cse537
Artificial Intelligence

Part 2: Data Mining

Professor Anita Wasilewska
Computer Science Department
Stony Brook University

Textbook

- Course Part 2 Textbook:

Jianwei Han, Micheline Kamber

DATA MINING

Concepts and Techniques

Morgan Kaufmann, 2006, 2011

Data Mining: Machine Learning

- **Part One:** Chapter 1: Intuitive Introduction and DM Overview of DM techniques

- **Part Two:** Chapter 2 (preprocessing),
- Chapter 6 (classification)
- Chapter 5 (association)
- Chapter 7 (clustering)

- **Part Three:** students presentations

Chapter 1: Introduction

Data Mining Main Objectives:

- Indentification of data as a source of useful information

- Use of discovered information for **competitive advantages** when working in business enviroment

Data – Information - Knowledge

- **Data** – as in databases

- **Information** or **knowledge** is a meta information ABOUT the **patterns hidden** in the data

- **The patterns** must be discovered **automatically**

Why Data Mining?

- **Data explosion problem**

  Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other data repositories

Why Data Mining?

- **Data explosion problem**
- We are drowning in data, but starving for knowledge!
- **Solution:** Data warehousing and data mining
- **Data Mining:**

  Extraction of interesting knowledge (rules, regularities, patterns, constraints) **from data in large databases**

What is Data Mining?

- There are many activities with the same name: CONFUSSION

- DM: Huge volumes of data
- DM: Potential hidden knowledge
- DM: Process of **discovery** of hidden **patterns** in data

DM: Intuitive Definition

DM is a process to extract previously unknown knowledge from large volumes of data

Requires both new technologies and new methods

Data Mining: Machine Learning

- **DM creates models (algorithms):**
- **classification** (chapter 6)
- **association** (chapter 5)
- **prediction** (chapter 6)
- **clustering** (chapter 7)
- **DM** often **presents** the knowledge as a set of rules of the form
- **IF.... THEN...**
- In this case it is called a **Descriptive Learning**
- **Finds other relationships in data**
- **Detects deviations**

DM: Business Advantages

- **DM** uses gathered data to
- **Predict** tendencies and waves
- **Classifiy** new data
- **Find** previously unknown patterns for the use for business advantages
- **Discover** unknown relationships

DM: Technologies

- **Many commercially avaible tools**
- **Many** methods (models, algorithms) for the **same task**
- TOOLS ALONE ARE NOT THE SOLUTION
- **The user must often be able to interpret the results**

DM: Technologies

- One of the **requirements** of DM is that

the results must be easily comprehensible to the user

- **This is the strenght** of **Descriptive Methods**

- Most often, especially when dealing with statistical methods separate analysts are needed to **interpret** the results (knowledge)

- **This is the weakness** of **Statistical Methods**

Data Mining vs Statistics

- Some statistical methods are considered as a part of **Data Mining** i.e. they are used as **Data Mining algorithms**, or as a part of **Data Mining algorithms**

- Some, like statistical prediction methods, different types of regression, and clustering methods are now considered as an **integral part** of **Data Mining** research and applications

Data Mining vs Data Marketing

- **Data Mining** methods apply to **many domains**

- **Data Marketing:**
- applications of Data Mining methods in which **the goal** is to find buying patterns in Transactional Data Bases

- **APPRIORI Algorithm**

Market Analysis and Management

- Determine customer purchasing patterns over time
  - Conversion of single to a joint bank account: when marriage occurs, etc.
- Cross-market analysis
  - **Associations/co-relations** between product sales
  - **Prediction** based on the association information

Market Analysis and Management

- Customer profiling
  - data mining can tell you what types of customers buy what products (clustering or classification)
- Identifying customer requirements
- identifying the best products for different customers

Corporate Analysis and Risk Management

- Finance planning and asset evaluation
  - cash flow analysis and prediction
  - contingent claim analysis to evaluate assets
  - cross-sectional and time series analysis (financial-ratio, trend analysis, etc.)
- Resource planning:
  - summarize and compare the resources and spending

Business Summary

- **Data Mining** helps to **improve** competitive advantage of organizations in dynamically changing environment;

- it **improves** clients retention and conversion

- Different **Data Mining** methods are required for different kind of data and different kinds of goals

