

SPSS Clementine for Data Mining in Institutional Research

C. R. Thulasi Kumar
Office of Information Management & Analysis
University of Northern Iowa
November 10-12, 2004

Overview

- What Data Mining IS and IS NOT?
- Various Data Mining Techniques
- Steps in the Data Mining Process
 - CRISP-DM
- Examples of Data Mining Applications
- Data Mining Issues
- Questions

What is Data Mining?

- The exploration and analysis of large quantities of data in order to discover meaningful patterns and rules (Berry and Linoff).
- 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.
- A relatively easy task that requires knowledge of the business problem/subject matter expertise.
- “Computers and algorithms don’t mine data; people do!”

Data Mining Applications in Institutional Research

- Student academic success/Retention and graduation
- Identify high risk students
- Predict course demand and pattern
- Profile good transfer candidates
- Application success rates
- Predict potential alumni donations

SPSS Data Mining Techniques

Technique	Method	Types
1. Predictive	<ol style="list-style-type: none">1. Neural Networks2. Rule Induction3. Linear & Logistic Regression4. Sequence Detection	C5.0 and C & R Tree
2. Clustering	<ol style="list-style-type: none">1. Kohonen Networks2. K-Means Clustering3. Two-Step Clustering	
3. Association Rules	<ol style="list-style-type: none">1. APRIORI2. GRI3. CARMA	

Selecting the Appropriate Modeling Technique

Categorize your students
Classification

- Rule Induction
- Classification and Regression Trees

Predict students success
Prediction

- Neural Networks
- Regression

Group similar students
Segmentation

- Kohonen Networks
- K-Means Clustering
- Two-Step Clustering

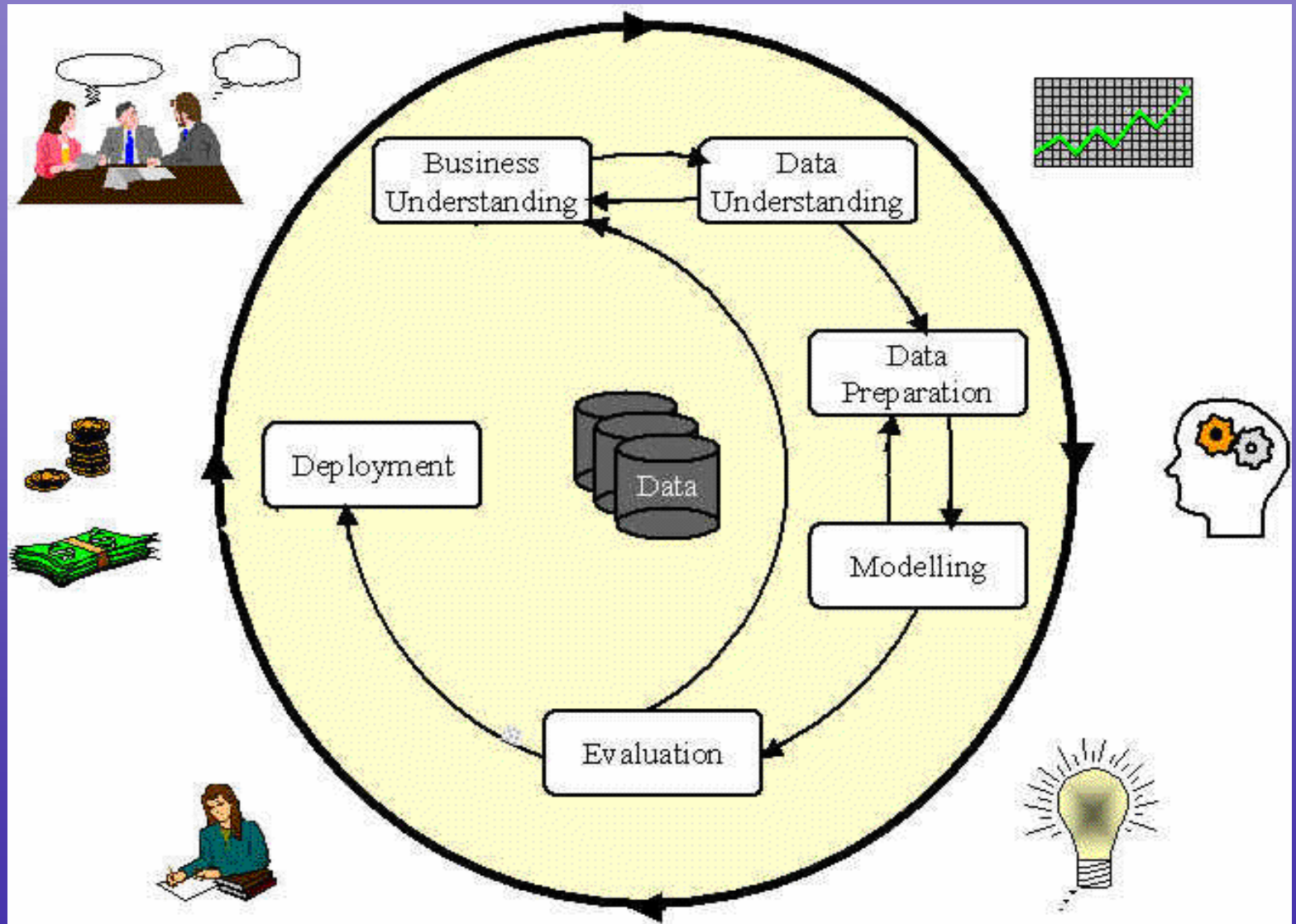
Identify courses that are taken together
Association

