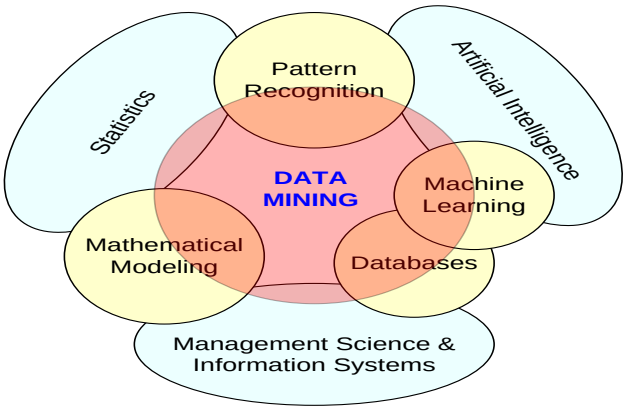
Why Data Mining?

- More intense competition at the global scale.
- Recognition of the value in data sources.
- Availability of quality data on customers, vendors, transactions, Web, etc.
- Consolidation and integration of data repositories into data warehouses.
- The exponential increase in data processing and storage capabilities; and decrease in cost.
- Movement toward conversion of information resources into nonphysical form.

Definition of Data Mining

- The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases. - *Fayyad et al., (1996)*
- Keywords in this definition: Process, nontrivial, valid, novel, potentially useful, understandable.
- Data mining: a misnomer?
- Other names: knowledge extraction, pattern analysis, knowledge discovery, information harvesting, pattern searching, data dredging,...

Data Mining is at the Intersection of Many Disciplines



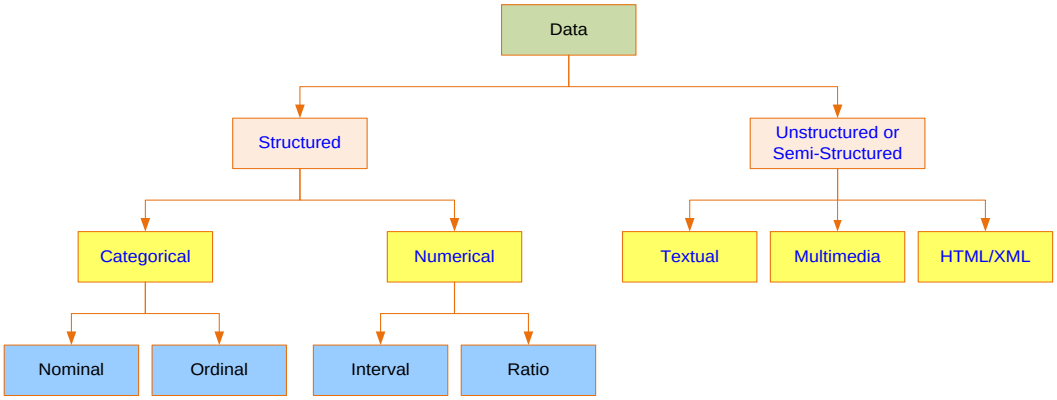
Data Mining Characteristics/Objectives

- Source of data for DM is often a consolidated data warehouse (not always!).
- DM environment is usually a client-server or a Web-based information systems architecture.
- Data is the most critical ingredient for DM which may include soft/unstructured data.

- The miner is often an end user
- Striking it rich requires creative thinking
- Data mining tools' capabilities and ease of use are essential (Web, Parallel processing, etc.)

Data in Data Mining

- Data: a collection of facts usually obtained as the result of experiences, observations, or experiments.
- Data may consist of numbers, words, images, ...
- Data: lowest level of abstraction (from which information and knowledge are derived).

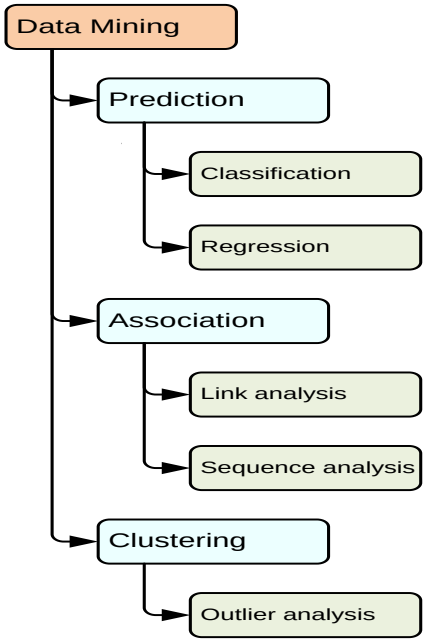


What Does DM Do?

