Introduction to Data Mining

Jay Urbain

Credits: Nazli Goharian & David Grossman @ IIT

Outline

- Introduction
- Data Pre-processing
- Data Mining Algorithms
  - Naïve Bayes
  - Decision Tree
  - Neural Network
  - Association Rules
  - Clustering
  - Support Vector Machines
  - Graphical (Probabilistic) Models

Introduction

- Mining potential useful knowledge from a large amount of data.
- Other terminologies:
  - Knowledge Discovery in Databases (KDD) - although DM is a step in KDD.
  - Knowledge Extraction
  - Data Analysis
  - Information Harvesting
  - Information extraction
  - Text mining

Data Mining versus KDD

- Knowledge Discovery in Databases (KDD) is the process of finding useful information and patterns in the data.
- Data Mining is the use of algorithms to find the useful information in the KDD process.

- KDD process is:
  - Data cleaning & integration (Data Pre-processing)
  - Creating a common data repository for all sources, such as data warehouse.
  - Data mining
  - Visualization for the generated results

Data Mining versus DB

- DB’s user knows what is looking for.
- DM’s user might/might not know what is looking for.
- DB’s answer to query is 100% accurate, if data correct.
- DM’s effort is to get the answer as accurate as possible.
- DB’s data are retrieved as stored.
- DM’s data need to be cleaned (some what) before producing results.
- DB’s results are subset of data.
- DM’s results are the analysis of the data.
- The meaningfulness of the results is not the concern of Database but it is the main issue in Data Mining.

Data Mining Applications

- Fraud Detection and Risk Analysis
  - Credit card fraud detection
  - Money laundry detection
  - Risk of loan payment
  - etc….
- Retail
  - Sale and Promotion
  - Coupon
  - etc…
- Stock Market Analysis

Data Mining Applications

- Identifying Criminals & Profiling terrorists
- Flood Prediction
- Telecommunications
- Medical Diagnosis & Treatment
- Biomedical & DNA Data Analysis
  - Which genes co-occur with other genes?
  - What are the sequence of genetic activities in stages of a disease?
- Web Mining
  - What are associations among different pages?
  - What are web page characteristics?
  - What is the distribution of information on web?

Data Mining Applications

- Text mining
  - Identifying named entities (concepts) in text.
  - Finding relationships between concepts.
  - Building graphs of concept relations.
  - E.g.:
    - protein function graphs
    - Gene expression
    - Identifying authors, scientists with expertise

Privacy Issues

- DM applications derive demographics about customers via
  - Credit card use
  - Store card
  - Subscription
  - Book, video, etc rental
  - Phone and web usage
  - Cell phone geographic information
  - and via more sources…
- As the DM results are deemed to be a good estimate or prediction, one has to be sensitive to the results not to violate privacy.

Commercial tools

- Problem: No a common model/ architecture, but there are initiatives – ACM SIGKDD.
  - Accessing different but not necessarily all type of data repositories.
  - Supporting one or more of the DM algorithms.
  - May/may not supporting all data types.
  - Supporting different but not all functionalities.
  - platform dependant.
  - => Each application might work with one commercial tool and not with the other tool.

Commercial tools

- Darwin (Oracle Corp.)
- MineSet (Silicon Graphics Inc. - SGI)
- Intelligent Miner (IBM Corp)
- Enterprise Miner (SAS Institute Inc.)
- Clementine (SPSS Inc – Integral Solutions)
- DMMiner (DBMiner Technology Inc.)
- Business Objects

Commercial tools

- BrainMaker (California Scientific Software)
- CART (Salford Systems)
- MARS (Salford Systems)
- Scenario (Cognos Inc.)
- Web Analyst (Megaputer Intelligence Inc.)
- SurfAid Analysis (IBM corp)
- Visualizer Workstation (Computer Science Innovations, Inc)
- etc….

Data sources

- Relational Database
- Data Warehouse
- Flat Files
- Web
- Object Oriented database
- Multi Media
- Docuemnts, journals, etc.