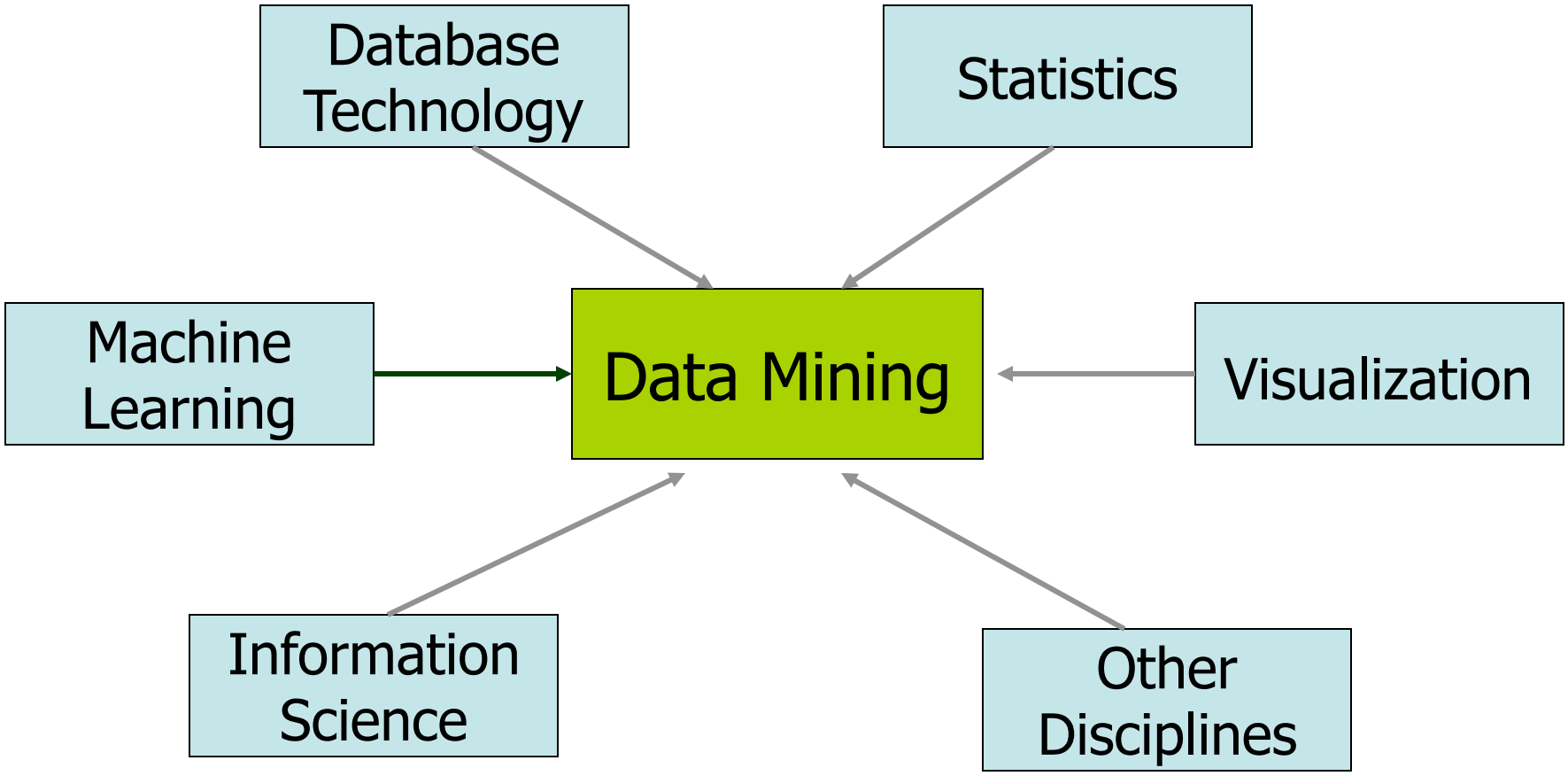
Scientific Applications

- Networks failure detection
- Controllers
- Geographic Information Systems
- Genome- Bioinformatics
- Intelligent robots
- Intelligent rooms
- etc… etc ….

What is NOT Data Mining

- Once patterns are found Data Mining process is **finished**
- Use of the patterns is not Data Mining
- Monitoring is not Data Mining
- Querries to the database are not DM