How Does it Work?

- DM extract patterns from data
  - Pattern? A mathematical (numeric and/or symbolic) relationship among data items
    - Types of patterns
  - Association
  - Prediction
  - Cluster (segmentation)
  - Sequential (or time series) relationships

A Taxonomy for Data Mining Tasks



Learning Method	Popular Algorithms
Supervised	Classification and Regression Trees, ANN, SVM, Genetic Algorithms
Supervised	Decision trees, ANN/MLP, SVM, Rough sets, Genetic Algorithms
Supervised	Linear/Nonlinear Regression, Regression trees, ANN/MLP, SVM
Unsupervised	Apriory, OneR, ZeroR, Eclat
Unsupervised	Expectation Maximization, Apriory Algorithm, Graph-based Matching
Unsupervised	Apriory Algorithm, FP-Growth technique
Unsupervised	K-means, ANN/SOM
Unsupervised	K-means, Expectation Maximization (EM)

Data Mining Tasks (cont.)

- Time-series forecasting
  - Part of sequence or link analysis?
- Visualization
  - Another data mining task?
    - Types of DM
  - Hypothesis-driven data mining
  - Discovery-driven data mining

Data Mining Applications

- Customer Relationship Management
  - Maximize return on marketing campaigns
  - Improve customer retention (churn analysis)
  - Maximize customer value (cross-, up-selling)
  - Identify and treat most valued customers
    - Banking & Other Financial
  - Automate the loan application process
  - Detecting fraudulent transactions
  - Maximize customer value (cross-, up-selling)
  - Optimizing cash reserves with forecasting
- Retailing and Logistics
  - Optimize inventory levels at different locations
  - Improve the store layout and sales promotions

- Optimize logistics by predicting seasonal effects
- Minimize losses due to limited shelf life
- Manufacturing and Maintenance
- Predict/prevent machinery failures
- Identify anomalies in production systems to optimize the use manufacturing capacity
- Discover novel patterns to improve product quality

- Brokerage and Securities Trading
  - Predict changes on certain bond prices
  - Forecast the direction of stock fluctuations
  - Assess the effect of events on market movements
  - Identify and prevent fraudulent activities in trading
  - Insurance
  - Forecast claim costs for better business planning
  - Determine optimal rate plans
  - Optimize marketing to specific customers
  - Identify and prevent fraudulent claim activities

- Computer hardware and software
- Science and engineering
- Government and defense
- Homeland security and law enforcement
- Travel industry
- Healthcare
- Medicine

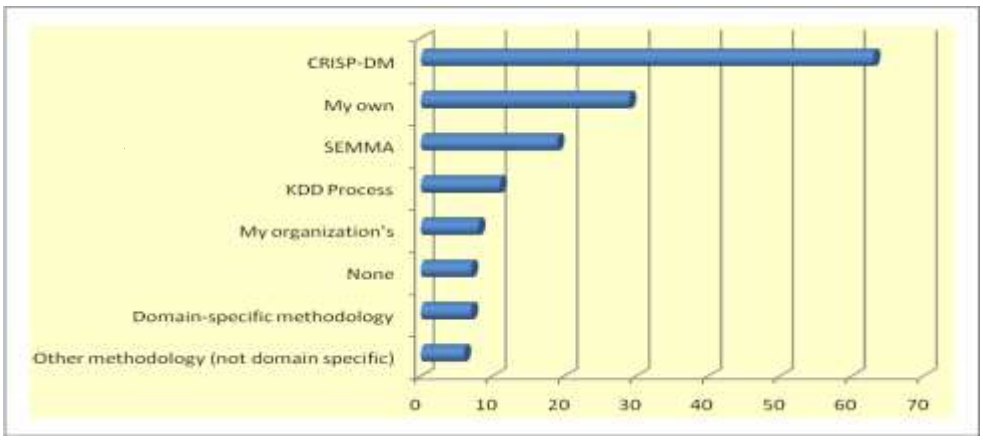
Increasingly more popular application areas for data mining

- Entertainment industry
- Sports
- Etc.