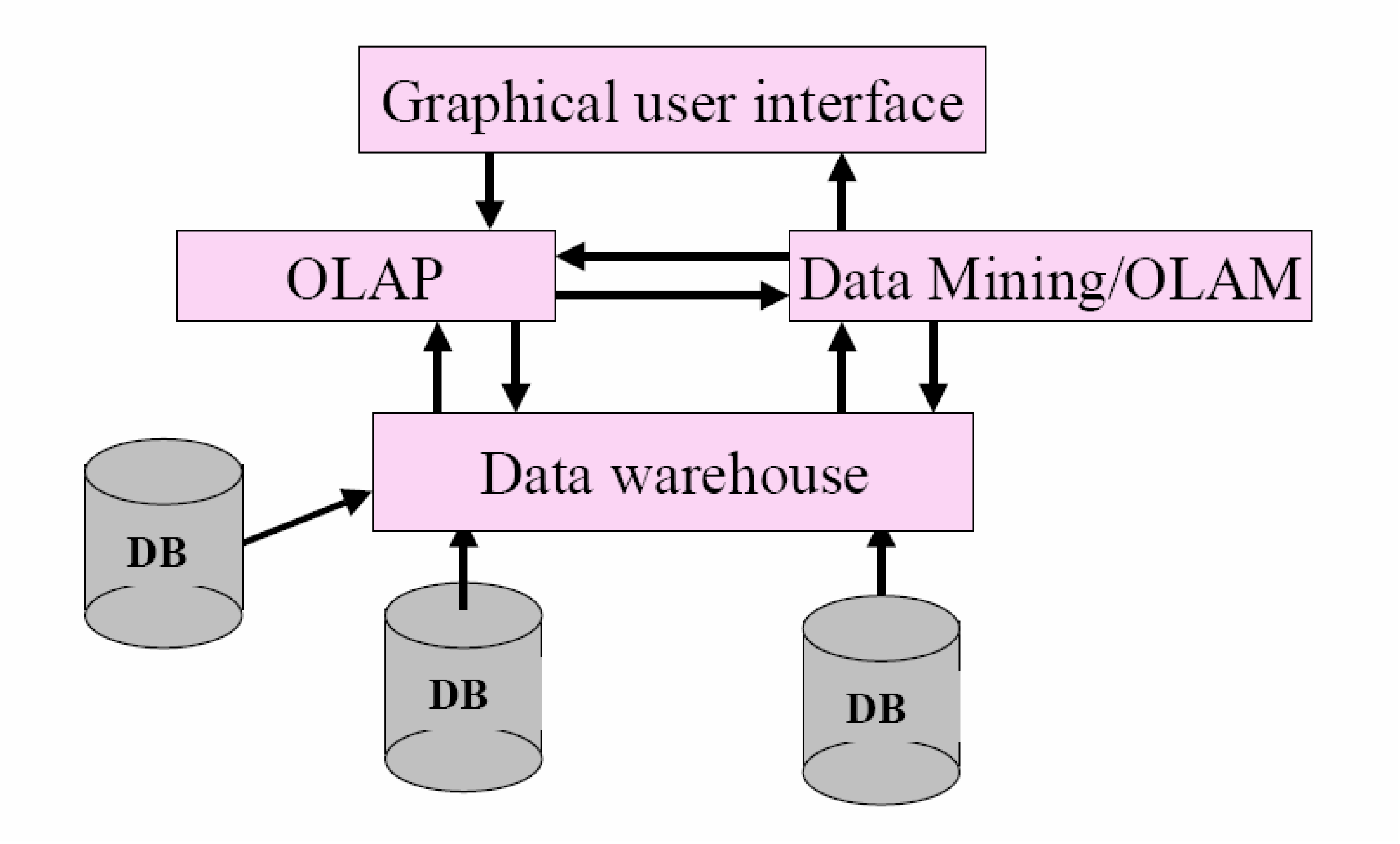
Data warehouse (cont.)

- To improve the performance in DW different techniques such as *Summarization* and *Denormalization* are used.
- Usually but not always DW is accessed by On-Line Analytical Processing (OLAP) query.
- SQL gives a precise answer to a user query.
- OLAP gives a multi-dimensional view of data and is as extension of some aggregate functions in SQL.
- OLAP Operations are Slice, Dice, Roll-up, and Drilldown.

Data warehouse (cont.)

- OLAP is a data summarization/ aggregation tool that facilitates the data analysis for the user by providing a multi-dimensional view of the data.
- Data Mining Tool provides an automated discovery of knowledge and gives more in-depth knowledge about data and hidden information.
- OLAM (OLAP Mining) is the integration of OLAP with Data Mining.

Data warehouse (cont.)



