Data Mining and Soft Computing

Francisco Herrera

Research Group on Soft Computing and Information Intelligent Systems (SCI^{2}S)
Dept. of Computer Science and A.I.
University of Granada, Spain

Email: herrera@decsai.ugr.es
http://sci2s.ugr.es
http://decsai.ugr.es/~herrera

DECSAI
Universidad de Granada

Summary

1. Introduction to Data Mining and Knowledge Discovery
2. Data Preparation
3. Introduction to Prediction, Classification, Clustering and Association
4. Data Mining - From the Top 10 Algorithms to the New Challenges
5. Introduction to Soft Computing. Focusing our attention in Fuzzy Logic and Evolutionary Computation
6. Soft Computing Techniques in Data Mining: Fuzzy Data Mining and Knowledge Extraction based on Evolutionary Learning
7. Genetic Fuzzy Systems: State of the Art and New Trends
8. Some Advanced Topics I: Classification with Imbalanced Data Sets
9. Some Advanced Topics II: Subgroup Discovery
10. Some advanced Topics III: Data Complexity
11. Final talk: How must I Do my Experimental Study? Design of Experiments in Data Mining/Computational Intelligence. Using Non-parametric Tests. Some Cases of Study.

Slides based on:

Introduction to Data Mining

Natasha Balac, Ph.D.

Lecture Notes for Chapter 1

Introduction to Data Mining

by

Tan, Steinbach, Kumar

Theory and Practice

of Data Mining

Mykola Pechenizkiy

Outline

- Motivation: Why Data Mining?
- What is Data Mining?
- History of Data Mining
- Data Mining Terminology / KDD Process
- Data Mining Applications
- Issues in Data Mining
- Concluding Remarks

Outline

- Motivation: Why Data Mining?
- What is Data Mining?
- History of Data Mining
- Data Mining Terminology / KDD Process
- Data Mining Applications
- Issues in Data Mining
- Concluding Remarks

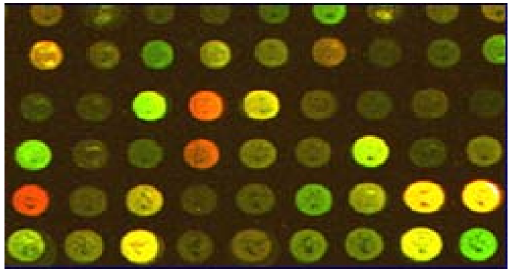
Why DATA MINING?

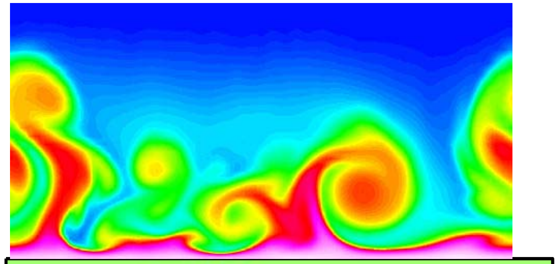
- Lots of data is being collected and warehoused
  - – Web data, e-commerce
  - – purchases at department/ grocery stores
  - – Bank/Credit Card transactions

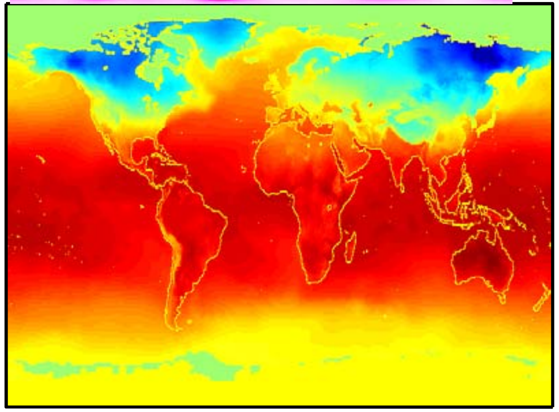
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - – Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

Why DATA MINING?

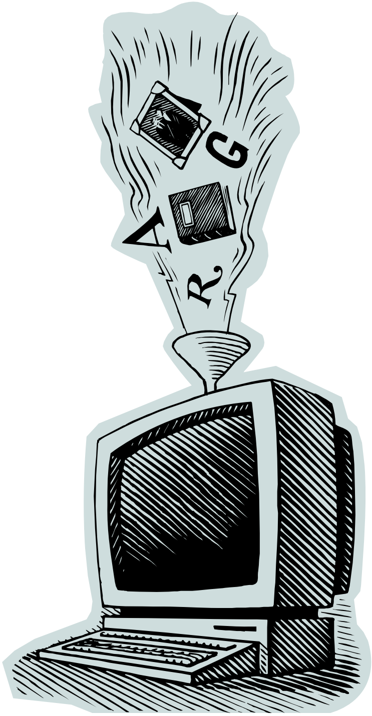
- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation







Why DATA MINING?

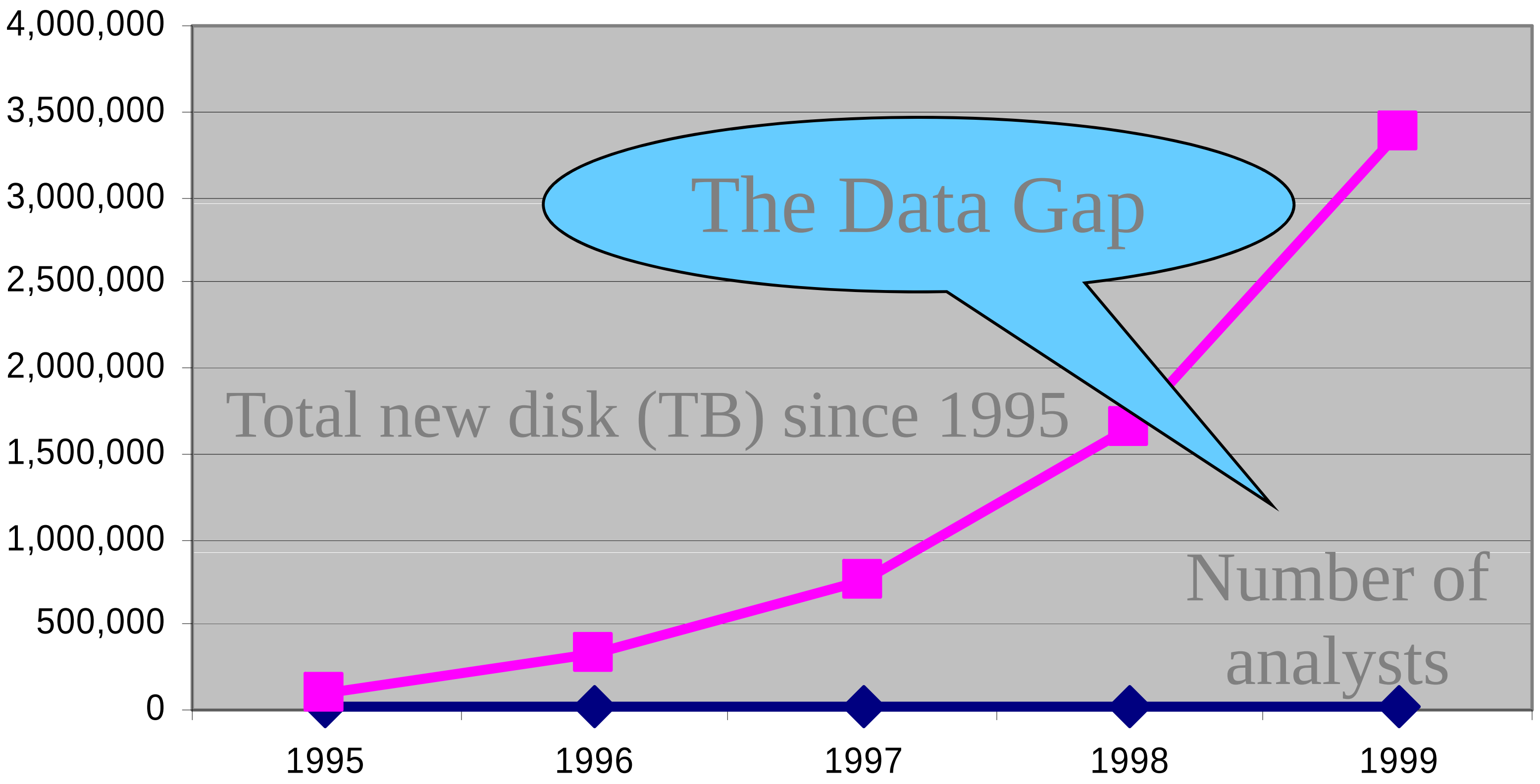


- Huge amounts of data

- Data rich – but information poor
- Lying hidden in all this data is information!