Data Mining: Confluence of Multiple Disciplines

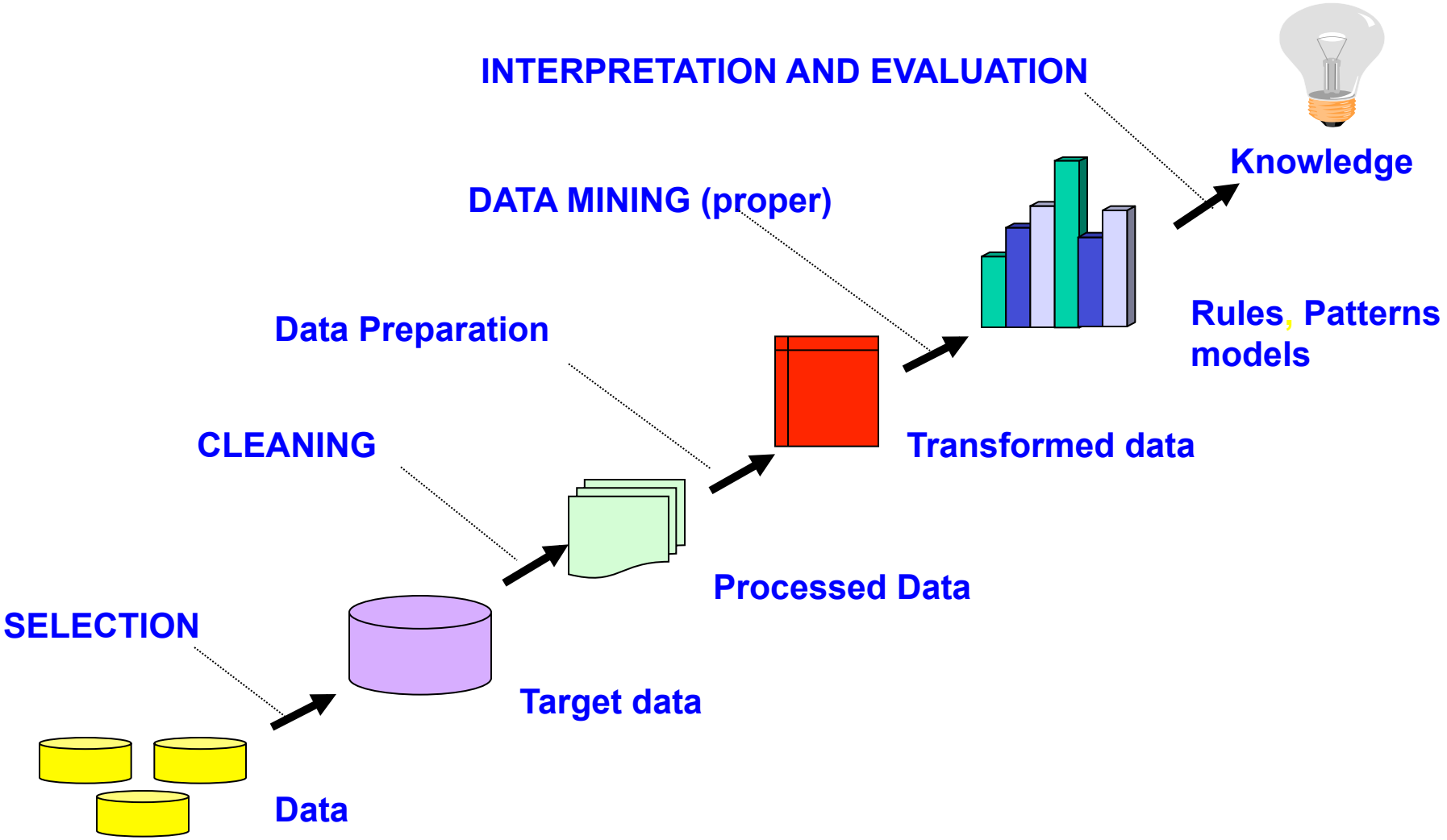


KDD process Definition

[Piatetsky-Shapiro 97]

- **KDD** is a non trivial process for identification of :
  - Valid
  - New
  - Potentially useful
  - Understable
  - patterns in data

KDD Process


INTERPRETATION AND EVALUATION
Knowledge
DATA MINING (proper)
Rules, Patterns
models
Data Preparation
Transformed data
CLEANING
Processed Data
SELECTION
Target data
Data

DM: Data Mining

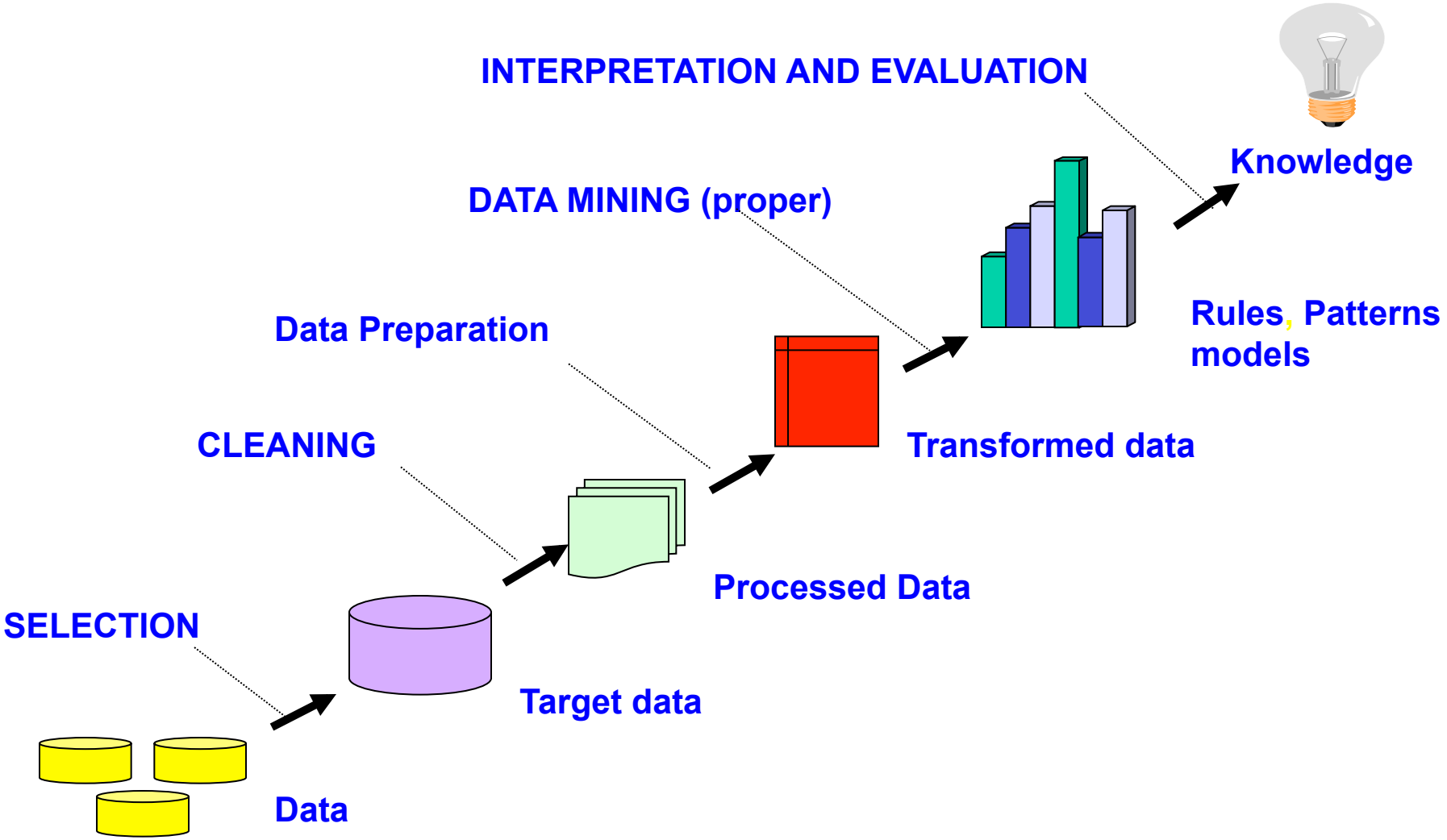
- **DM** is a step in the **KDD process** in which algorithms are applied to look for **patterns in data**
- We use term **DATA MINING PROPER** for DM step in KDD Process
- We usually use term **DM process** term for KDD process

