Data Mining

Vijay Raghavan

raghavan@louisiana.edu

The Center for Advanced Computer Studies

University of Louisiana at Lafayette

Lafayette, La., USA

CONTENTS

- The Motivation
- Knowledge Discovery in Databases (KDD)
- Data Mining
  - Related Fields
  - Research Issues
  - Tasks
- Association Mining Problem
- Classification Mining Problem
- Conclusions

THE MOTIVATION

"We are drowning in information, but starving for knowledge."

John Naisbett

KNOWLEDGE DISCOVERY IN DATABASES- Definition

- A hot buzzword for a class of database applications that look for patterns or relationships in data that are:
  - Hidden,
  - Previously unknown and
  - Potentially useful

KDD: Definition

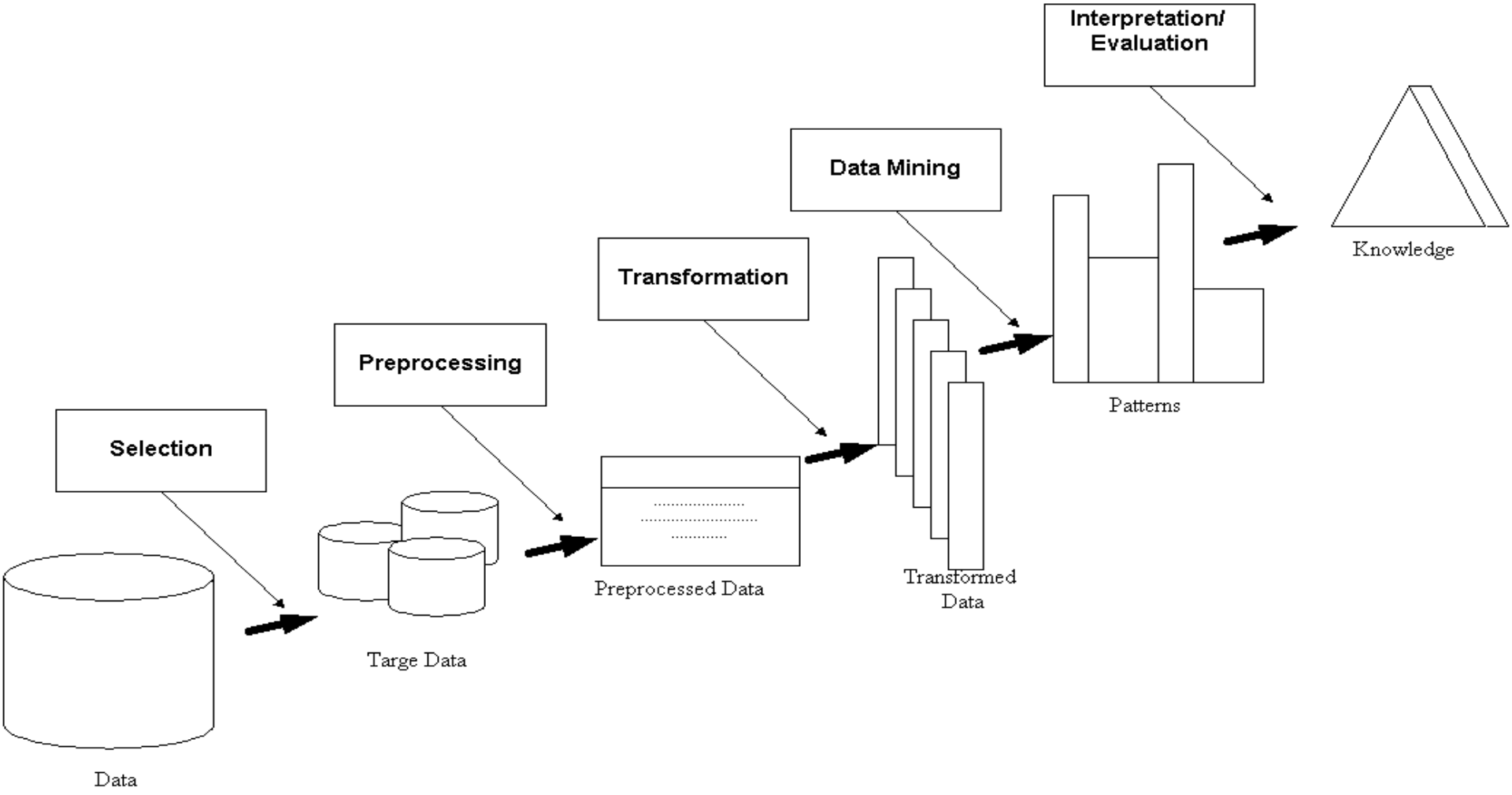
- Extract (discover):
  - interesting and
  - previously unknown

