How to Choose a Data Mining System?

- Commercial data mining systems have little in common
  - Different data mining functionality or methodology
  - May even work with completely different kinds of data sets
- Need multiple dimensional view in selection
- Data types: relational, transactional, text, time sequence, spatial?
- System issues
  - running on only one or on several operating systems?
  - a client/server architecture?
  - Provide Web-based interfaces and allow XML data as input and/or output?
How to Choose a Data Mining System? (2)

- **Data sources**
  - ASCII text files, multiple relational data sources
  - support ODBC connections (OLE DB, JDBC)?

- **Data mining functions and methodologies**
  - One vs. multiple data mining functions
  - One vs. variety of methods per function
    - More data mining functions and methods per function provide the user with greater flexibility and analysis power

- **Coupling with DB and/or data warehouse systems**
  - Four forms of coupling: no coupling, loose coupling, semitight coupling, and tight coupling
    - Ideally, a data mining system should be tightly coupled with a database system

How to Choose a Data Mining System? (3)

- **Scalability**
  - Row (or database size) scalability
  - Column (or dimension) scalability
  - Curse of dimensionality: it is much more challenging to make a system column scalable that row scalable

- **Visualization tools**
  - “A picture is worth a thousand words”
  - Visualization categories: data visualization, mining result visualization, mining process visualization, and visual data mining

- **Data mining query language and graphical user interface**
  - Easy-to-use and high-quality graphical user interface
  - Essential for user-guided, highly interactive data mining
Examples of Data Mining Systems (1)

- **IBM Intelligent Miner**
  - A wide range of data mining algorithms
  - Scalable mining algorithms
  - Toolkits: neural network algorithms, statistical methods, data preparation, and data visualization tools
  - Tight integration with IBM's DB2 relational database system

- **SAS Enterprise Miner**
  - A variety of statistical analysis tools
  - Data warehouse tools and multiple data mining algorithms

- **Microsoft SQLServer 2000**
  - Integrate DB and OLAP with mining
  - Support OLEDB for DM standard

Examples of Data Mining Systems (2)

- **SGI MineSet**
  - Multiple data mining algorithms and advanced statistics
  - Advanced visualization tools

- **Clementine (SPSS)**
  - An integrated data mining development environment for end-users and developers
  - Multiple data mining algorithms and visualization tools

- **DBMiner (DBMiner Technology Inc.)**
  - Multiple data mining modules: discovery-driven OLAP analysis, association, classification, and clustering
  - Efficient, association and sequential-pattern mining functions, and visual classification tool
  - Mining both relational databases and data warehouses
CRISP-DM: Data Mining Process

- Cross-Industry Standard Process for Data Mining (CRISP-DM)
- European Community funded effort to develop framework for data mining tasks
- Goals:
  - Encourage interoperable tools across entire data mining process
  - Take the mystery/high-priced expertise out of simple data mining tasks

Why Should There be a Standard Process?

- 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”

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

- 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, ...

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: Overview

CRISP-DM: Phases

- **Business Understanding**
  - Understanding project objectives and requirements
  - Data mining problem definition
- **Data Understanding**
  - Initial data collection and familiarization
  - Identify data quality issues
  - Initial, obvious results
- **Data Preparation**
  - Record and attribute selection
  - Data cleansing
- **Modeling**
  - Run the data mining tools
- **Evaluation**
  - Determine if results meet business objectives
  - Identify business issues that should have been addressed earlier
- **Deployment**
  - Put the resulting models into practice
  - Set up for repeated/continuous mining of the data
Phases and Tasks

### Business Understanding
- Determine Business Objectives
  - Background
  - Business Objectives
  - Business Success Criteria
- Situational Assessment
  - Inventory of Resources
  - Requirements
  - Assumptions
  - 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 Assessment 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
  - 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
- Reformat 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
  - 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 (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 the 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

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
Phases in the DM Process (5)

- Model Evaluation
  - Evaluation of model: how well it 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

CS590D 18

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

CS590D 19
CRISP-DM: Details

• Available on-line: www.crisp-dm.org
  – 20 pages model (overview)
  – 30 page user guide (step-by-step process, hints)
  – 10 page “output” (suggested outline for a report on a data mining project)

• Has SPSS written all over it
  – But not a plug for a product (or even customized toward that product)

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