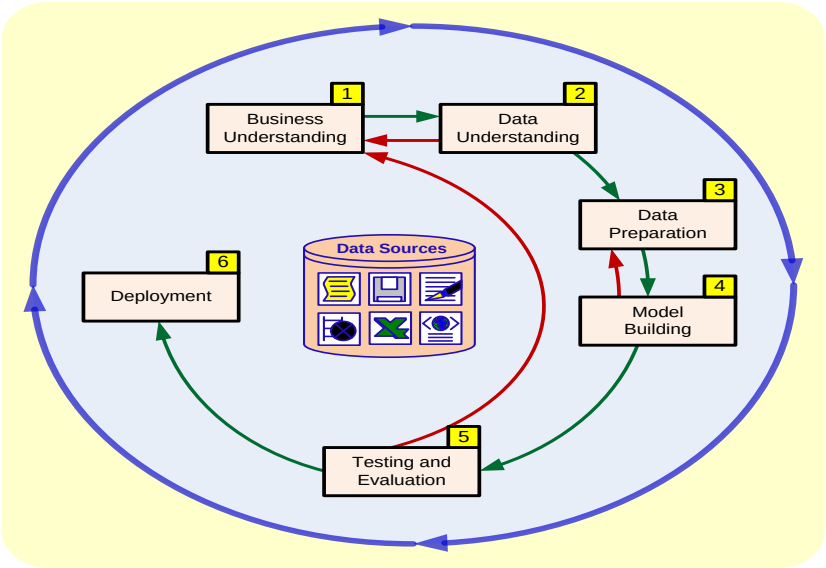
Data Mining Process

- A manifestation of best practices
- A systematic way to conduct DM projects
- Different groups has different versions
- Most common standard processes:
  - CRISP-DM (Cross-Industry Standard Process for Data Mining)
  - SEMMA (Sample, Explore, Modify, Model, and Assess)
  - KDD (Knowledge Discovery in Databases)

Data Mining Process



Data Mining Process: CRISP-DM



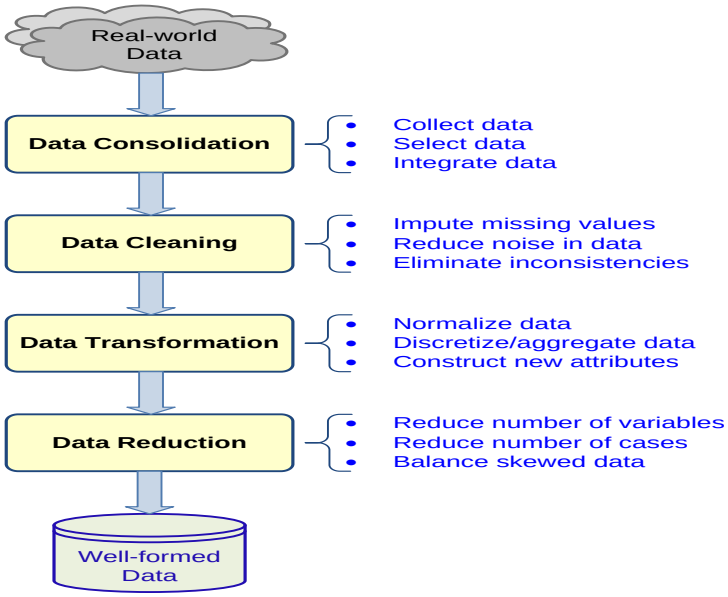
Data Mining Process: CRISP-DM

Step 1: Business Understanding
Step 2: Data Understanding
Step 3: Data Preparation (!)
Step 4: Model Building
Step 5: Testing and Evaluation
Step 6: Deployment