Mining Large Data Sets - Motivation

- **There is often information "hidden" in the data that is not readily evident**
- **Human analysts may take weeks to discover useful information**
- **Much of the data is never analyzed at all**



From: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

Data vs. Information

- Society produces massive amounts of data
  - business, science, medicine, economics, sports, …
- Potentially valuable resource
- Raw data is useless
  - need techniques to automatically extract information
  - Data: recorded facts
  - Information: patterns underlying the data

Outline

- Motivation: Why Data Mining?
- What is Data Mining?
- History of Data Mining
- Data Mining Terminology / KDD Process
- Data Mining Applications
- Issues in Data Mining
- Concluding Remarks

What is DATA MINING?

Many Definitions

- *Extracting or “mining” knowledge from large amounts of data*

- Data -driven discovery and modeling of hidden patterns (we never new existed) in large volumes of data

- Extraction of implicit, previously unknown and unexpected, potentially extremely useful information from data

What Is Data Mining?

- **Data mining:**
  - Extraction of interesting **(non-trivial, implicit, previously unknown** and **potentially useful**) information or patterns from data in large databases

Data Mining is NOT

- Data Warehousing
- (Deductive) query processing
  - SQL/ Reporting
- Software Agents
- Expert Systems
- Online Analytical Processing (OLAP)
- Statistical Analysis Tool
- Data visualization

- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about "Amazon"

Machine Learning Techniques

Technical basis for data mining:

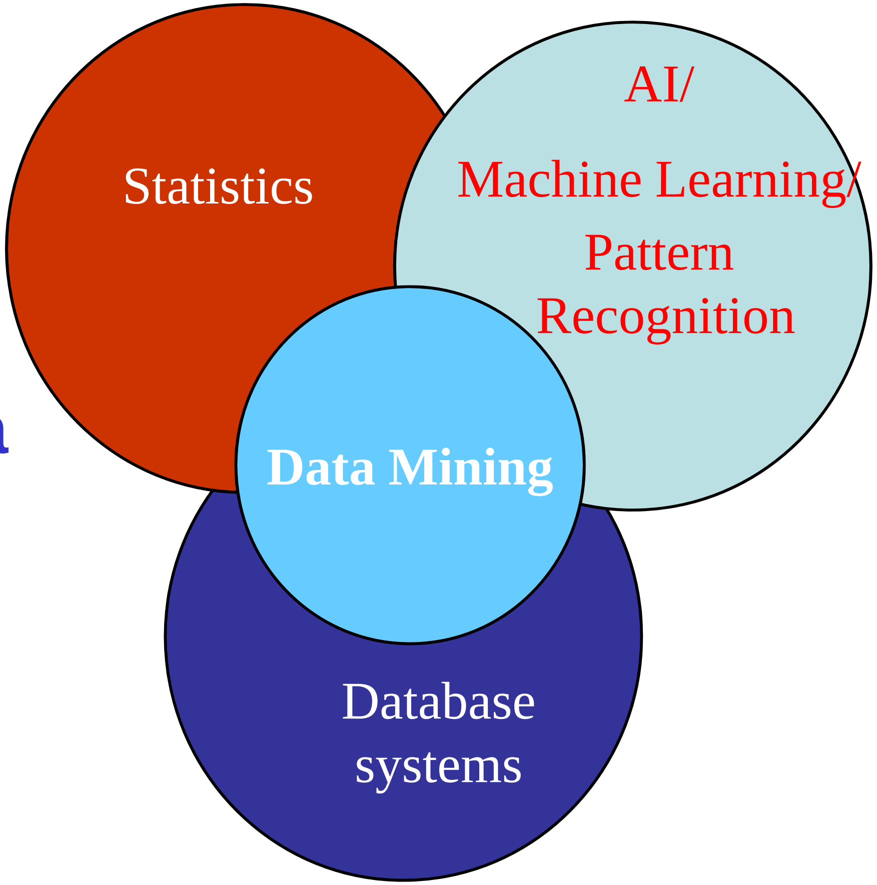
- algorithms for acquiring structural descriptions from examples

- Methods originate from artificial intelligence, machine learning, statistics, and research on databases

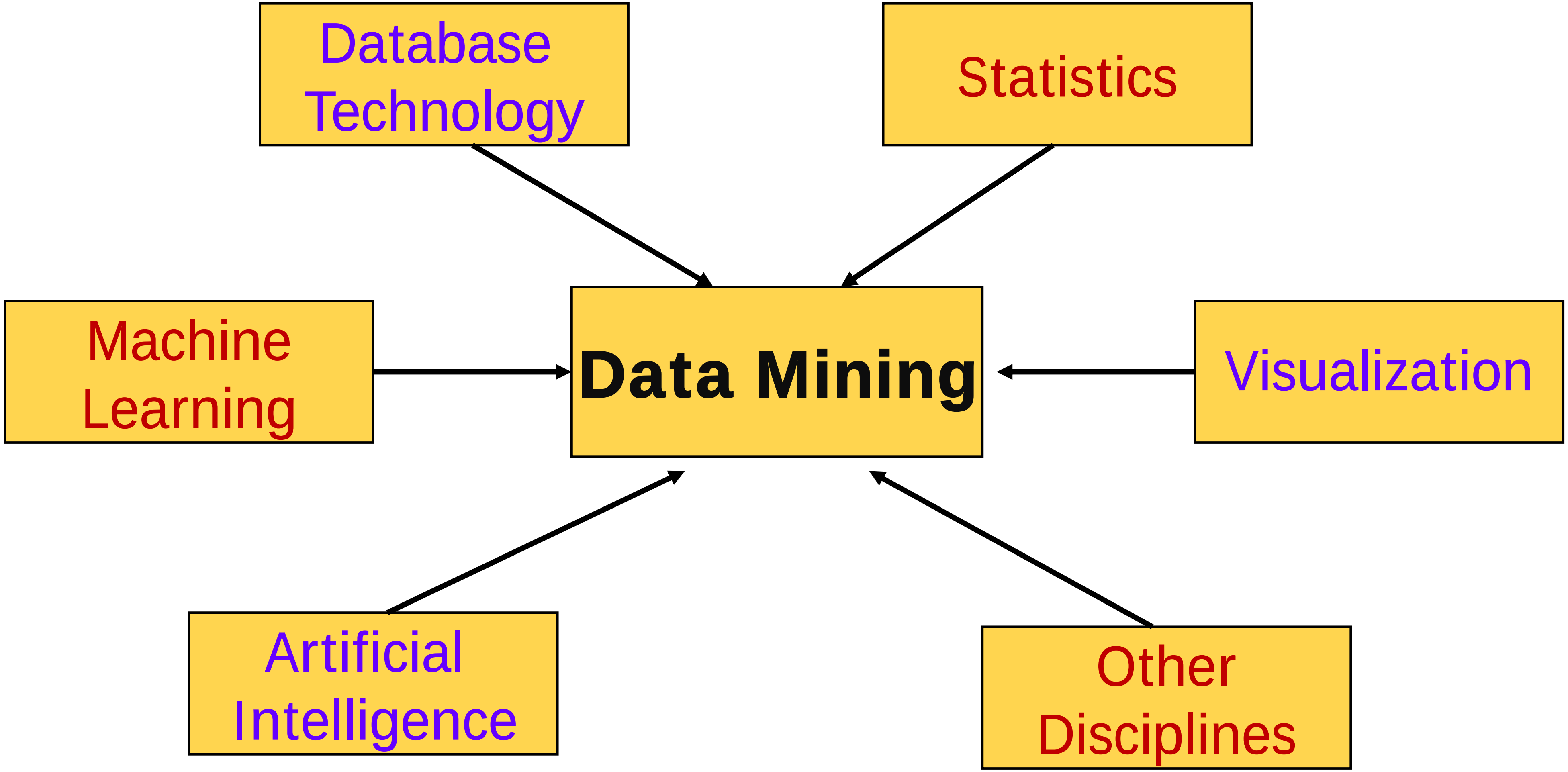
Origins of Data Mining

- **Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems**

- **Traditional Techniques may be unsuitable due to**
  - **Enormity of data**
  - **High dimensionality of data**
  - **Heterogeneous, distributed nature of data**