- APRIORI
- GRI
- CARMA

Find patterns and trends over time
Sequence

- Capri
- Rule Induction

Phases in the DM Process: CRISP-DM



Phases and Tasks

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
<p>Determine Business Objectives Background Business Objectives Business Success Criteria</p> <p>Situation Assessment Inventory of Resources Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits</p> <p>Determine Data Mining Goal Data Mining Goals Data Mining Success Criteria</p> <p>Produce Project Plan Project Plan Initial Assessment of Tools and Techniques</p>	<p>Collect Initial Data Initial Data Collection Report</p> <p>Describe Data Data Description Report</p> <p>Explore Data Data Exploration Report</p> <p>Verify Data Quality Data Quality Report</p>	<p>Data Set Data Set Description</p> <p>Select Data Rationale for Inclusion / Exclusion</p> <p>Clean Data Data Cleaning Report</p> <p>Construct Data Derived Attributes Generated Records</p> <p>Integrate Data Merged Data</p> <p>Format Data Reformatted Data</p>	<p>Select Modeling Technique Modeling Technique Modeling Assumptions</p> <p>Generate Test Design Test Design</p> <p>Build Model Parameter Settings Models Model Description</p> <p>Assess Model Model Assessment Revised Parameter Settings</p>	<p>Evaluate Results Assessment of Data Mining Results w.r.t. Business Success Criteria Approved Models</p> <p>Review Process Review of Process</p> <p>Determine Next Steps List of Possible Actions Decision</p>	<p>Plan Deployment Deployment Plan</p> <p>Plan Monitoring and Maintenance Monitoring and Maintenance Plan</p> <p>Produce Final Report Final Report Final Presentation</p> <p>Review Project Experience Documentation</p>

Model Building

- Predictive or Descriptive
- Selecting data mining tools
- Transforming data if needed
- Generating samples (as necessary) for training, testing and validating the model
- Build, test and select models.

Collecting, Cleaning, and Preparing Data

- Obtain necessary data from various internal and external sources.
- Resolve representation and encoding differences.
- Join data from various tables to create a homogeneous source.
- Check and resolve data conflicts, outliers (unusual or exception values), missing data, and ambiguity.
- Use conversions and combinations to generate new data fields such as ratios or rolled-up summaries..

Validating the Models

- Test the model for accuracy on an independent dataset, one that has not been used to create the model.
- Assess the sensitivity of a model.
- Pilot test the model for usability.