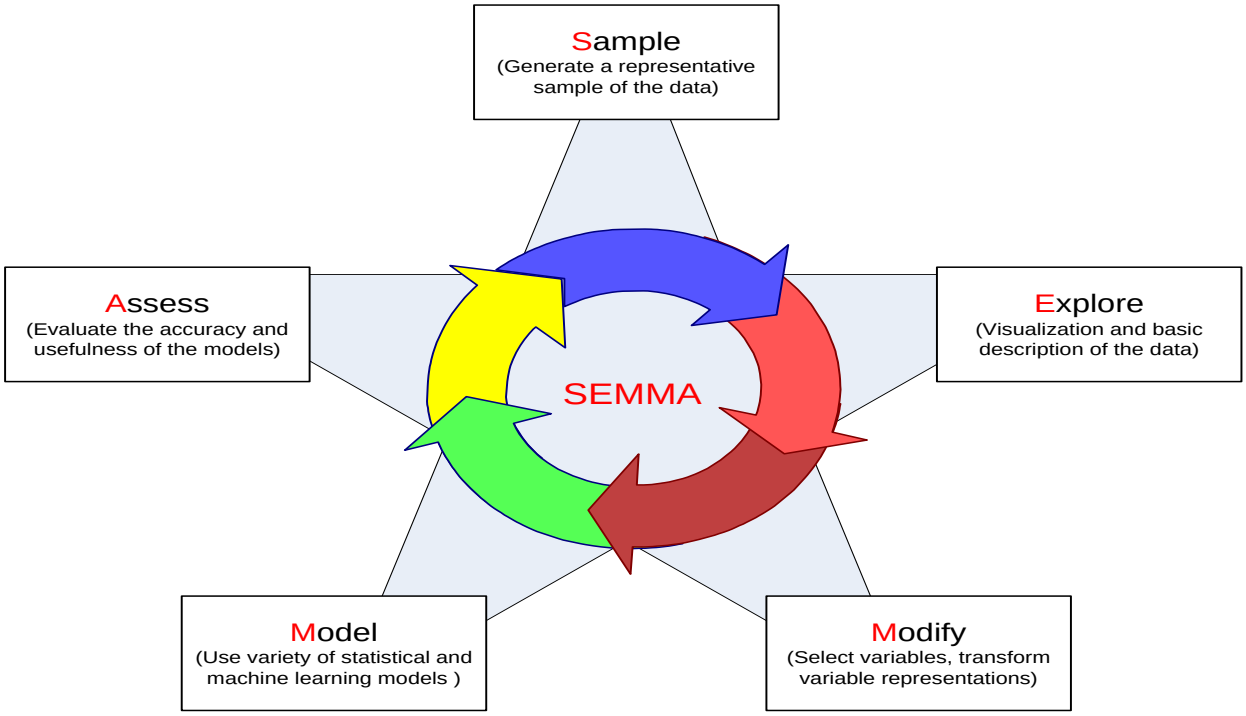
Accounts for ~85% of total project time

- The process is highly repetitive and experimental (DM: art versus science?)

Data Preparation – A Critical DM Task



Data Mining Process: SEMMA



Data Mining Methods: Classification

- Most frequently used DM method
- Part of the machine-learning family
- Employ supervised learning

- Learn from past data, classify new data
- The output variable is categorical (nominal or ordinal) in nature
- Classification versus regression?
- Classification versus clustering?

Assessment Methods for Classification

- Predictive accuracy
  - Hit rate
- Speed
  - Model building; predicting
- Robustness
- Scalability
- Interpretability
  - Transparency, explainability

Accuracy of Classification Models

- In classification problems, the primary source for accuracy estimation is the confusion matrix

		True Class: Positive	True Class: Negative
Predicted Class	Positive	True Positive Count (TP)	False Positive Count (FP)
Predicted Class	Negative	False Negative Count (FN)	True Negative Count (TN)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

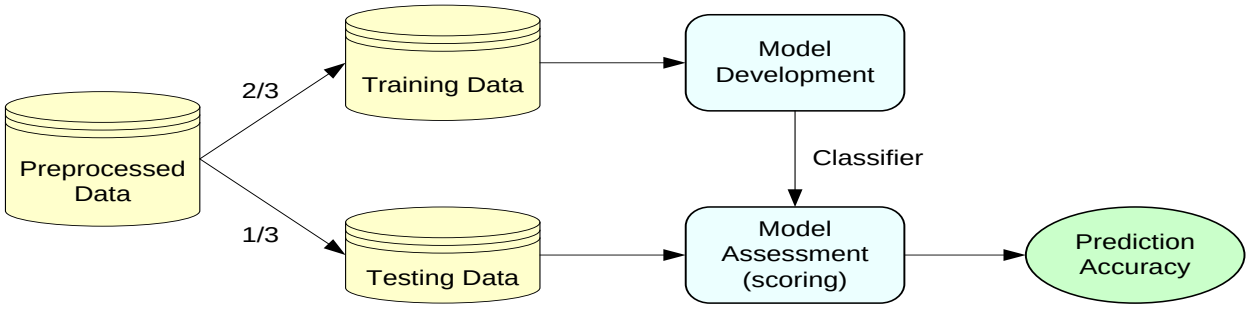
$$True\ Positive\ Rate = \frac{TP}{TP + FN}$$

$$True\ Negative\ Rate = \frac{TN}{TN + FP}$$

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

Estimation Methodologies for Classification

- Simple split (or holdout or test sample estimation)
  - Split the data into 2 mutually exclusive sets training (~70%) and testing (30%)