DM: Data Mining Process

- **Rememeber**
- It is necessary to apply first
- the preprocessing operations to clean and preprocess the data in order to obtain significant patterns

- DM Process can be re-iterated- and usually is

Data Mining Process



KDD vs DM

- **KDD** is a term used by **academia**
- **DM** is a often a **commercial** term
- **DM** term is also being used in academia, as it has become a "brand name" for both KDD process and its DM sub-process
- The important point is to see Data Mining as **a process** with Data Mining Proper as part of it

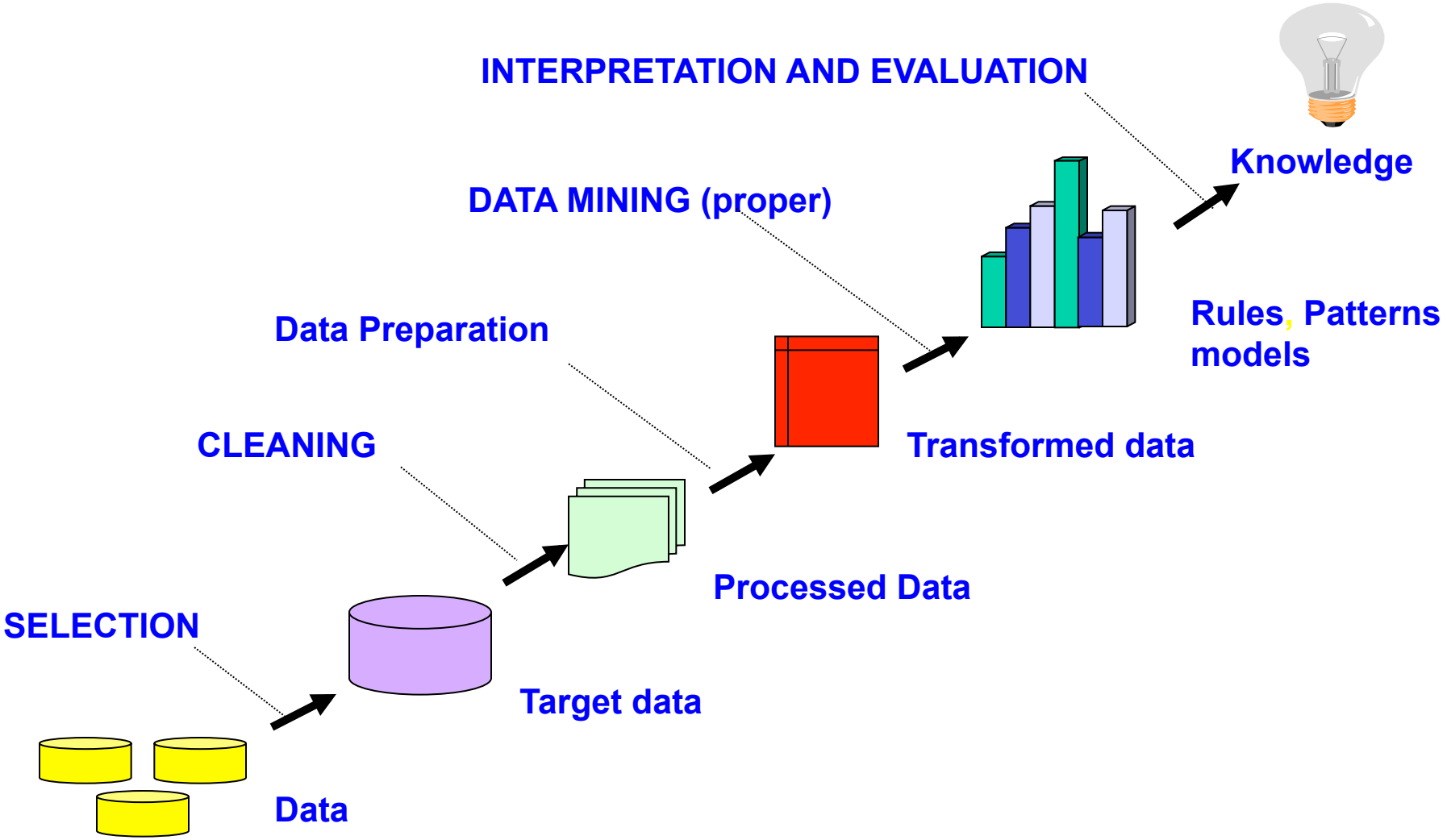
Steps of the KDD or DM process

- **Preprocessing:** includes all the operations that have to be performed **before** a data mining **algorithm is applied**