Multidisciplinary Field

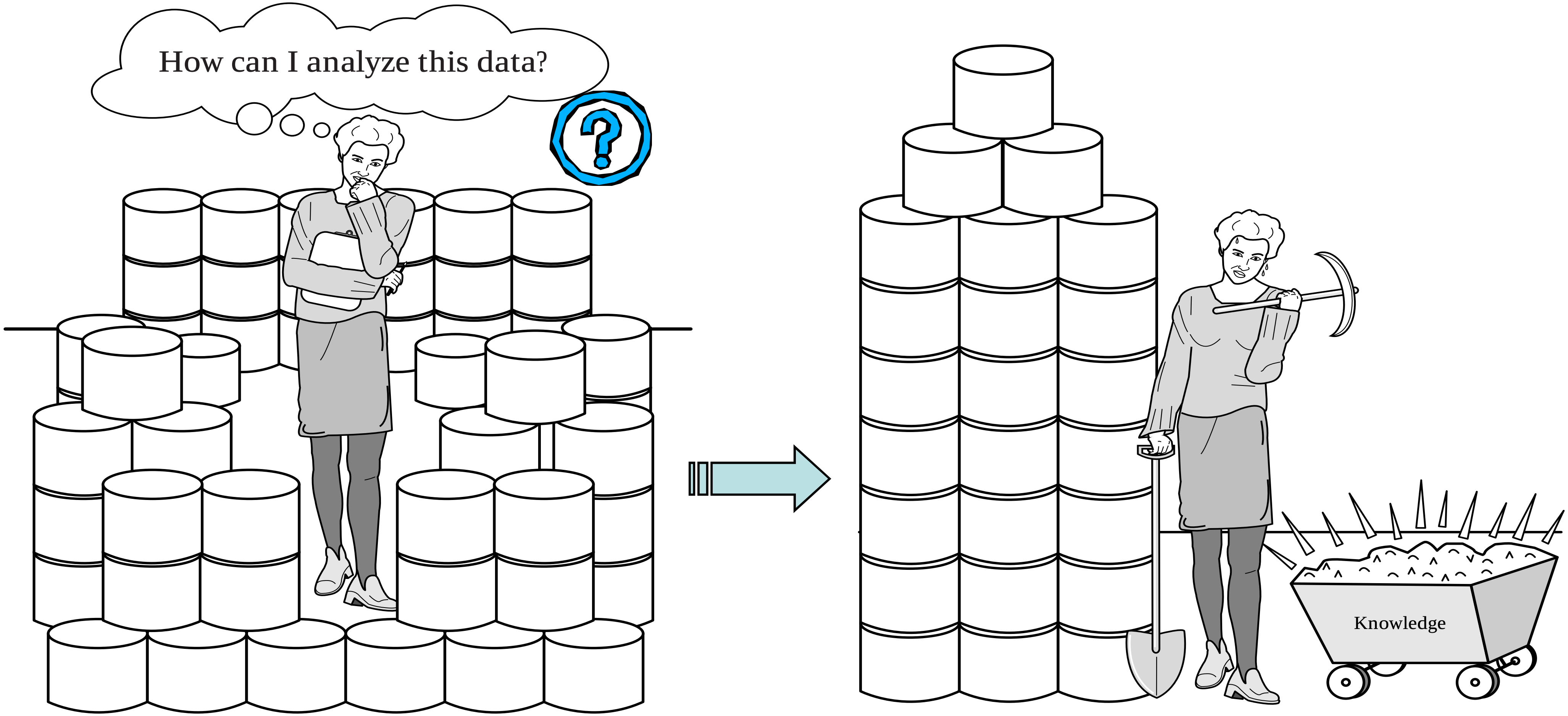


Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
- Description Methods
  - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Summary in a Picture



**We have rich data,
but poor information**

**Data mining-searching for knowledge
(interesting patterns) in your data.**

**J. Han, M. Kamber. Data Mining. Concepts and Techniques
Morgan Kaufmann, 2006 (Second Edition)**

Outline

- Motivation: Why Data Mining?
- What is Data Mining?
- History of Data Mining
- Data Mining Terminology / KDD Process
- Data Mining Applications
- Issues in Data Mining

History

- Emerged late 1980s
- Flourished –1990s
- Consolidation – 2000s

A Brief History of Data Mining Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases (Piatetsky-Shapiro)
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)

  1993 – Association Rule Mining Algorithm APRIORI proposed by Agraval, Imielinski and Swami.
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- 1998 ACM SIGKDD, SIGKDD'1999-2001 conferences, and SIGKDD Explorations
- More conferences on data mining
  - PAKDD, PKDD, SIAM-Data Mining, (IEEE) ICDM, etc.

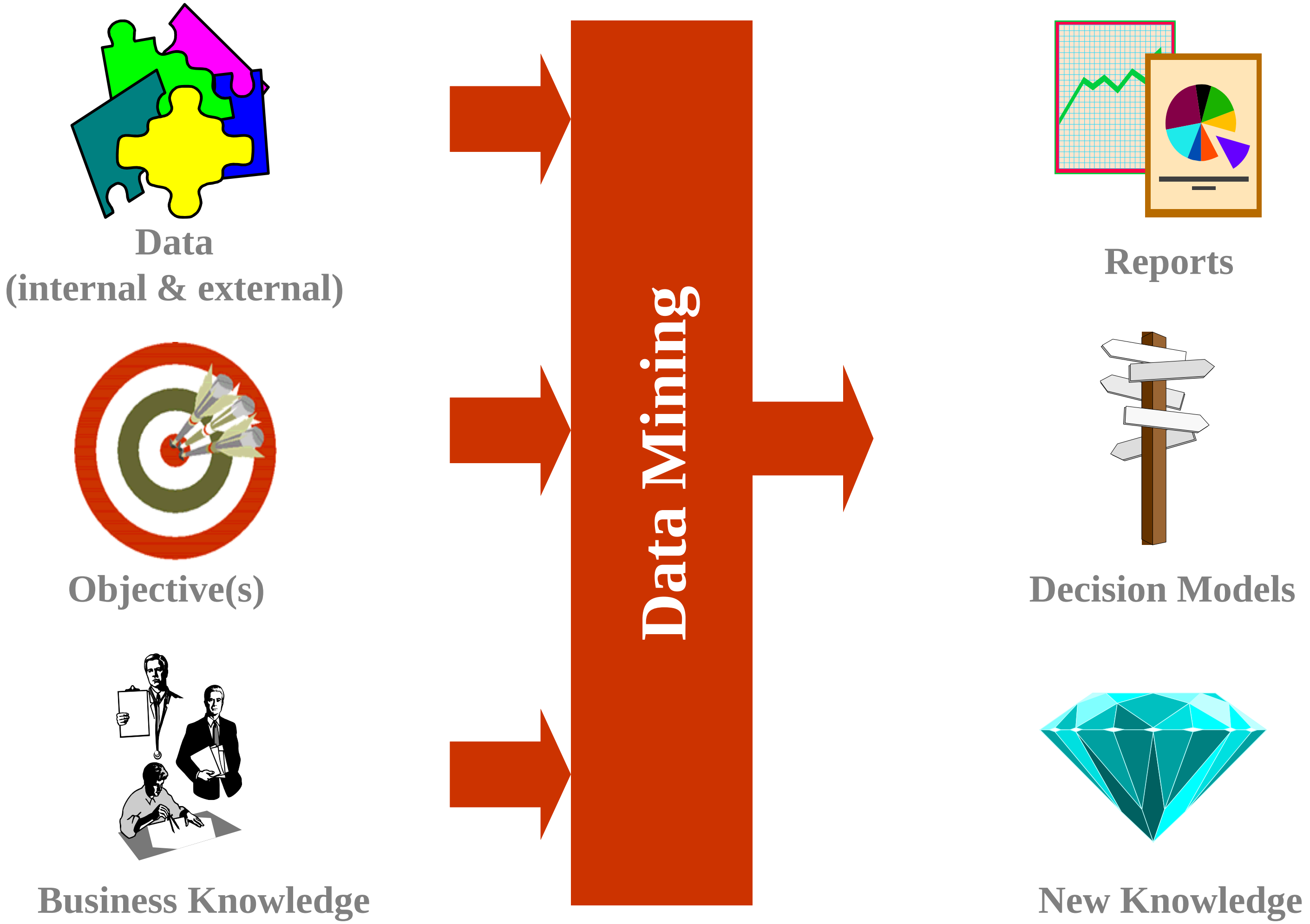
Outline

- Motivation: Why Data Mining?
- What is Data Mining?
- History of Data Mining
- Data Mining Terminology / KDD Process
- Data Mining Applications
- Issues in Data Mining
- Concluding Remarks

Terminology

- Gold Mining
- Knowledge mining from databases
- Knowledge extraction
- Data/pattern analysis
- Knowledge Discovery Databases or KDD
- Business intelligence

Input-Output View



Data
(internal & external)

Objective(s)