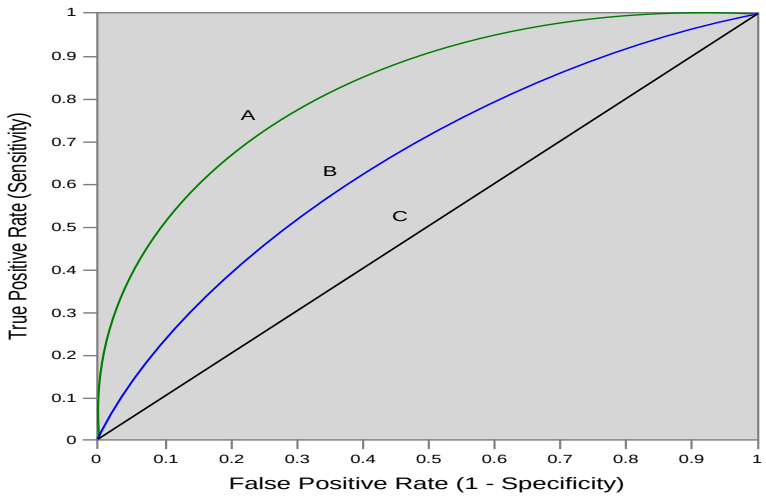


  - For ANN, the data is split into three sub-sets (training [~60%], validation [~20%], testing [~20%])

Estimation Methodologies for Classification

- *k*-Fold Cross Validation (rotation estimation)
  - Split the data into *k* mutually exclusive subsets
  - Use each subset as testing while using the rest of the subsets as training
  - Repeat the experimentation for *k* times
  - Aggregate the test results for true estimation of prediction accuracy training
- Other estimation methodologies
  - Leave-one-out, bootstrapping, jackknifing
  - Area under the ROC curve

Estimation Methodologies for Classification – ROC Curve



Classification Techniques

- Decision tree analysis
- Statistical analysis
- Neural networks
- Support vector machines
- Case-based reasoning
- Bayesian classifiers
- Genetic algorithms
- Rough sets

Decision Trees

- Employs the divide and conquer method
- Recursively divides a training set until each division consists of examples from one class

A general algorithm for decision tree building

1. Create a root node and assign all of the training data to it.
2. Select the best splitting attribute.
3. Add a branch to the root node for each value of the split. Split the data into mutually exclusive subsets along the lines of the specific split.
4. Repeat the steps 2 and 3 for each and every leaf node until the stopping criteria is reached.

Decision Trees

- DT algorithms mainly differ on
  1. Splitting criteria
     - Which variable, what value, etc.
  2. Stopping criteria
     - When to stop building the tree
  3. Pruning (generalization method)
     - Pre-pruning versus post-pruning
- Most popular DT algorithms include
  1. ID3, C4.5, C5; CART; CHAID; M5

Decision Trees

- Alternative splitting criteria
  - Gini index determines the purity of a specific class as a result of a decision to branch along a particular attribute/value
    - Used in CART
  - Information gain uses entropy to measure the extent of uncertainty or randomness of a particular attribute/value split
    - Used in ID3, C4.5, C5
  - Chi-square statistics (used in CHAID)

Cluster Analysis for Data Mining

- Used for automatic identification of natural groupings of things
- Part of the machine-learning family
- Employ unsupervised learning
- Learns the clusters of things from past data, then assigns new instances
- There is not an output variable

- Also known as segmentation

Cluster Analysis for Data Mining

- Clustering results may be used to
  - Identify natural groupings of customers
  - Identify rules for assigning new cases to classes for targeting/diagnostic purposes
  - Provide characterization, definition, labeling of populations
  - Decrease the size and complexity of problems for other data mining methods
  - Identify outliers in a specific domain (e.g., rare-event detection)

Cluster Analysis for Data Mining

- How many clusters?
  - There is not a "truly optimal" way to calculate it
  - Heuristics are often used
- Most cluster analysis methods involve the use of a distance measure to calculate the closeness between pairs of items.
  - Euclidian versus Manhattan/Rectilinear distance

Cluster Analysis for Data Mining

- *k*-Means Clustering Algorithm
  - *k* : pre-determined number of clusters
  - Algorithm (Step 0: determine value of *k*)