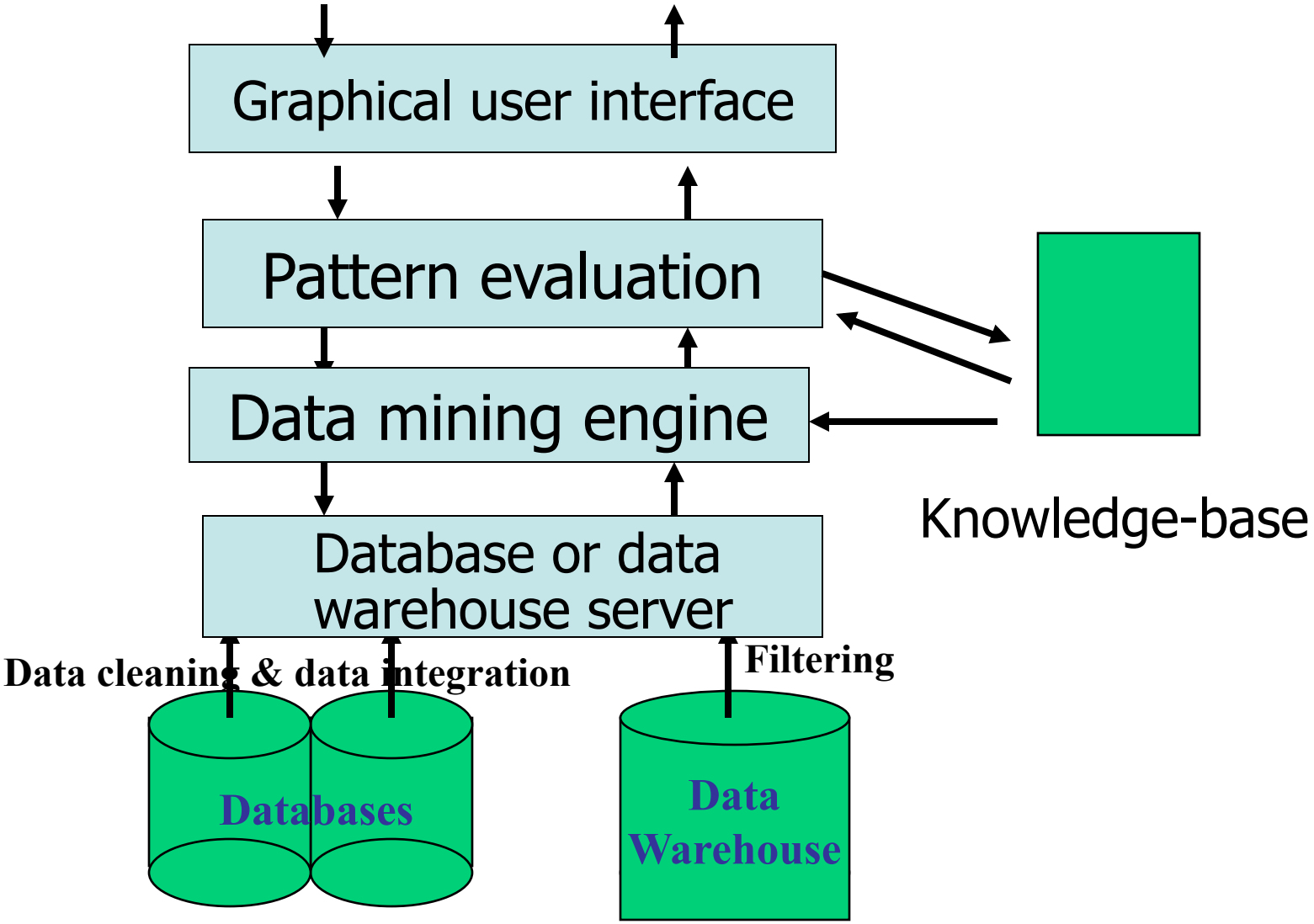
- **Data Mining (proper):** knowledge discovery algorithms are applied in order to **obtain the patterns**

- **Interpretation**: discovered patterns are presented in a proper format and the user decides if it is neccesary to re-iterate the algorthms

Data Mining Process



Architecture of a Typical Data Mining System



What Kind of Data?

- Relational Databases
- Data warehouses
- Transactional databases
- Advanced DB and information repositories
  - Object-oriented and object-relational databases
  - Spatial databases
  - Time-series data and temporal data
  - Text databases and multimedia databases
  - Heterogeneous and legacy databases
  - WWW

Descriptive Data Mining:
Concept Description

- **concept** – is defined **semantically** as **any subset of records**
- We often **define** the **concept** by distinguishing in our database **an attribute c** and **its value v**
- In this case the **concept description** is written **syntactically** as : **c=v**
- **We define** a concept **C** with the description **c=v** as

***C = {records:* c=v}**

Descriptive Data Mining:

Concept Description

- Let **C** be a concept with the description **c=v**, i.e.

