CS 590D - Data Mining
Spring 1998
http://www.cs.purdue.edu/homes/ramakris/CS590D

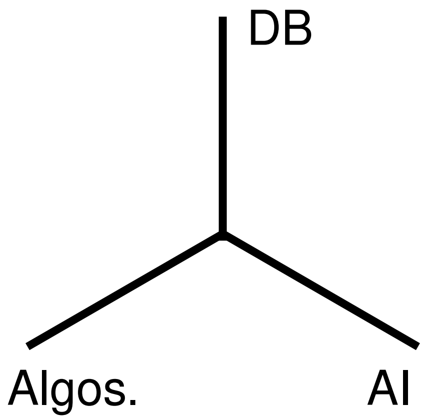
Naren Ramakrishnan and Ananth Y. Grama
Department of Computer Sciences
Purdue University, IN 47907

naren@cs.purdue.edu, CS164B, 494-6530, Off. Hours: TBA
ayg@cs.purdue.edu, CS164D, 494-6964, Off. Hours: TBA

Outline of Course

- **Objectives**
  - provide broad background in the three “core” areas
  - create specialized tools for real-life situations
  - utilize standard off-the-shelf components, tools and software
- **Review background knowledge**
  - Databases
  - Algorithms (Statistics, Data Analysis, Visualization)
  - AI (Pattern Recognition, Machine Learning)
- **The ‘Works’**
  - Full fledged Knowledge Discovery systems
  - Parallel Data Mining
  - Integrated Discovery Systems
  - Case Studies from real world projects

What is Data Mining?



- a.k.a.
  - Exploratory Data Analysis
  - Unsupervised pattern recognition
  - Data Driven Discovery, Deductive learning etc.

- Determining patterns in data
  - non-trivial
  - valid
  - novel
  - potentially useful
  - interesting
  - general and simple
  - understandable

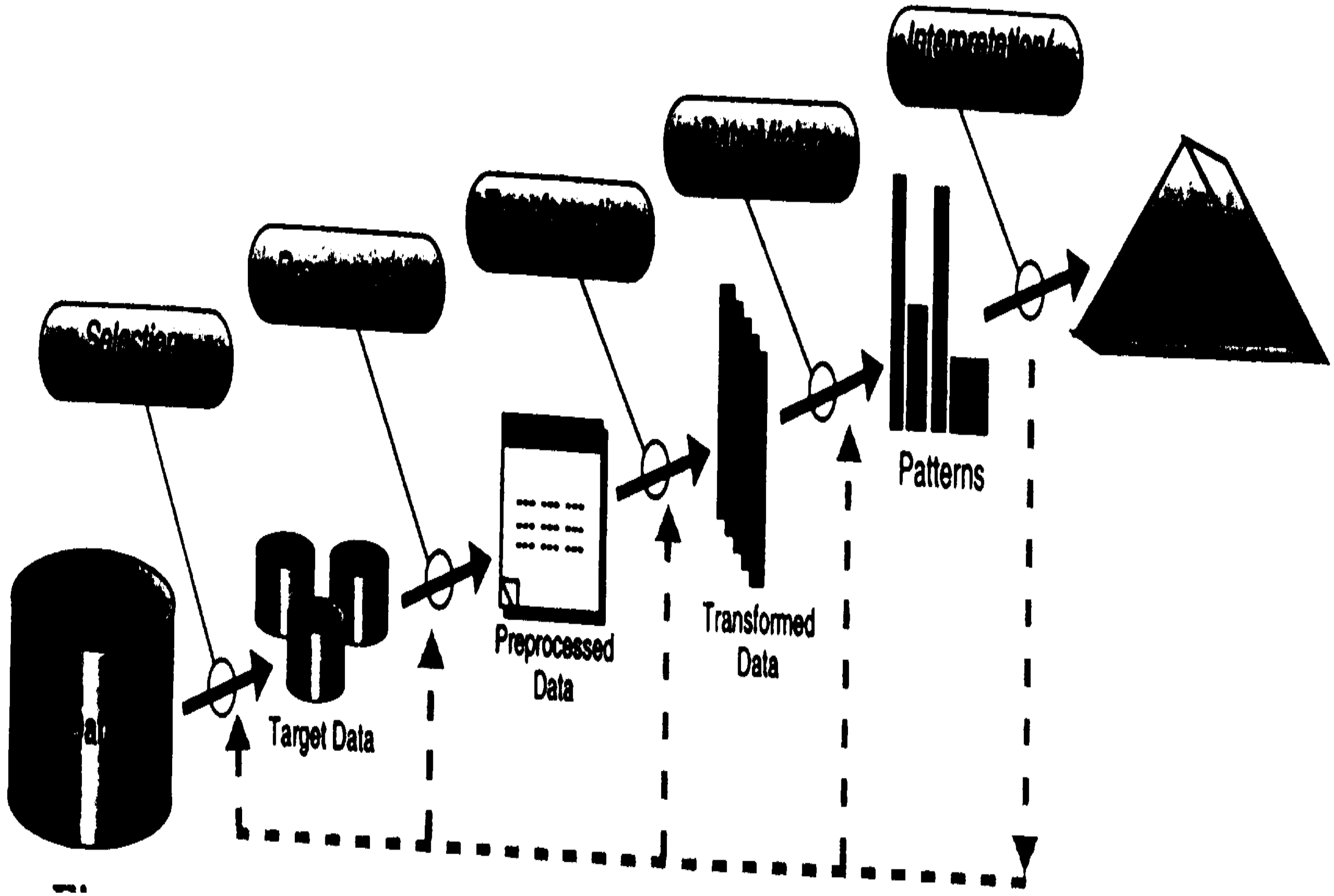
Difficulties in Data Mining

- ❏ Data rarely collected for the purpose of mining
- ❏ Limited Depth, Bias and the Curse of Dimensionality
  - ❏ lack of ‘controls’
- ❏ Huge dimensionality
- ❏ Representation of Causal Relationships
- ❏ Closed world assumption
- ❏ Overfitting, Statistical Significance
- ❏ The missing data/attributes problem
- ❏ Understandability of patterns

More Difficulties ...

- Mining the obvious (AM program)
- Database integration, models of processes
- Non-conventional forms of data
- Incremental Mining, Changing Data
- Fishing, Dredging or “Mining”

The Data Mining Process



Issues in Data Mining

- Application domain feasibility study
- Database schema, retrieval, data models
- Data transformation, cleaning
- Algorithms for Data Mining
- Interpretation and Visualization of Results
- Integration of Data Mining into Processes
- Incremental Mining, Scalability and Repeatability

Some Generic Tasks of Data Mining

- Classification and Clustering
- Regression, Discrimination and Prediction
- Density Estimation, Parameter/Model Selection
- Summarization, Rule Derivation
- Dependency Modeling
- Change/Abnormality Monitoring and Deviation Detection
- Link/Sequence Analysis

A Simple Characterization of Data Mining Approaches

- What is the model space?
- What is the preference criterion?
- What is the search algorithm?

Examples

- Commercial Data Processing
- Analyzing numerical data
- Mining the Web (The Harvest System)
- Mathematical Discovery Systems - BACON
- A Database perspective
- A second visualization marathon
- An integrated discovery system