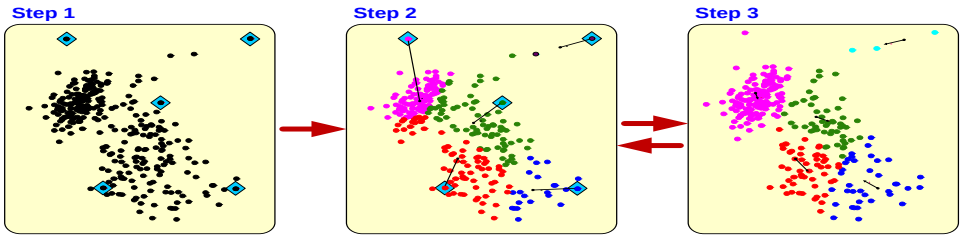
Step 1: Randomly generate *k* random points as initial cluster centers.

Step 2: Assign each point to the nearest cluster center.

Step 3: Re-compute the new cluster centers.

Repetition step: Repeat steps 3 and 4 until some convergence criterion is met (usually that the assignment of points to clusters becomes stable).

Cluster Analysis for Data Mining - *k*-Means Clustering Algorithm



Association Rule Mining

- A very popular DM method in business
- Finds interesting relationships (affinities) between variables (items or events)
- Part of machine learning family

- Employs unsupervised learning
- There is no output variable
- Also known as market basket analysis
- Often used as an example to describe DM to ordinary people, such as the famous "relationship between diapers and beers!"

Association Rule Mining

- A representative applications of association rule mining include
  - In business: cross-marketing, cross-selling, store design, catalog design, e-commerce site design, optimization of online advertising, product pricing, and sales/promotion configuration
  - In medicine: relationships between symptoms and illnesses; diagnosis and patient characteristics and treatments (to be used in medical DSS); and genes and their functions (to be used in genomics projects)

Association Rule Mining

- Are all association rules interesting and useful?

  A Generic Rule: **X ⇒ Y [S%, C%]**

  X, Y: products and/or services

  X: Left-hand-side (LHS)

  Y: Right-hand-side (RHS)

  S: Support: how often **X** and **Y** go together

  C: Confidence: how often **Y** go together with the **X**

  Example: {Laptop Computer, Antivirus Software} ⇒ {Extended Service Plan} [30%, 70%]
- Algorithms are available for generating association rules
  - Apriori
  - Eclat
  - FP-Growth
  - + Derivatives and hybrids of the three
- The algorithms help identify the frequent item sets, which are, then converted to association rules
- Apriori Algorithm
  - Finds subsets that are common to at least a minimum number of the itemsets
  - Uses a bottom-up approach
    - frequent subsets are extended one item at a time (the size of frequent subsets increases from one-item subsets to two-item subsets, then three-item subsets, and so on), and
    - groups of candidates at each level are tested against the data for minimum support.

    (*see the figure*) → --

Association Rule Mining

Apriori Algorithm

Raw Transaction Data

Transaction No	SKUs (Item No)
1	1, 2, 3, 4
1	2, 3, 4
1	2, 3
1	1, 2, 4
1	1, 2, 3, 4
1	2, 4

One-item Itemsets

Itemset (SKUs)	Support
1	3
2	6
3	4
4	5

Two-item Itemsets

Itemset (SKUs)	Support
1, 2	3
1, 3	2
1, 4	3
2, 3	4
2, 4	5
3, 4	3

Three-item Itemsets

Itemset (SKUs)	Support
1, 2, 4	3
2, 3, 4	3

Data Mining Software

- Commercial
  - IBM SPSS Modeler (formerly Clementine)
  - SAS - Enterprise Miner
  - IBM - Intelligent Miner
  - StatSoft – Statistica Data Miner
  - ... many more
- Free and/or Open Source
  - R
  - RapidMiner
  - Weka...

Data Mining Myths

- Data mining ...
  - provides instant solutions/predictions
  - is not yet viable for business applications
  - requires a separate, dedicated database
  - can only be done by those with advanced degrees
  - is only for large firms that have lots of customer data
  - is another name for the good-old statistics

Common Data Mining Blunders

1. Selecting the wrong problem for data mining
2. Ignoring what your sponsor thinks data mining is and what it really can/cannot do
3. Not leaving insufficient time for data acquisition, selection and preparation
4. Looking only at aggregated results and not at individual records/predictions
5. Being sloppy about keeping track of the data mining procedure and results
6. ...more in the book