knowledge from very large real world databases.

KDD: Definition

- More formally:
  - Valid,
  - Novel, Potentially useful or Desired
  - Ultimately understandable.

KDD- PROCESS


Interpretation/ Evaluation
Data Mining
Transformation
Preprocessing
Selection
Knowledge
Patterns
Transformed Data
Preprocessed Data
Targe Data
Data

KDD vs. DATA MINING

- Synonyms (?)
- KDD
  - More than just finding pattern
  - Mining, dredging and fishing

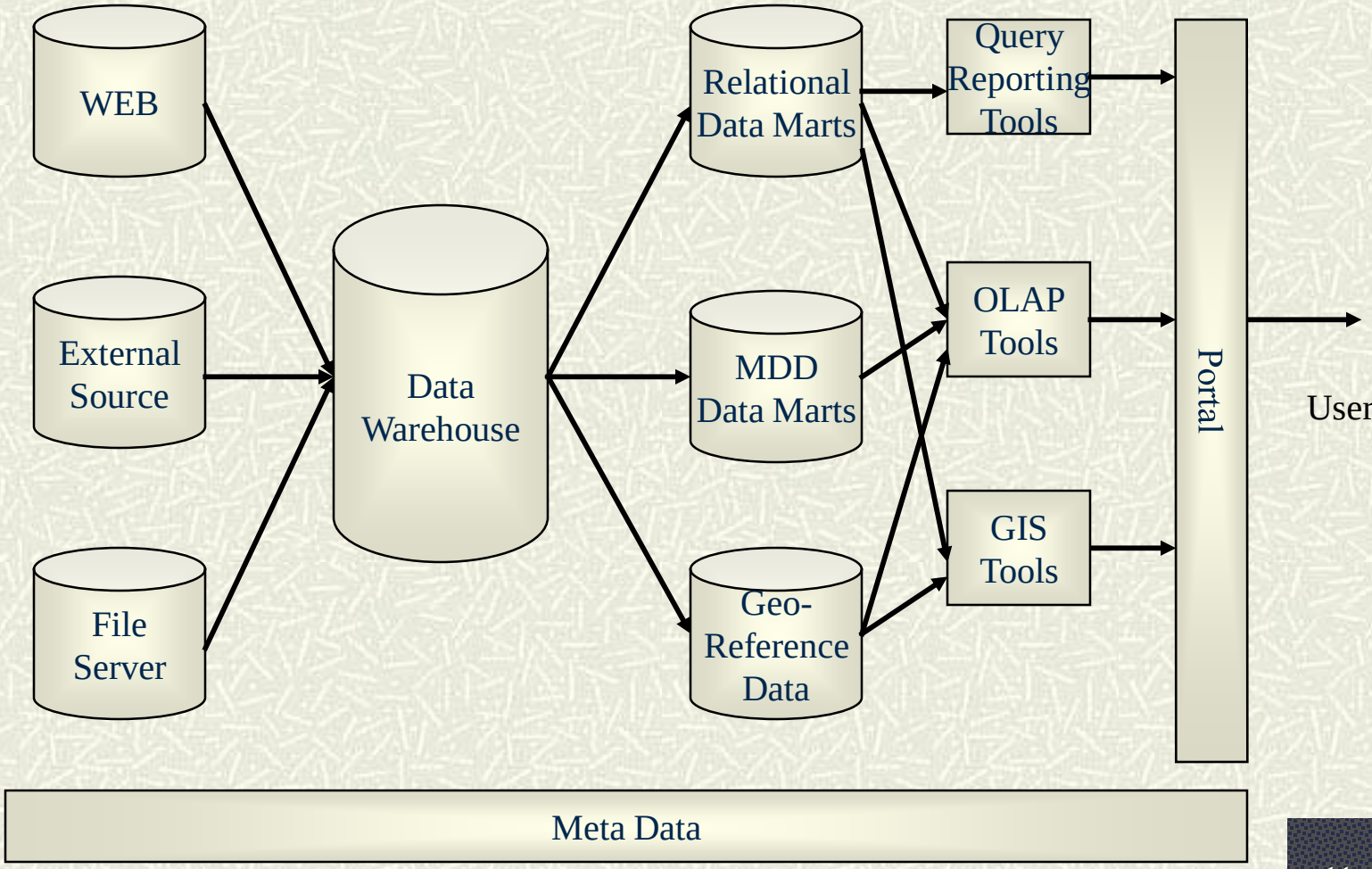
KDD- Related Fields

- Data Warehousing
- On-Line Analytical Processing (OLAP)
- Database Marketing
- Exploratory Data Analysis (EDA)