  C = {records: c=v}

- We call such attribute **c** a **CLASS** attribute, or a **decision** attribute, or a **classification** attribute

- The description **c=v** is called a **class** description

Descriptive Data Mining:

Concept, Class Description

- **For example:**
- ***climate=wet*** *is a description of the concept of WET CLIMATE and*
- ***WET CLIMAT = {records:*** *climate=wet***}**
- **We use words:** decision attribute, class attribute, concept attribute
- We talk about a concept or a class **description**
- **REMEMBER: all definitions are relative to the database we deal with.**

Desctiptive DM

- Let **C** be a **class** with a description **c=v**, i.e.

  C = {records: c=v}

- *The **class C characteristics** is a set of attributes a1, a2, … ak, and their respective values v1, v2, …. vk that are **characteristic** for a given class C, i.e. such that*

- ***{**records: a1=v1 & a2=v2&…..ak=vk**}** /\ **C** = non empty set*

- ***Characteristics description** of **C** is then written as*

  a1=v1 & a2=v2&…..ak=vk

Characterization

- ***Describes the process which aim is to find rules that describe characteristic properties of a class. They take the form***

If *class* ***then*** *characteristics*

If c = v then *a2=v2&…..ak=vk*

C=1 → A=1 & B=3 **25%** (support: there are 25% of the records for which the rule is true)

- C=1 → A=1 & B=4 **17%**
- C=1 → A=0 & B=2 **16%**

Discrimination

- ***It is the process which aim is to find rules that allow us to discriminate the objects (records) belonging to a given concept (one class) from the rest of records***

If *characteristics* ***then*** *class*

If *a2=v2&…..ak=vk* ***then*** ***c = v***

- **A=0 & B=1 → C=1 33% 83% (support, confidence: the conditional probability of the concept given the characteristics)**
- **A=2 & B=0 → C=1 27% 80%**
- **A=1 & B=1 → C=1 12% 76%**

Classification - Supervised Learning

- **Classification**
- Finding models (rules) that **describe** (characterize) or/ and **distinguish** (discriminate) classes or concepts for future prediction
- **Example:** classify countries based on climate (characteristics)
- classify cars based on gas mileage and use it to predict classification of a new car

Classification Algorithms
Models, Basic Classifiers

- **Decision Trees** (ID3, C4.5) –descriptive
- **Neural Networks-** statistical
- **Bayesian Networks** - statistical
- **Rough Sets** - descriptive
- **Genetic Algorithms** – descriptive or statistical but mainly an optimization method
- **Classification by Association** – descriptive

Classification Algorithms
Models, Basic Classifiers

- **Presentation of results:**
- characteristic and /or discriminant rules
- In case of **descriptive DM**
- **converged network** (Neural, Bayes) in case of **statistical DM**

Statistical DM, Clustering

- **Statistical Prediction** - predict some unknown or missing numerical values
- **Cluster analysis (statistical)**
  - Class label is unknown
  - Goal: group data to form new classes
  - It is called **unsupervised learning**
  - **For example:** cluster houses to find distribution patterns
  - **Clustering** is based on the principle:
  - **maximizing** the intra-class similarity and **minimizing** the interclass similarity

Statistical DM

- **Outlier analysis**
  - **Outlier:** a data object that does not comply with the general behavior of the data
  - It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis

Statistical DM

- **Trend and evolution analysis (statistical)**
  - Trend and deviation: regression analysis
  - Sequential pattern mining, periodicity analysis
  - Similarity-based analysis
- **Other pattern-directed or statistical analyses**

Classification: Chapter 6

- Decision Trees (ID3, C4.5) – descriptive
- Neural Networks -statistical
- Rough Sets – descriptive
- Bayesian Networks- statistical
- Genetic Algorithms- can be both, but is mainly an optimization method

Association: Chapter 5
Problem Statement

- **I**={i1, i2, ..., in} a set of **items**
- **Transaction T:** set of items, **T** is subset of **I**
- **Data Base:** set of transactions
- An association rule is an implication of the form : **X-> Y,** where **X, Y** are **disjoint** subsets of **T**
- **Problem: Find association rules that have support and confidence greater that user-specified minimum support and minimun confidence**

Association Rules

- **Confidence:** a rule **X->Y** holds in the database **D** with a confidence c if the c % of transactions in **D** that contain **X** also contain **Y**

- **Support:** a rule **X->Y** has a support s in **D** if s% of transactions contain **XUY**

Association Rules

- Association rules presentation (predicate presentation)

  Multi-dimensional:

  age(X, "20..29") ^ income(X, "20..29K") → buys(X, "PC") [support = 2%, confidence = 60%]

  Single-dimensional:

  – **buys(x, "computer") → buys(x, "software")** [1%, 75%]

  Association rules presentation (non-predicate presentation)

  Age = 20..29 /\ income=20..29K → buys=PC (2%, 60%)

  Buys=computer → buys=software (1%,75%)

Clustering- Chapter 7

- Database segmentation
- Given a set of objects (records) the algorithm obtains a division of the objects into clusters in which the **distance** of objects **inside** a claster is minimal and the distance among objects of **diferent clusters** is maximal
- Unsupervised learning

Other Statistical Methods chapter 6

- Regression
- Temporal Series
- Lazy learners
- Support Vector Machines

.....