Business Knowledge

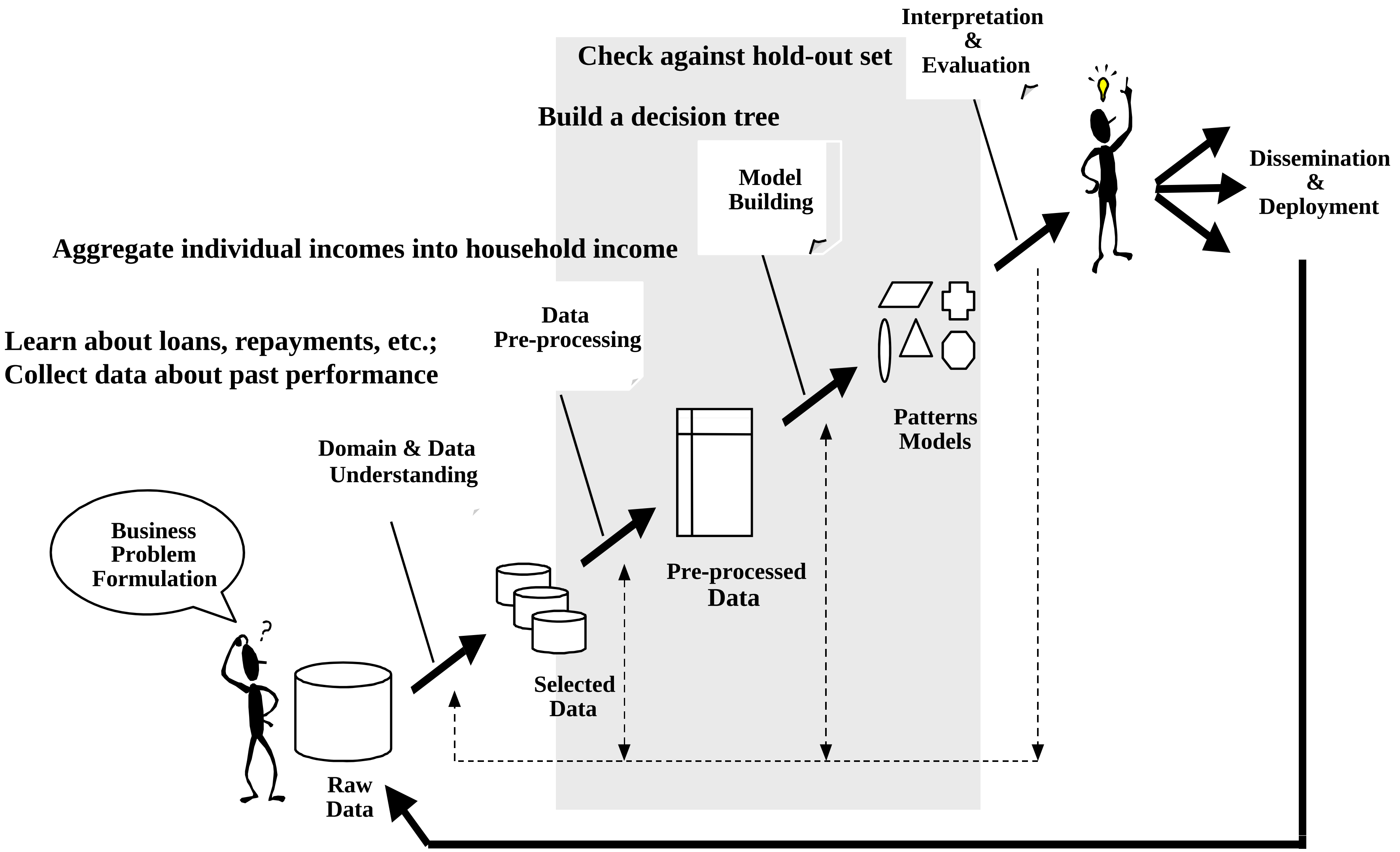
Data Mining

Reports

Decision Models

New Knowledge

Process View



Interpretation
&
Evaluation

Check against hold-out set

Build a decision tree

Dissemination
&
Deployment

Model
Building

Aggregate individual incomes into household income

Data
Pre-processing

Learn about loans, repayments, etc.;
Collect data about past performance

Patterns
Models

Domain & Data
Understanding

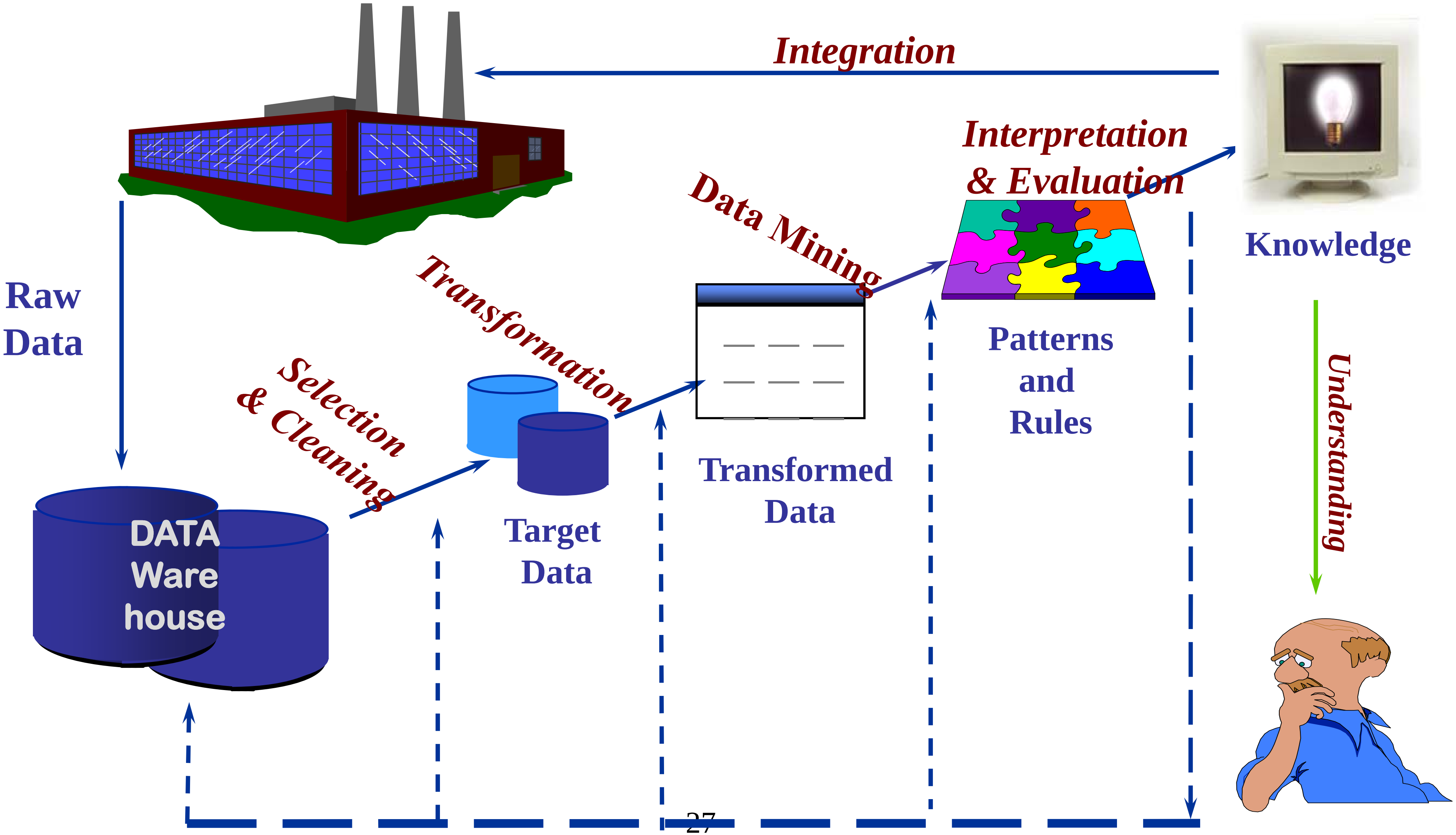
Business
Problem
Formulation

Pre-processed
Data

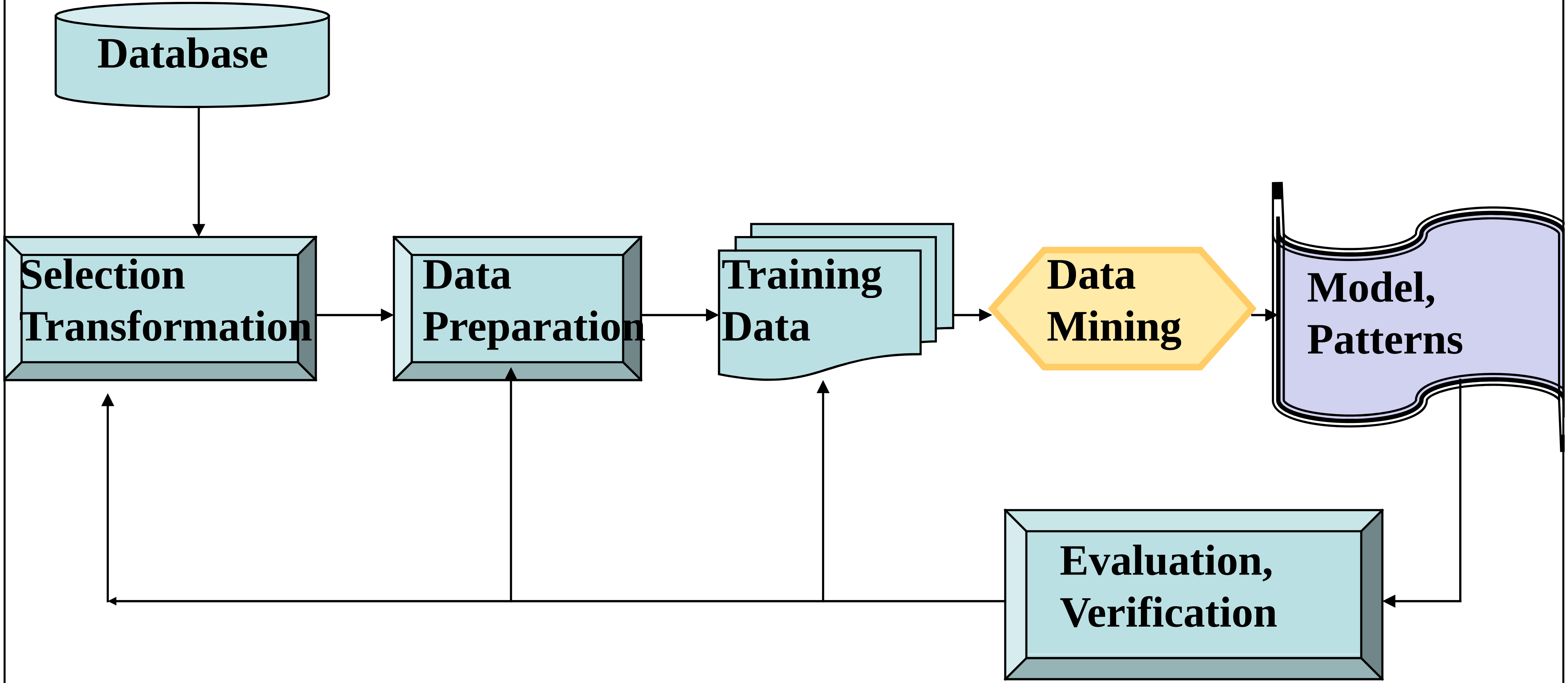
Selected
Data

Raw
Data

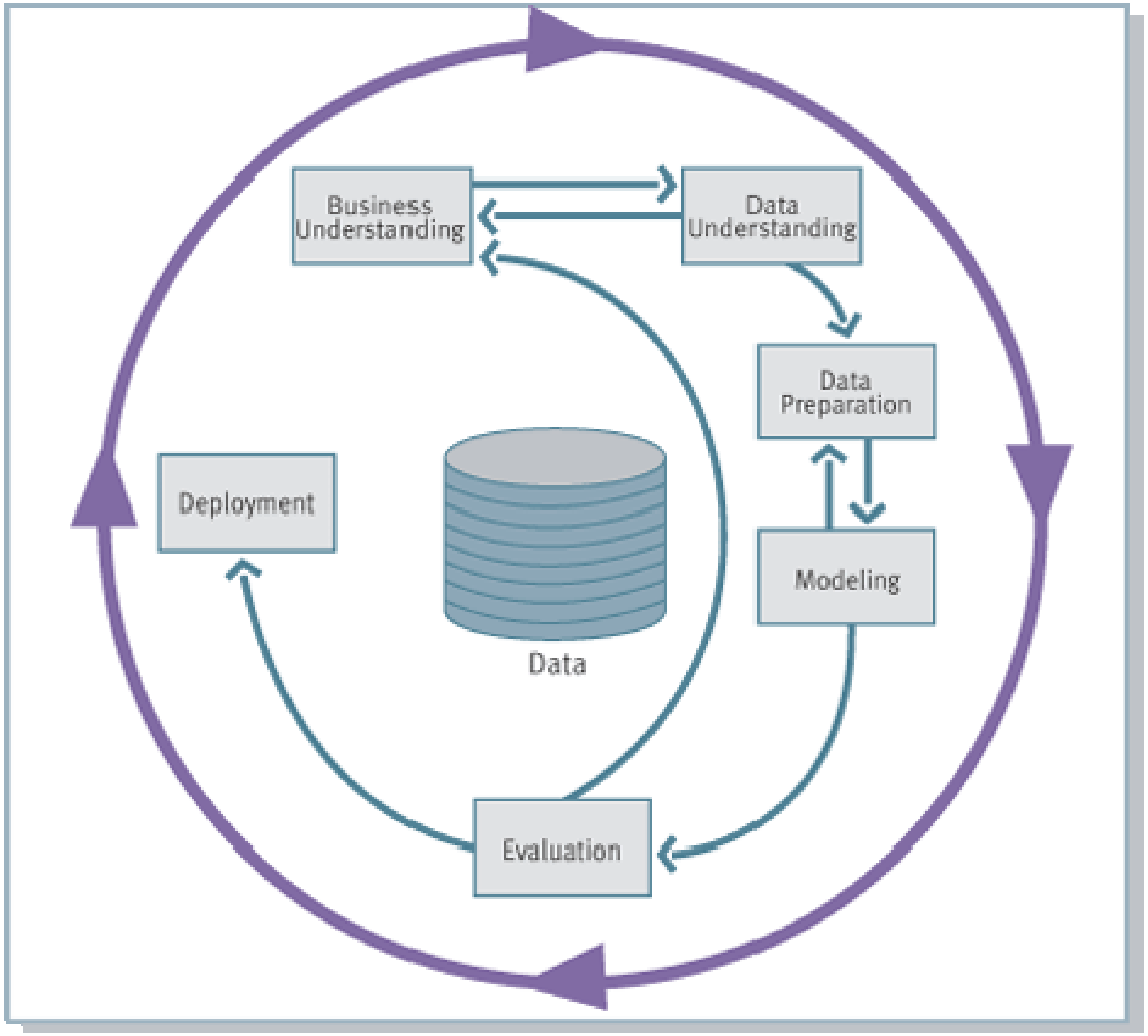