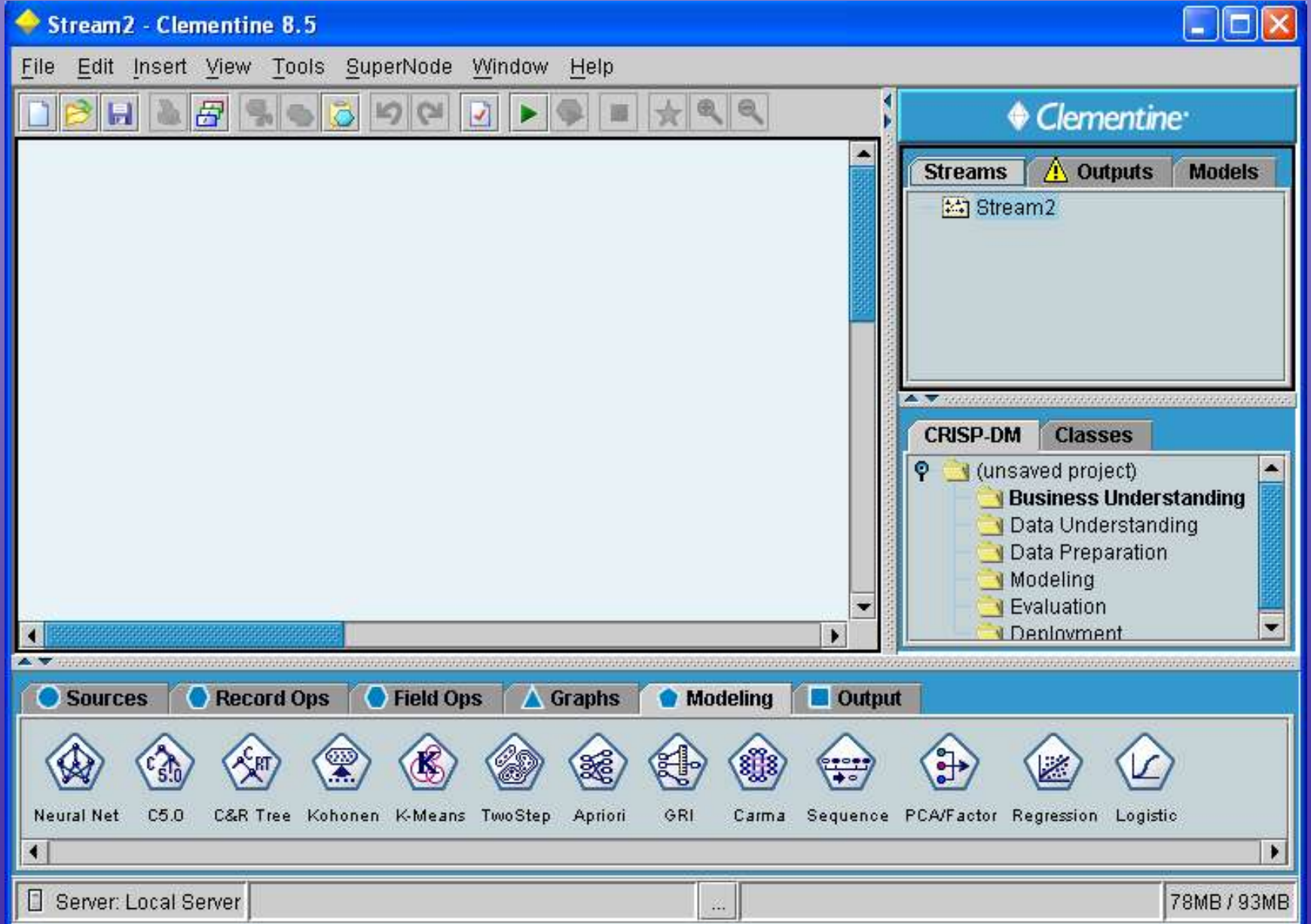
Deploying the Model

- For a predictive model, use the model to predict results for new cases, then use the prediction to alter organizational behavior.
- Deployment may require building computerized systems that capture the appropriate data and generate a prediction in real time so that a decision maker can apply the prediction. For example, a model can determine if a credit card transaction is likely to be fraudulent.

Monitoring

- Whatever you are modeling, it is likely to change over time.
- Monitoring models requires constant revalidation of the model on new data to assess if the model is still appropriate.