Major Issues in Data Mining (1)

Book Slide

- Mining methodology and user interaction
  - Mining different kinds of knowledge in databases
  - Interactive mining of knowledge at multiple levels of abstraction
  - Incorporation of background knowledge
  - Data mining query languages and ad-hoc data mining
  - Expression and visualization of data mining results

Major Issues in Data Mining (2)

Book Slide

- Handling noise and incomplete data
- Pattern evaluation: the interestingness problem

- Performance and scalability

  Efficiency and scalability of data mining algorithms

  Parallel, distributed and incremental mining methods

Major Issues in Data Mining (3)

Book Slide

- Issues relating to the diversity of data types
  - Handling relational and complex types of data
  - Mining information from heterogeneous databases and global information systems (WWW)
- Issues related to applications and social impacts
  - Application of discovered knowledge
    - Domain-specific data mining tools
    - Intelligent query answering
    - Process control and decision making
  - Integration of the discovered knowledge with existing knowledge: A knowledge fusion problem
  - Protection of data security, integrity, and privacy

Summary

- **Data mining:** discovering comprehensible, interesting patterns from large amounts of data
- **A natural evolution of database technology, in great demand, with wide applications**
- **A KDD process**, or **DM process** includes data cleaning, data integration, data selection, transformation, data mining proper, pattern evaluation, and knowledge presentation
- **Mining can be performed in a variety of information repositories**

Summary

- **Data mining functionalities**: characterization, discrimination, association, classification, clustering, outlier and trend analysis
- Classification of data mining systems
- Major issues in data mining

Preprocesing
Introduction to chapter 2

Preprocesing

- Select, **integrate**, and **clean** the data
- Decide which kind of patterns are needed
- Decide which algorithm is the best for your goal
- It depends on many factors
- Prepare data for algorithms
- Different algorithms accept differnt data format

Implementaion Preparation

- Identify the problem to be solved.
- Study problem it in detail
- Explore the solution space
- Find one acceptable solution (feasible to implement)
- Specify the solution
- Prepare the data

Preparation

- Remember GIGO! (garbage in gabage out)
- Add some data, if necessary
- Structure the data in a proper form
- Be careful with incomplete and noisy data

Some implementation preparation rules to follow

- Select the problem
- Specify the problem
- Study the data
- The problem must guide the search for tools and technologies
- Search for the simpliest model (algorithm, method)
- Define for each data the solution is valid, where it is not valid at all and where it is valid with some constraints

Studying the data

- The surrounding world consists of objects ,(data) and the DM problem is to find the relationships among objects
- The objects are characterized by properties:
- attributes, values of attributes
- that have to be analized
- **The results** (rules, descriptions) are valid

(true) under certain circumstances (data) and in certain moments (avaible data at the moment)

Measures

- Type of data decides a way in which data are analized and preprocessed
  - Names (attributes)
  - Categories, classes, class attributes
  - Ordered values of attributes
  - Intervals of values of attributes
  - Types of values of attributes

Types of data

- Generally we distinguish:
  - Quantitative Data
  - Qualitative Data
- Bivaluated: often very useful
- Null Values are not applicable
- Missing data usually **not acceptable**
- NNetworks, and Bayes accepts some missing data.

What to take into account

- Eliminate redundant records
- Eliminate out of range values of attributes
- **Decide a generalization level**
- Consistency

Other preprocessing tasks

- Generalization vs specification
- Discretization
- Sampling
- Reducing number of attributes at the preprocessing stage

Summary

- The preprocessing is required and is an **essential** part of the DM process
- **If preprocessing is not performed patterns obtained could be of no use**
- Preprocessing is a tedious task that could even take more time that DM proper

APPROACHES TO DATA MINING

Aproaches

- **Mathematics**: Consist in the creation of mathematical models, algorithms, methods, to extract rules, regularities and patterns
- Rough Sets is the most precise model

- **Statistics:** They are focused in the creation of statistical models to analise data
- Regression, Bayesian networks, NN, Clustering

Statistical methods

- Numerical data are needed
- Statistical methods are also often used in preprocessing steps to study the sample
- Hypothesis validation and regression analisys are used in data mining steps of the process

Decision trees

- Discovering **discriminant rules**, i.e.
- Descriptive Data Mining
- **Method:** succesive division of the set of data
- This is a classification algorithm
- Works better when attributes