Knowledge Discovery Process



KDD Process


Database
Selection Transformation
Data Preparation
Training Data
Data Mining
Model, Patterns
Evaluation, Verification

Knowledge Discovery Process flow, according to CRISP-DM



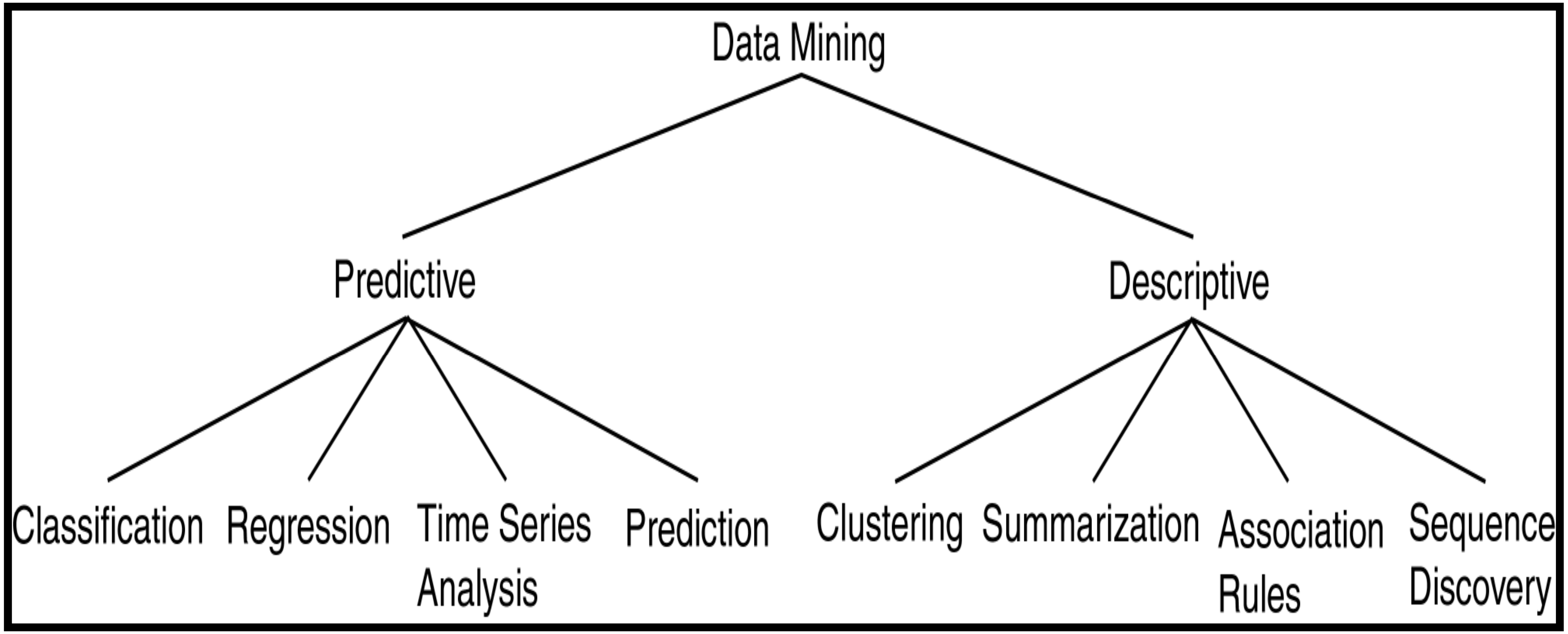
see www.crisp-dm.org for more information

Outline

- Motivation: Why Data Mining?
- What is Data Mining?
- History of Data Mining
- Data Mining Terminology / KDD Process
- Data Mining Applications
- Issues in Data Mining
- Concluding Remarks

Data Mining Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
- Description Methods
  - Find human-interpretable patterns that describe the data.



Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]
- Time Series [Predictive]
- Summarization [Descriptive]

Classification: Definition

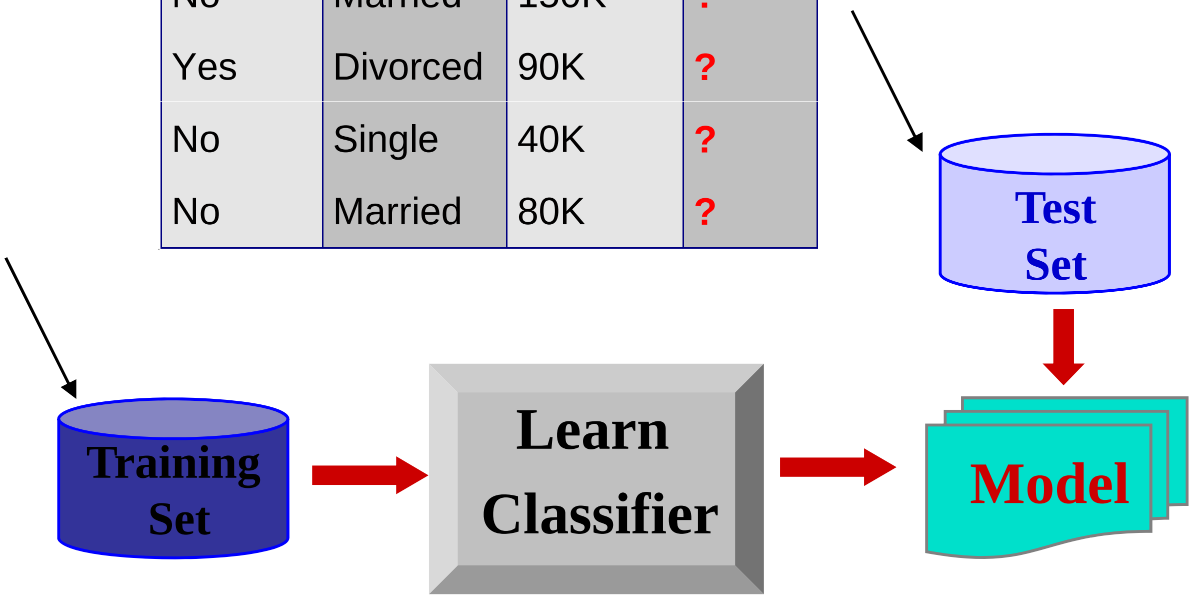
- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: <u>previously unseen</u> records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification Example

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Test Set

Training Set

Learn Classifier

Model

Classification: Application 1

- **Direct Marketing**
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don’t buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

Classification: Application 2

- **Fraud Detection**
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 3

- **Customer Attrition/Churn:**
  - Goal: To predict whether a customer is likely to be lost to a competitor.
  - Approach:
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal.
    - Find a model for loyalty.

Classification: Application 4

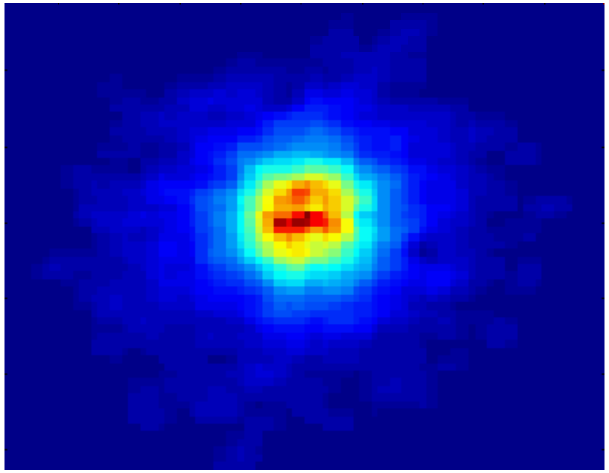
- **Sky Survey Cataloging**
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Classifying Galaxies

Courtesy: http://aps.umn.edu

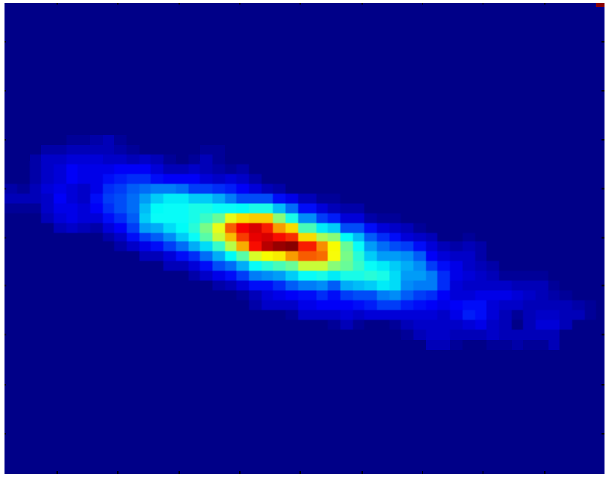
Early



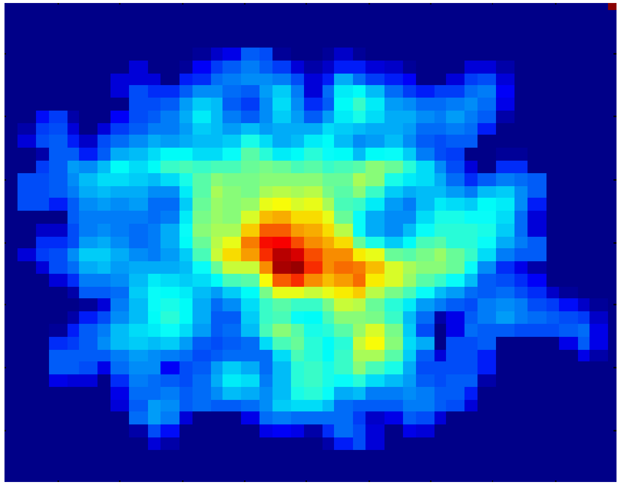
Class:
- Stages of Formation

Attributes:
- Image features,
- Characteristics of light waves received, etc.

Intermediate



Late



Data Size:
- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

Clustering Definition

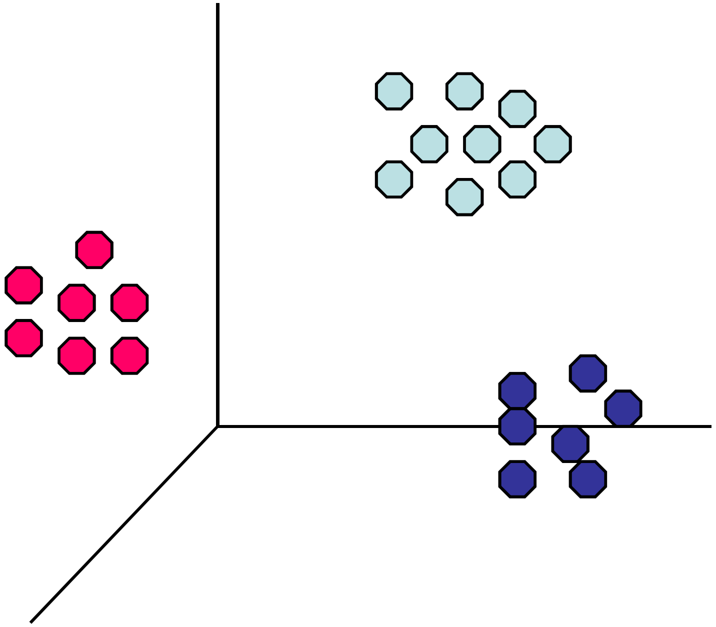
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Illustrating Clustering

☒Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



Clustering: Application 1

- **Market Segmentation:**
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

- **Document Clustering:**
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Illustrating Document Clustering

- Clustering Points: 3204 Articles of Los Angeles Times.
- Similarity Measure: How many words are common in these documents (after some word filtering).

Category	*Total Articles*	*Correctly Placed*
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association Rule Discovery: Application 1

- **Marketing and Sales Promotion:**
  - Let the rule discovered be

    {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Association Rule Discovery: Application 2

- **Supermarket shelf management.**
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is very likely to buy beer.
    - So, don’t be surprised if you find six-packs stacked next to diapers!

Association Rule Discovery: Application 3

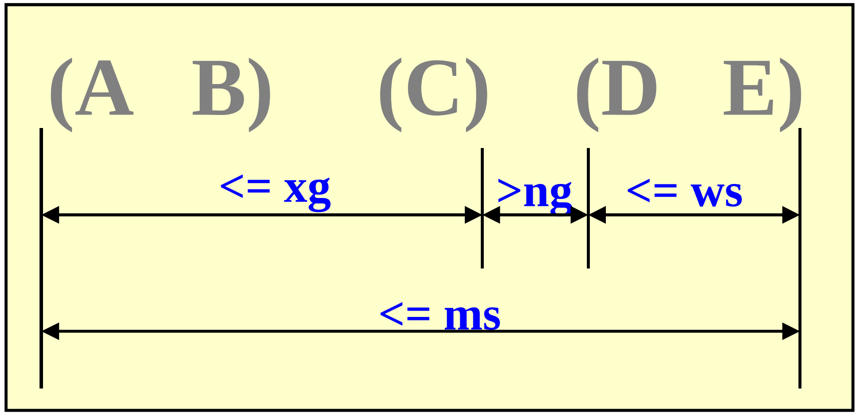
- **Inventory Management:**
  - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
  - Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Sequential Pattern Discovery: Definition

- Given is a set of *objects*, with each object associated with its own *timeline of events*, find rules that predict strong sequential dependencies among different events.