Clementine Screen Shot 1



Clementine Screen Shot 2

The screenshot displays the Clementine 8.5 software interface. The main workspace contains a workflow diagram with the following components and connections:

- insclaim.dat** (Source) connects to a **type** node.
- The **type** node connects to a **CLAIM** node (yellow diamond).
- The **type** node also connects to a **table** node.
- The **type** node connects to a **CLAIM** node (blue pentagon).
- The **type** node connects to a **CLAIM** node (blue triangle).
- The **CLAIM** node (yellow diamond) connects to a **DIFF** node.
- The **DIFF** node connects to a **sort** node.
- The **sort** node connects to a **table** node.
- The **CLAIM** node (yellow diamond) connects to a **Merge** node.
- The **Merge** node connects to a **Matrix** node.
- The **Merge** node connects to a **CLAIM** node (blue triangle).
- The **Merge** node connects to a **(generated)** node.
- The **Merge** node connects to a **\$N-CLAIM v. \$E-CLAIM** node (blue triangle).
- The **(generated)** node connects to a **Table** node.
- The **CLAIM** node (blue pentagon) connects to a **CLAIM** node (blue triangle).
- The **CLAIM** node (blue triangle) connects to a **CLAIM v. \$E-CLAIM** node (blue triangle).

The right sidebar shows the **Clementine** interface with the following sections:

- Streams**: Contains a **CLAIM** node.
- CRISP-DM Classes**: Lists the following classes:
 - (unsaved project)
 - Business Understanding
 - Data Understanding
 - Data Preparation
 - Modeling
 - Evaluation
 - Deployment

The bottom toolbar includes the following nodes: Type, Filter, Derive, Filler, Reclassify, Binning, SetToFlag, History, and Field Reorder.

The status bar at the bottom indicates: Server: Local Server, 32MB / 41MB.

Specific Data Mining Applications

- U.S. News College Ranking Survey
 - Graduation Rate Prediction
- College Student Experience Questionnaire (CSEQ UNI Data)
 - Best Predictors of Writing Skills

Clementine Screen Shot 3

The screenshot displays the Clementine 8.5 data mining software interface. The main workspace contains a complex workflow diagram starting with a data source 'USNEWS 1995 FINAL5.s...' which feeds into a 'Type' node. From this 'Type' node, the workflow branches into several paths:

- A path through a 'Table' node to a '(600)' node, which then connects to another 'Type' node.
- A path through a 'Sort' node to a 'Select' node.
- A path through a '(generated)' node to a 'Table' node.
- A path through a 'State' node to a 'Graduation rate' node.
- A path through a 'Yield v. Graduation...' node to a 'Table' node.
- A path through a 'Graduation rate' node to an 'Analysis' node.

The 'Select' node leads to a 'Graduation rate' node, which then connects to a 'Merge' node. The 'Merge' node receives input from a '(generated)' node and a 'Table' node. The 'Merge' node's output goes to a 'Derive2' node, which then connects to a 'Derive3' node. The 'Derive3' node outputs to a 'Table' node. The 'Merge' node also connects to an 'Analysis' node. The 'Derive2' node connects to a 'Table' node. The 'Derive3' node connects to a 'Table' node. The 'Merge' node also connects to a 'Table' node. The 'Merge' node also connects to a 'Table' node. The 'Merge' node also connects to a 'Table' node.

The interface includes a menu bar (File, Edit, Insert, View, Tools, SuperNode, Window, Help), a toolbar with various icons, and a right-hand sidebar with 'Streams', 'Outputs', and 'Models' tabs. The 'CRISP-DM Classes' panel shows a project structure with folders for Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The bottom status bar shows 'Server: Local Server' and '74MB / 109MB'.

Clementine Screen Shot 4

The screenshot displays the Clementine 8.5 software interface. The main workspace contains a workflow diagram starting with a data source 'CSEQ 99-01-03-6.sav'. This source feeds into '4 Fields' and 'Quality' nodes. The workflow then branches into several paths: one leading to a 'Table' node, another to '(generated)' nodes, and a central 'Type' node. From the 'Type' node, multiple 'GNWRITE-Recode' nodes are connected, some of which are further processed by 'Merge' and 'GNWRITE-Recode x \$R...' nodes. The right-hand side of the interface features a 'Clementine' sidebar with 'Streams', 'Outputs', and 'Models' tabs, and a 'CRISP-DM' section showing a project structure with folders for Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The bottom of the window includes a toolbar with icons for Plot, Distribution, Histogram, Collection, Multiplot, Web, and Evaluation, and a status bar at the very bottom showing 'Server: Local Server' and '46MB / 79MB'.