have a small set of values

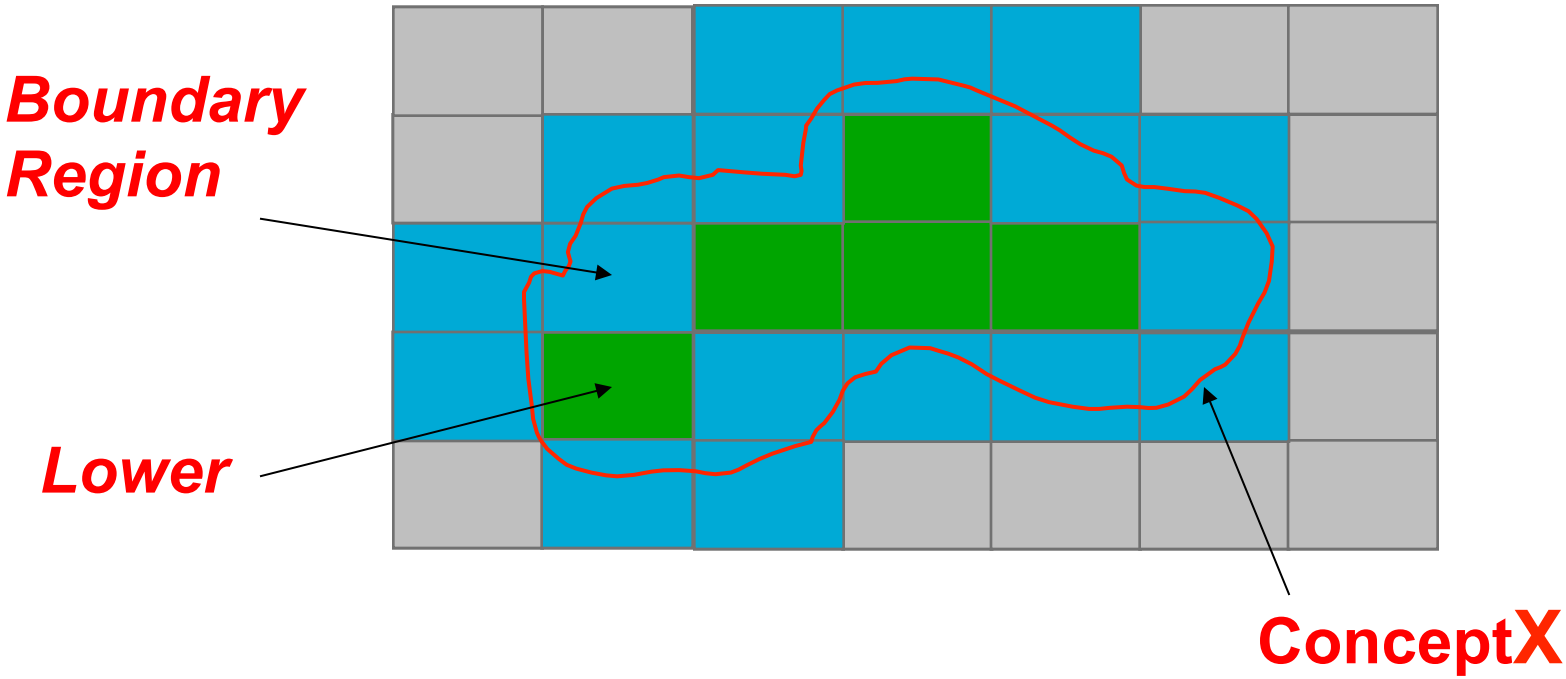
Apriori Algorithm

- Agrawal, Imielinski (IBM S. José. California)
- It is an intuitive and efficient algorithm to extract associations from transactions
- Also used as classification algorithm
- classification by association
- **Method:**
- Iterates until the associations obtained don't have the required support

Rough Sets
Descriptive Classification

- Approximation space A=(U,IND(B)):
  - *Lower Approximation* $\underline{X}_B = \{o \in U / [o] \subseteq X\}$
  - *Upper Approximation* $\overline{X}_B = \{o \in U / [o] \cap X \neq \emptyset\}$
  - *Boundary* Region $\mathrm{Bnd(X)_B} = \overline{X}_B - \underline{X}_B$
  - Positive Region: $\mathrm{POS_B(D)} = \bigcup \{\overline{X} : X \in IND(D)\}$

Rough Sets



Boundary + *Lower* = *Upper*

Variable Precision Rough Set Model

$$c(X,Y) = \begin{cases} 0 & \text{if } card(X) = 0 \\ 1 - card(X \cap Y)/card(X) & \text{if } card(X) > 0 \end{cases}$$

Rough Sets in SQL

```
Begin UPPER
setdb(dbName);
exec(conn,"BEGIN");

"DECLARE clases CLASES FOR
SELECT C1,....,CN, D, COUNT (*) AS cnt
FROM R
GROUP BY C1,....,CN, D
ORDER BY C1,....,CN, D, CNT desc");

while  not_end_records() do
    equ_class=exec("FETCH 1 IN cursor");
    first_decision_value=get_value(equ_class("D"));
    insert(equ_class,upper[first_decision_value]);
    while (equ_class == exec("FETCH 1 IN cursor") do
        decision_value=get_value(equ_class("D"));
        insert(equ_class,upper[first_decision_value]);
    end while
end while
End UPPER
```

Statistical Methods

- **Neural Network:** statistical CLASSIFICATION algorithm
- the network is trained to obtain **classification patterns**

- **Clustering:** form groups of objects without any previous hypothesis

Genetic Algorithms

- **Optimization method**
- They should be used when the goal is to find an optimal solution in solution space
- They often are used together with neural netwoks, or other methods to produce more understable (optimal) outputs
- They also are used to find the **optimal set of discriminant and/or characteristic rules** for a given database and a given class

Classification: requirements

- Decision attribute; called also class attribute, concept attribute
- Condition attributes: rest of the attributes or its subset
- Some require numerical data but there are algorithms to deal with any kind of data

Asociation: requirements

- Transactional data
- There is not needed to specify right and left side of the rules
- There are algorithms to tackle any kind of data
- Minimum support
- Maximum number of rules to be obtained

Clustering: requirements

- Set of attributes
- Maximum number of clusters
- Number of iterations
- Mimimun number of elements in any cluster