(A B) (C) → (D E)

- Rules are formed by first disovering patterns. Event occurrences in the patterns are governed by timing constraints.



Sequential Pattern Discovery: Examples

- **In telecommunications alarm logs,**
  - (Inverter_Problem Excessive_Line_Current)
    (Rectifier_Alarm) --> (Fire_Alarm)
- **In point-of-sale transaction sequences,**
  - Computer Bookstore:
    (Intro_To_Visual_C) (C++_Primer) -->

    (Perl_for_dummies,Tcl_Tk)
  - Athletic Apparel Store:
    (Shoes) (Racket, Racketball) --> (Sports_Jacket)

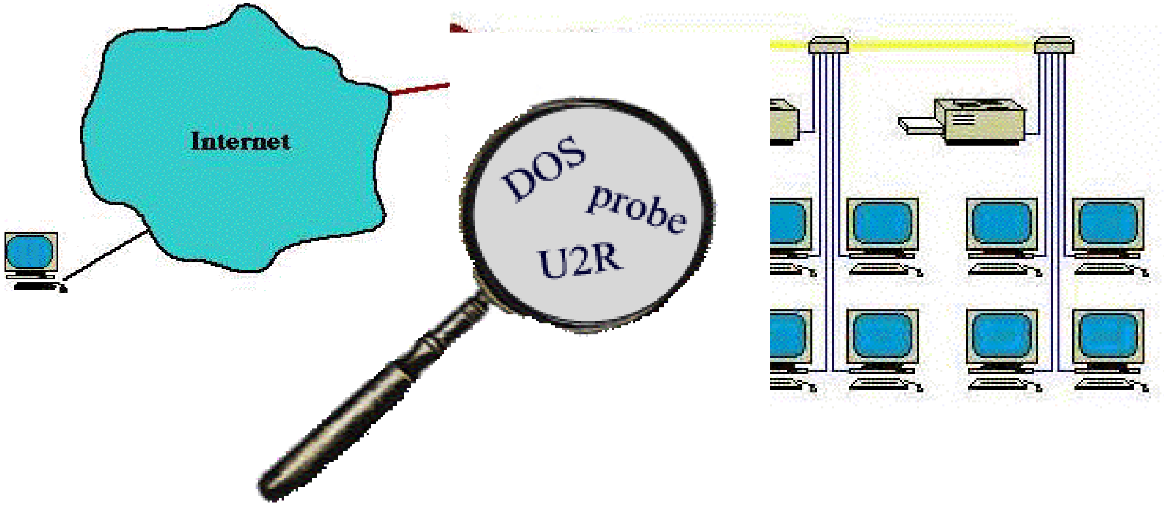
Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advetising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Deviation/Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection

  - Network Intrusion Detection

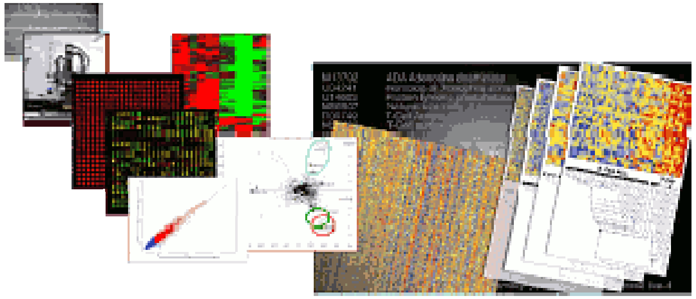


Typical network traffic at University level may reach over 100 million connections per day

Data Mining Applications

- **Science: Chemistry, Physics, Medicine**
  - Biochemical analysis
  - Remote sensors on a satellite
  - Telescopes – star galaxy classification
  - Medical Image analysis

- **Bioscience**
  - Sequence-based analysis
  - Protein structure and function prediction
  - Protein family classification
  - Microarray gene expression



Data Mining Applications

- **Pharmaceutical companies, Insurance and Health care, Medicine**
  - Drug development
  - Identify successful medical therapies
  - Claims analysis, fraudulent behavior
  - Medical diagnostic tools
  - Predict office visits

Data Mining Applications

- **Financial Industry, Banks, Businesses, E-commerce**
  - Stock and investment analysis
  - Identify loyal customers vs. risky customer
  - Predict customer spending
  - Risk management
  - Sales forecasting
- **Retail and Marketing**
  - Customer buying patterns/demographic characteristics
  - Mailing campaigns
  - Market basket analysis
  - Trend analysis

Data Mining Applications

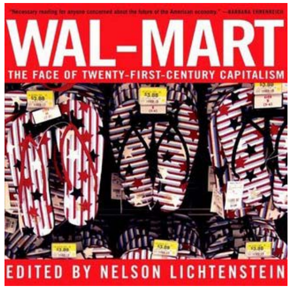
- **Database analysis and decision support**
  - Market analysis and management
    - target marketing, customer relation management, market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and management

Data Mining Applications

- **Sports and Entertainment**
  - IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage for New York Knicks and Miami Heat
- **Astronomy**
  - JPL and the Palomar Observatory discovered 22 quasars with the help of data mining

DATA MINING EXAMPLE

Beer and Nappies/Diapers -- A Data Mining Urban Legend Introduction


WAL-MART
THE FACE OF TWENTY-FIRST-CENTURY CAPITALISM
EDITED BY NELSON LICHTENSTEIN

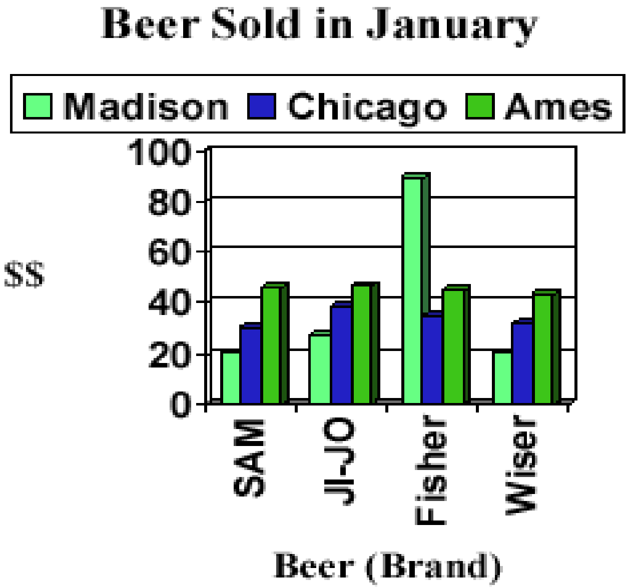
There is a story that a large supermarket chain, usually Wal-Mart, did an analysis of customers' buying habits and found a statistically significant correlation between purchases of beer and purchases of nappies (diapers in the US).

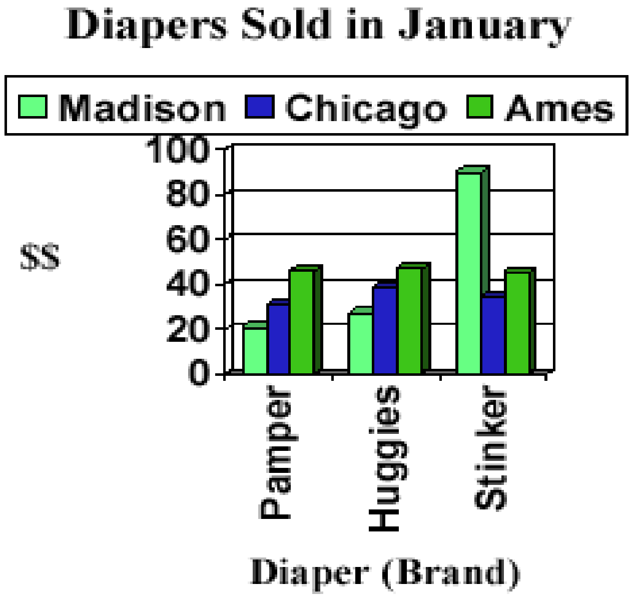
It was theorized that the reason for this was that fathers were stopping off at Wal-Mart to buy nappies for their babies, and since they could no longer go down to the pub as often, would buy beer as well. As a result of this finding, the supermarket chain is alleged to have the nappies next to the beer, resulting in increased sales of both.

The Original Version

DATA MINING EXAMPLE

Beer and Nappies/Diapers -- A Data Mining Urban Legend





What is the information?