Data Mining Issues

- Cost
- Complexity
- Difficulty/Technicality
- Usefulness
- Accuracy
- Ease of use

How Much Does it Cost?

- Clementine (SPSS)
 - Price varies
- Enterpriser Miner (SAS)
 - Academic server license \$40K-100K
- Affinium Model (Unica)
- Intelligent Miner (IBM)
 - Free for Higher Education
- XLMiner
 - Standard academic version \$199 for two-years
- GhostMiner
 - \$2.5K-30K + Maintenance fee
- CART and MARS (Salford Systems)
 - \$1,000 (3 year license)
- Insightful Miner (Insightful)
 - Small/fraction of other mining tools

Resources

Web Sites

- <http://www.kdnuggets.com/>
- <http://www.uni.edu/instrsch/dm/index.html>

Training


- <http://www.the-modeling-agency.com>



Data Mining, Knowledge Discovery, Genomic Mining, Web Mining

Data Mining Consulting | Data Mining Jobs | Advertising | Contact Us

Full in-database mining with **Clementine 8.5**



Full in-database mining with Clementine 8.5

KDnuggets News, Data Mining & Knowledge Discovery

newsletter: data mining news, jobs, software, courses, ...

[2004 issues](#) | [Schedule](#) | [Archive](#) | [Submit](#) | [Subscribe!](#)

Current Issue: **NEW!** [04:20, Oct 26, Your preference: Bush or Kerry? Mining books and films \(32 items\)](#)

Match in: [help](#)

Software:
[Classification](#), [Suites](#), [Text](#)

Jobs: **NEW!** [Yahoo!](#)
[Industry](#), [Academic](#)

Solutions:
[Bioinformatics](#), [CRM](#), [Web](#)

Courses: [Nov](#), [Dec](#), [Jan](#)
[Education](#)

Companies:
[Consulting](#), [Products](#)

NEW! [College Data Mining Course](#)
by G. Piatetsky-Shapiro and G. Parker

Web sites:
[AI](#), [Bio](#), [Data Mining](#)

Meetings
Conferences

FAQ
[Data Mining](#)

Publications:
[Books](#), [Surveys](#), [Business](#)

ACM SIGKDD:
[Data Mining Society](#)

Polls
NEW! [Your Primary Occupation](#)

Datasets:
[Competitions](#), [KDD Cup](#)

Poll

In US Elections, are you

- US voter, for Bush
- US voter, for Kerry
- US voter, for other
- US voter, undecided
- non-US, for Bush
- non-US, for Kerry
- non-US, for other
- non-US, undecided

[View Results](#)

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites Internet Options

Address <http://www.uni.edu/instrsch/dm/index.html> Go Links

Data Mining in Institutional Research

a beginner's resource

[Data Mining in IR](#) [Tutorials](#) [Books](#) [Training](#) [Reports](#) [Software](#) [Organizations](#) [Journals](#) [IR Applications](#)

What is Data Mining?

- The process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data stored in repositories and by using pattern recognition technologies as well as statistical and mathematical techniques (The Gartner Group).
- The Nontrivial extraction of implicit, previously unknown and potentially useful information from data (Frawley, Paitestky-Shapiro and Mathews).

Data Mining in Institutional Research

- Data analysis for institutional research (IR) has evolved from simple retrospective data delivery in the 1960's to retrospective dynamic data delivery at multiple levels in the 1990's. Unlike the past methodologies, data mining is prospective and proactive in data analysis and information delivery. With a blend of tools and techniques from disciplines such as statistics, computer science, mathematics, biology and engineering, data mining provides new opportunities for institutional research professionals to provide decision support data. This site provides a collection of resources from an introductory perspective for institutional research professionals interested in data mining.
- As this area is still in its infant stages, real world examples of IR applications are difficult to find, let alone emulate. As more and more examples in IR become available, this site will be updated. Until that time, most of the examples refer to the current data mining applications in the business and industry sectors.
- Data mining has been used by universities in a number of areas, including but not limited to enrollment management, retention and graduation analysis, survey data analysis, and donation prediction (alumni contribution).

Comments or Suggestions?
Email Dr. Kumar, Information Management & Analysis
Last Modified: March 25, 2004

Copyright 2004 University of Northern Iowa, Office of Information Management & Analysis

Done Internet

Conclusions