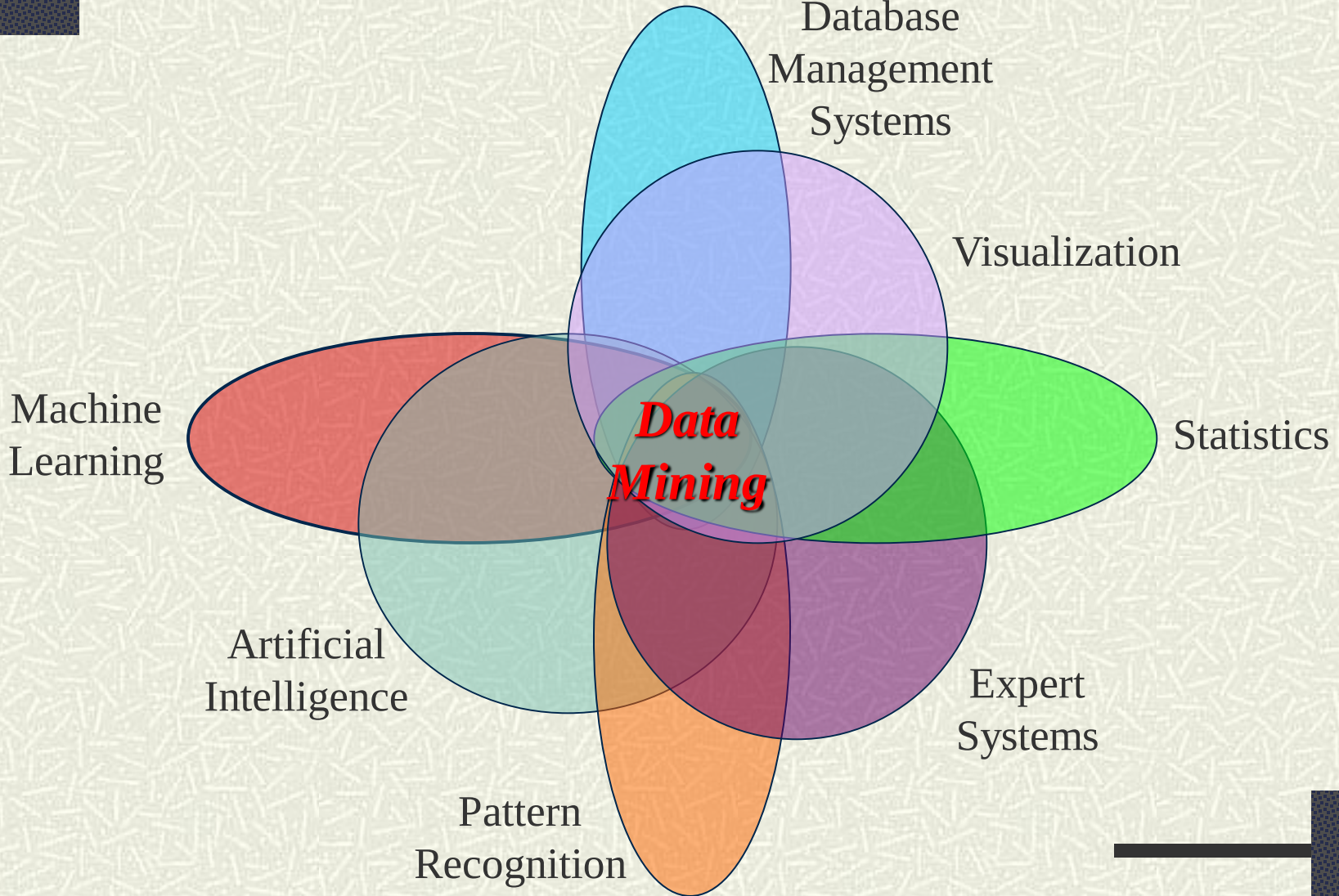
Data Warehousing

- A data warehouse is a subject-oriented, integrated, time-variant and nonvolatile collection of data in support of management's decision making process.

OLAP and Data Warehousing



Data Mining: Related Areas



Database versus Data Mining

- Query
  - DB: Well Defined & SQL
  - DM: Poorly Defined & Various Languages
- Data
  - DB: Operational (and generally relational)
  - DM: Not Operational.
- Output
  - DB: Precise, subset of the database.
  - DM: Varies.

Examples

- Database
  - Find all people with last name Raghavan.
  - Identify all customers who have bought more than 10,000 dollars
- Data Mining
  - Find those who have poor credit
  - Find all those who like the same cars
  - Find all items that are often (frequently) purchased with milk.
  - Predict the value of the housing market.

Statistics

- Simple descriptive models
- Traditionally:
  - A model created from a sample of the data to the entire dataset.
- Exploratory Data Analysis:
  - Data can actually drive the creation of the model
  - Opposite of traditional statistical view.
- Presupposes a distribution

Machine Learning

- Machine Learning: area of AI that examines how to write programs that can learn.
- Types of models
  - Classification
  - Prediction (Regression)
- Types of Learning:
  - Supervised
  - Unsupervised
- Traditionally
  - Small Datasets
  - ‘Complete’ Data