DM and Other Disciplines

- Statistical Concepts
  - Bayes Theorem
  - Graphical Models
  - Regression
  - etc…

- Machine Learning
  - Support Vector Machines
  - Neural Network
  - Genetic Algorithm
  - Clustering
  - Association Rule

Scalability

- Statistical approach deal with small data sets.
  - Believe that all data must be cleaned and reduced.
- Machine Learning deal with small data sets.
  - Goal is to make machine learn.
  - Applications such as Chess Playing rather than applications that deal market analysis.
- Real life data to be mined is huge, thus need scalable algorithms.
- Need to compromise on scalability versus algorithm.

DM Algorithms

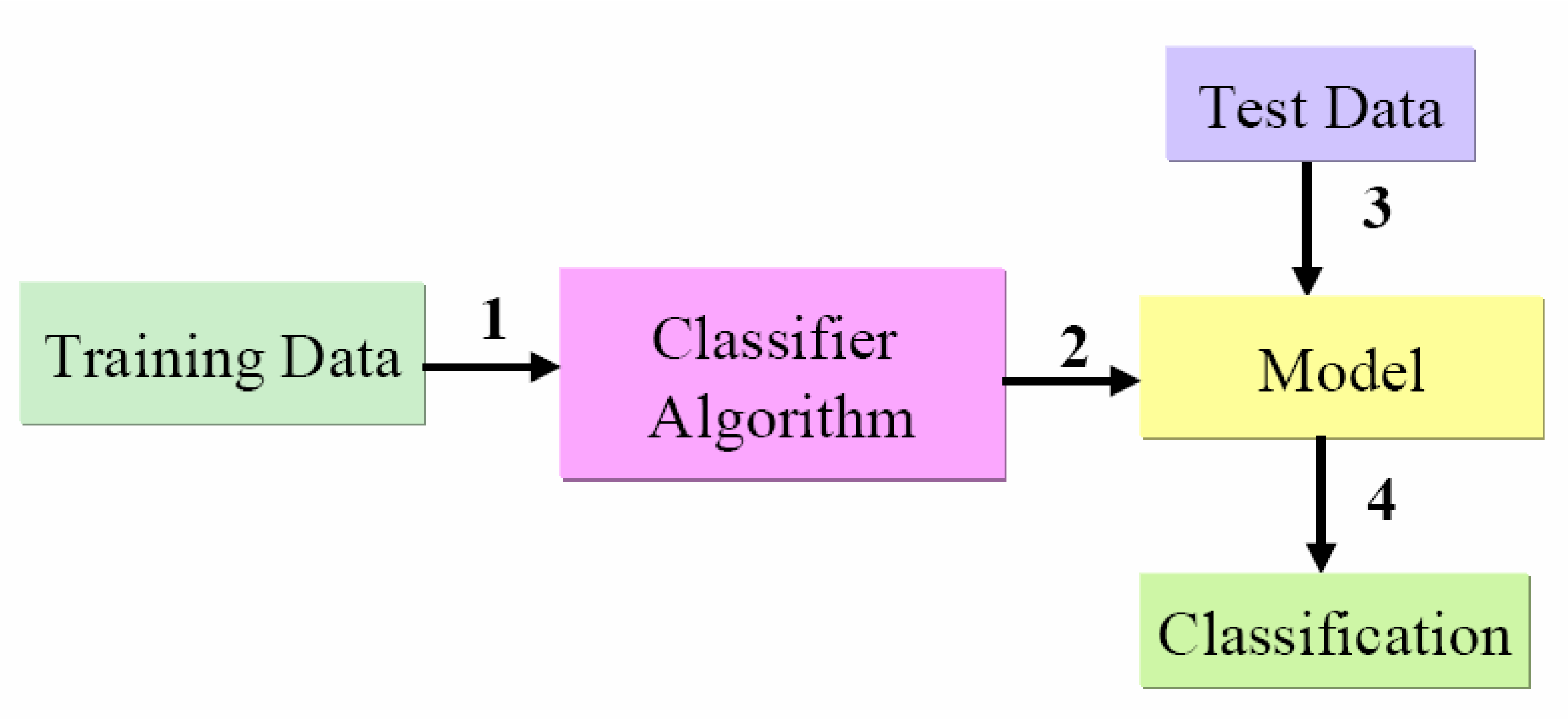
- Supervised (Classification)
  - Bayesian
  - Neural Network
  - Decision Tree
  - Others: Genetic Algorithms, Fuzzy Set, K Nearest Neighbor
- Unsupervised
  - Association Rules
  - Clustering

Supervised vs. Unsupervised

- Supervised algorithms
  - Learning by example:
  - Use training data which has correct answers (class label attribute)
  - Create a model by running the algorithm on the training data
  - Identify a class label for the incoming new data

- Unsupervised algorithms
  - Do not use training data.
  - Classes may not be known in advance.