Diapers ➔ Bear

DATA MINING EXAMPLE

Beer and Nappies/Diapers -- A Data Mining Urban Legend

The stories are embellished with plausible sounding factoids, e.g. increased sales as a percentage:

"The man decided to buy a six pack of beer while he was stopped anyway. The discount chain moved the beer and snacks such as peanuts and pretzels next to the disposable diapers and increased sales on peanuts and pretzels by more that 27%."

The now legendary revelation that men who bought diapers on Friday night at convenience stores were also likely to buy beer is an example of market basket analysis.

Outline

- Motivation: Why Data Mining?
- What is Data Mining?
- History of Data Mining
- Data Mining Terminology / KDD Process
- Data Mining Applications
- Issues in Data Mining
- Concluding Remarks

Major Issues in Data Mining

- **Mining methodology and user interaction**
  - Mining different kinds of knowledge in databases
  - Incorporation of background knowledge
  - Handling noise and incomplete data
  - Pattern evaluation: the interestingness problem
  - Expression and visualization of data mining results

Major Issues in Data Mining

- **Performance and scalability**
  - Efficiency of data mining algorithms
  - Parallel, distributed and incremental mining methods

- **Issues relating to the diversity of data types**
  - Handling relational and complex types of data
  - Mining information from diverse databases

Major Issues in Data Mining

- **Issues related to applications and social impacts**
  - Application of discovered knowledge
    - Domain-specific data mining tools
    - Intelligent query answering
    - Expert systems
    - Process control and decision making
  - A knowledge fusion problem
  - Protection of data security, integrity, and privacy

Outline

- Motivation: Why Data Mining?
- What is Data Mining?
- History of Data Mining
- Data Mining Terminology / KDD Process
- Data Mining Applications
- Issues in Data Mining
- Concluding Remarks

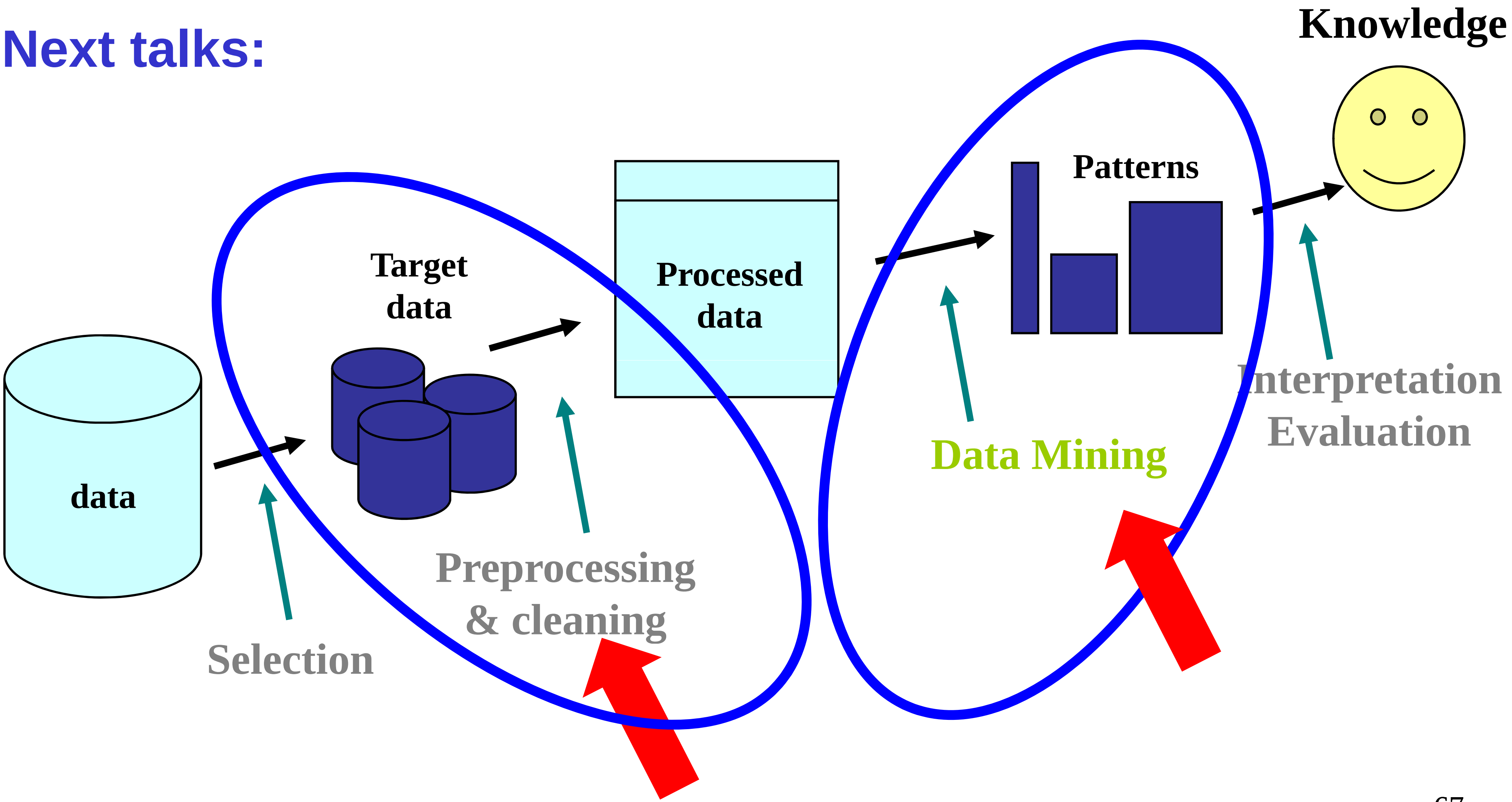
Concluding Remarks

Summary

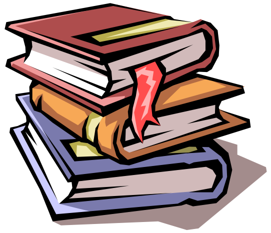
- Data mining: discovering interesting patterns from large amounts of data
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Mining can be performed in a variety of information repositories
- Data mining functionalities: characterization, association, classification, clustering, outlier and trend analysis, etc.

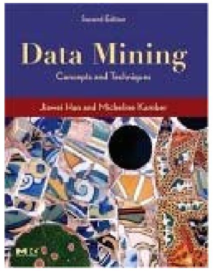
Concluding Remarks

Next talks:



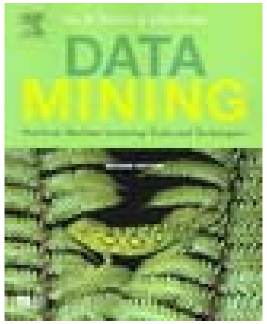
Bibliography

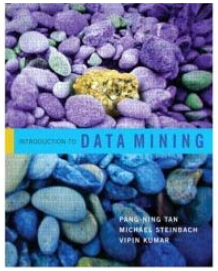




J. Han, M. Kamber.
Data Mining. Concepts and Techniques
Morgan Kaufmann, 2006 (Second Edition)
http://www.cs.sfu.ca/~han/dmbook

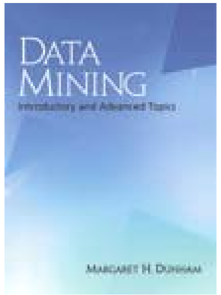
I.H. Witten, E. Frank.
Data Mining: Practical Machine Learning Tools and Techniques,
Second Edition,Morgan Kaufmann, 2005.
http://www.cs.waikato.ac.nz/~ml/weka/book.html

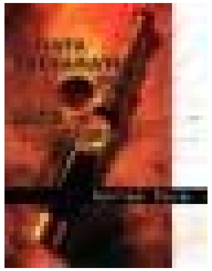




Pang-Ning Tan, Michael Steinbach, and Vipin Kumar
Introduction to Data Mining (First Edition)
Addison Wesley, (May 2, 2005)
http://www-users.cs.umn.edu/~kumar/dmbook/index.php

Margaret H. Dunham
Data Mining: Introductory and Advanced Topics
Prentice Hall, 2003
http://lyle.smu.edu/~mhd/book





Dorian Pyle
Data Preparation for Data Mining
Morgan Kaufmann, Mar 15, 1999

Mamdouh Refaat
Data Preparation for Data Mining Using SAS
Morgan Kaufmann, Sep. 29, 2006)

Summary

1. Introduction to Data Mining and Knowledge Discovery
2. Data Preparation
3. Introduction to Prediction, Classification, Clustering and Association
4. Data Mining - From the Top 10 Algorithms to the New Challenges
5. Introduction to Soft Computing. Focusing our attention in Fuzzy Logic and Evolutionary Computation
6. Soft Computing Techniques in Data Mining: Fuzzy Data Mining and Knowledge Extraction based on Evolutionary Learning
7. Genetic Fuzzy Systems: State of the Art and New Trends
8. Some Advanced Topics I: Classification with Imbalanced Data Sets
9. Some Advanced Topics II: Subgroup Discovery
10. Some advanced Topics III: Data Complexity
11. Final talk: How must I Do my Experimental Study? Design of Experiments in Data Mining/Computational Intelligence. Using Non-parametric Tests. Some Cases of Study.