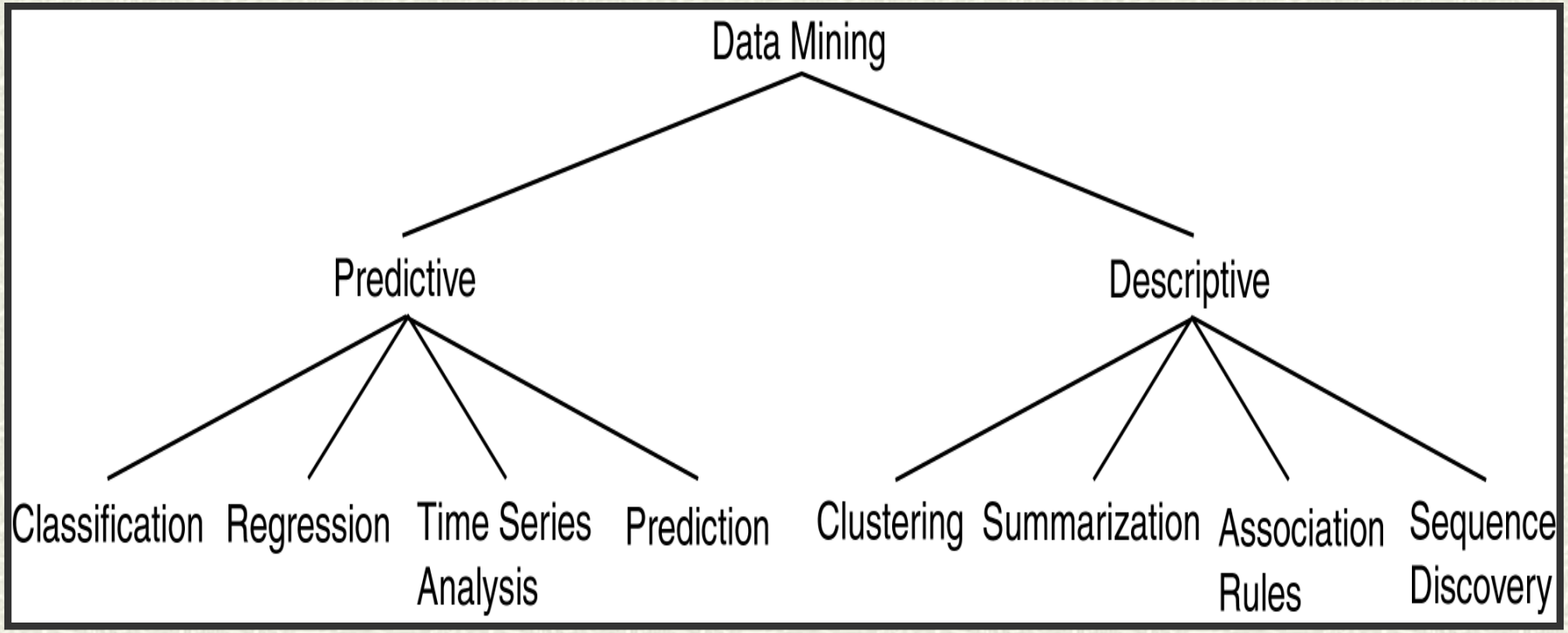
Data Mining: Research Issues

- Ultra large data
- Noisy data
- Null values
- Incomplete data
- Redundant data
- Dynamic aspects of data

Data Mining: Tasks

- Association
- Classification
- Clustering
- Estimation
- Data Visualization
- Deviation Analysis
- etc

Data Mining Models and Tasks



ASSOCIATION MINING PROBLEM

Deriving association rules from data:

Given a set of items $I = \{i_1, i_2, \ldots, i_n\}$ and a set of transactions $S = \{s_1, s_2, \ldots, s_m\}$, each transaction $s_i \in S$, such that $s_i \subseteq I$,

an **association rule** is defined as $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \varnothing$, describes the existence of a relationship between the two itemsets X and Y.

Measurements

Measures to define the strength of the relationship between two itemsets X and Y

Measure of Confidence

$$\text{Confidence}(X \Rightarrow Y) = \frac{P(X, Y)}{P(X)}$$

The percentage of transactions that contain *Y* among those transaction containing *X*.

Applications of Associations

- I = Products, S = Baskets
- I = Cited Articles, S = Technical Articles
- I = Incoming Links, S = Web pages
- I = Keywords, S = Documents
- I = Term papers, S = Sentences

Classification Mining Problem

- Pattern Recognition and Machine Learning communities
- Generally aimed at models of the data.
- Often includes both
  - Categorization
  - Prediction (Regression)
- Supervised.

Clustering Mining Problem

- Assumption: Data, naturally, falls into groups.
  - Overlapping or Non-Overlapping
- What are the groups?
  - And what data falls within each group.
- Unsupervised.

Measures

- Error
  - Categorization
    - Number Bad Assignments/Total Assignments
  - Prediction
    - Mean Squared Error
- In truth, a number of measures have been proposed.

Note about ‘Data’

- Various types:
  - Text
  - Strings
  - Numeric
  - Sound
  - Image
  - Relations
  - Etc.

CONCLUSIONS

- KDD has interesting problems
- It is an inter-disciplinary field
- No matter your expertise, you can find an interesting niche
- Many high-demand applications (e.g. CRM)