Supervised Algorithms



DM Evaluation Metrics

- Not always straightforward.

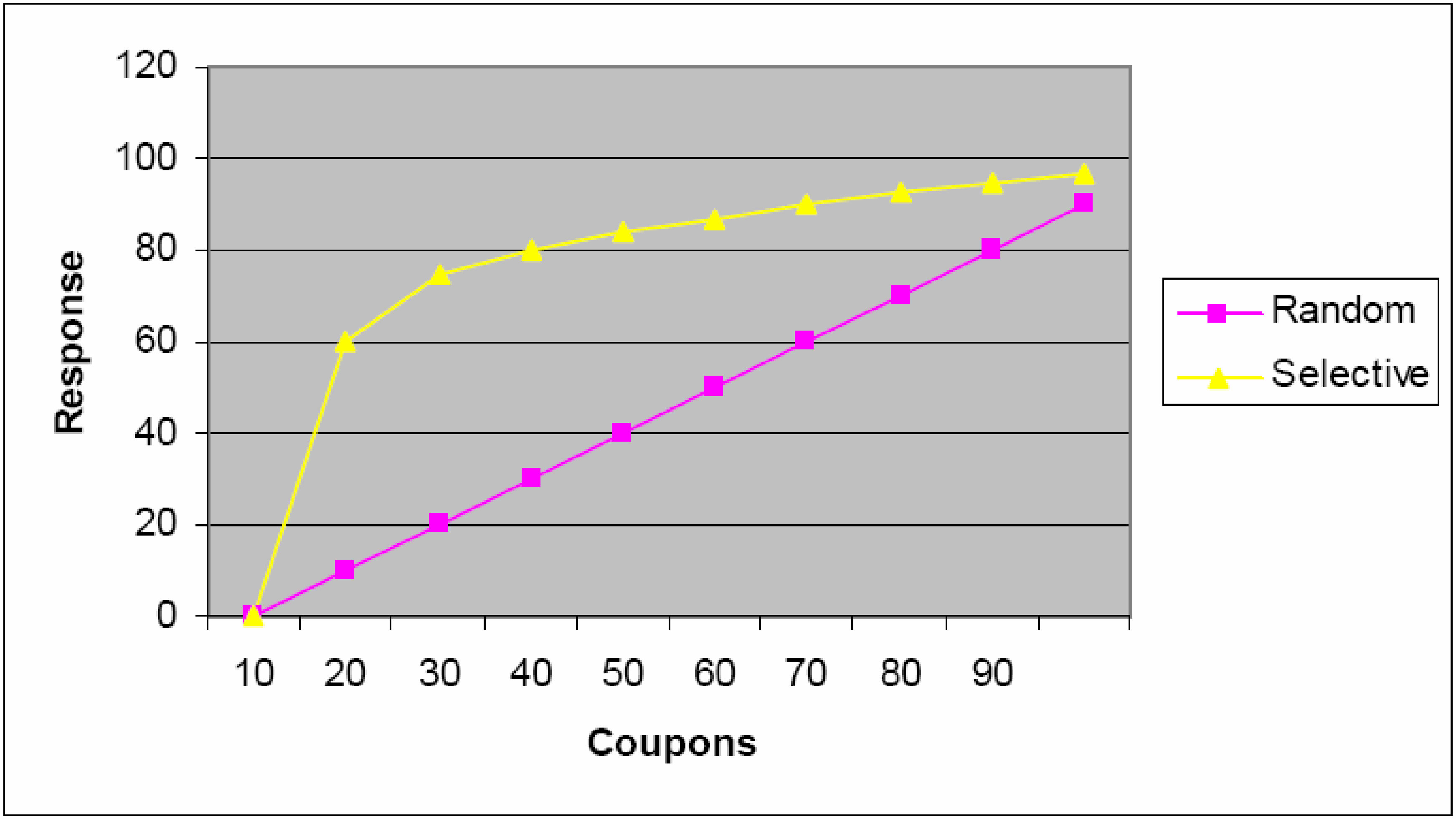
- ROI (Return on Investment) used in business to measure
- benefit of using Data Mining.

- Lift Chart used to visualize and measure response modeling performance.

- Traditional Computer Science evaluation metrics are ***space*** requirement and ***time complexity*** to compare the algorithms.

- Measuring accuracy of DM results:
  - Use of Cross-Validation in Supervised algorithms.
  - Information Retrieval measures of Precision & recall.
  - Various accuracy measures based on each algorithm.

Lift Chart


120
100
80
60
40
20
0
Response
10
20
30
40
50
60
70
80
90
Coupons
Random
Selective

Cross Validation

- Goal: We want to measure the effectiveness of a classification (supervised) algorithm.
  - Take the training dataset
  - Build model
  - Test using the training dataset

- This usually leads to a very *optimistic* result that has little ability to predict the real accuracy of the model.

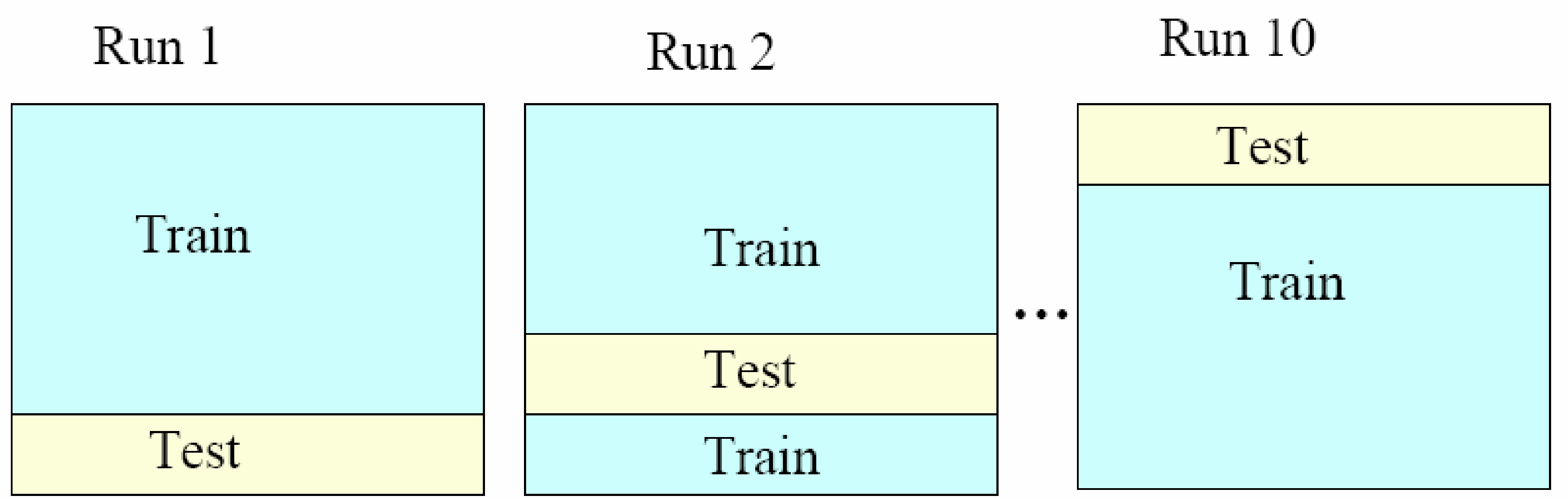
Cross Validation

- Another approach is to take the training data set, cut it in half and use half on training and half for testing.

- This leads to potential errors in estimating the real classification rate because the half we hold for testing may be very different that the half we used for training.

10-fold Cross Validation

- Take the data set and use the first 90 percent of the data for training and then test on the final ten percent. Then use the next 10 percent for testing, etc.



10-fold Cross Validation

- Each run will result in a particular classification rate.
- Ex: If we classified 50/100 of the test records correctly our classification rate for that run is 50%.
- Choose the model that generated the highest classification rate. The final classification rate for the model is the average of the ten classification rates.

Summary

- Data Mining algorithms are used to detect the information that we do not know.
- There are various data sources, types,formats and applications for Data Mining.
- Usually Data Mining is used on a Data Warehouse.
- There are many Data Mining algorithms.
- Scalability of Data Mining differentiates it from statistical and Machine Learning approach, as in Data Mining we deal with huge amount of data.