cse643, cse590
Data Mining

Anita Wasilewska
Computer Science Department
Stony Brook University
Stony Brook NY 11794

Course Textbook

- Course Textbook:

Jianwei Han, Micheline Kamber

DATA MINING

Concepts and Techniques

Morgan Kaufmann, 2006

Second Edition

Course structure

- **Part One:** chapter 1: Intuitive Introduction and DM Overview
- **Part Two:** Textbook chapter 2 (preprocessing),
- Chapter 6 (classification)
- Chapter 5 (association)
- Chapter 7 (clustering)
- **Part Three:** students presentations 1
- book chapter 3 (data warehouse) and some parts of chapters 2, 5, 6, 7, 8, 10, 11
- **Part Four :** students presentations 2
- Research papers of students' choice
- **Part Five:** project presentation (10 minutes)

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, datawarehouses and other information repositories

Why DM? (c.d.)

- **Data explosion problem** (c.d.)
- We are drowning in data, but starving for knowledge!
- **Solution:** Data warehousing and 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 volumes of data**
- **Requires both new technologies and methods**

Data Mining

- **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 (descriptive DM)**
  IF.... THEN...
- **Finds other relationships in data**
- **Detects deviations**

DM Some Applications

- **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

DM Other Applications

- **Other Applications**
  - Text mining (news group, email, documents) and Web analysis.
  - Intelligent query answering
  - Scientific Applications

DM: Business Advantages

- Data Mining uses gathered data to
- **Predicts** tendencies and waves
- **Classifies** new data
- Find previously unknown patterns
- 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 be able to interpret the results; one of the requirements of DM is:**

“the results must be easily comprehensible to the user”

Strenght of Descriptive Methods

- **Most often, especially when dealing with statistical methods analysts are needed to interpret the knowledge – 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 of different types of regression and clustering methods are now considered as an integral part of Data Mining research and applications

Bussiness Applications

- Buying patterns
- Fraud detection
- Customer Campaings
- Decision support
- Medical aplications
- Marketing
- and more

Fraud Detection and Management (B1)

- **Applications**
  - **widely used in health care, retail, credit card services, telecommunications (phone card fraud), etc.**
- **Approach**
  - **use historical data to build models of fraudulent behavior and use data mining to help identify similar instances**

Fraud Detection and Management (B2)

- Examples
  - **auto insurance**: detect characteristics of group of people who stage accidents to collect on insurance
  - **money laundering**: detect characteristics of suspicious money transactions (US Treasury's Financial Crimes Enforcement Network)
  - **medical insurance**: detect characteristics of fraudolant patients and doctors

Fraud Detection and Management (B3)

- **Detecting inappropriate medical treatment**
  - Australian Health Insurance Commission detected that in many cases blanket screening tests were requested (save Australian $1m/yr).
- **Detecting telephone fraud**
  - DM builds telephone call model: destination of the call, duration, time of day or week. Detects patterns that deviate from an expected norm.
  - British Telecom identified discrete groups of callers with frequent intra-group calls, especially mobile phones, and broke a multimillion dollar fraud.

Fraud Detection and Management (B4)

- **Retail**
  - Analysts used Data Mining techniques to estimate that 38% of retail shrink is due to dishonest employees
  - and more….

Data Mining vs Data Marketing

- Data Mining methods apply to many domains
- Applications of Data Mining methods in which the goal is to find buying patterns in Transactional Data Bases has been named: Data Marketing

Market Analysis and Management (MA1)

- Where are the data sources for analysis?
  - Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies
- Target marketing
  - DM finds clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.

Market Analysis and Management (MA2)

- 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 (MA3)

- 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 (CA1)

- 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

Corporate Analysis and Risk Management (CA2)

- Competition:
  - monitor competitors and market directions
  - group customers into classes and a class-based pricing procedure
  - set pricing strategy in a highly competitive market

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 requiered 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 ….

Other Applications

- Sports
  - 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
  - And more …..

What is NOT Data Mining

- Once the patterns are found Data Mining process is finished
- The use of the patterns is not Data Mining
- Monitoring is not analysis
- Querries to the database are not DM

Evolution of Database Technology

- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation

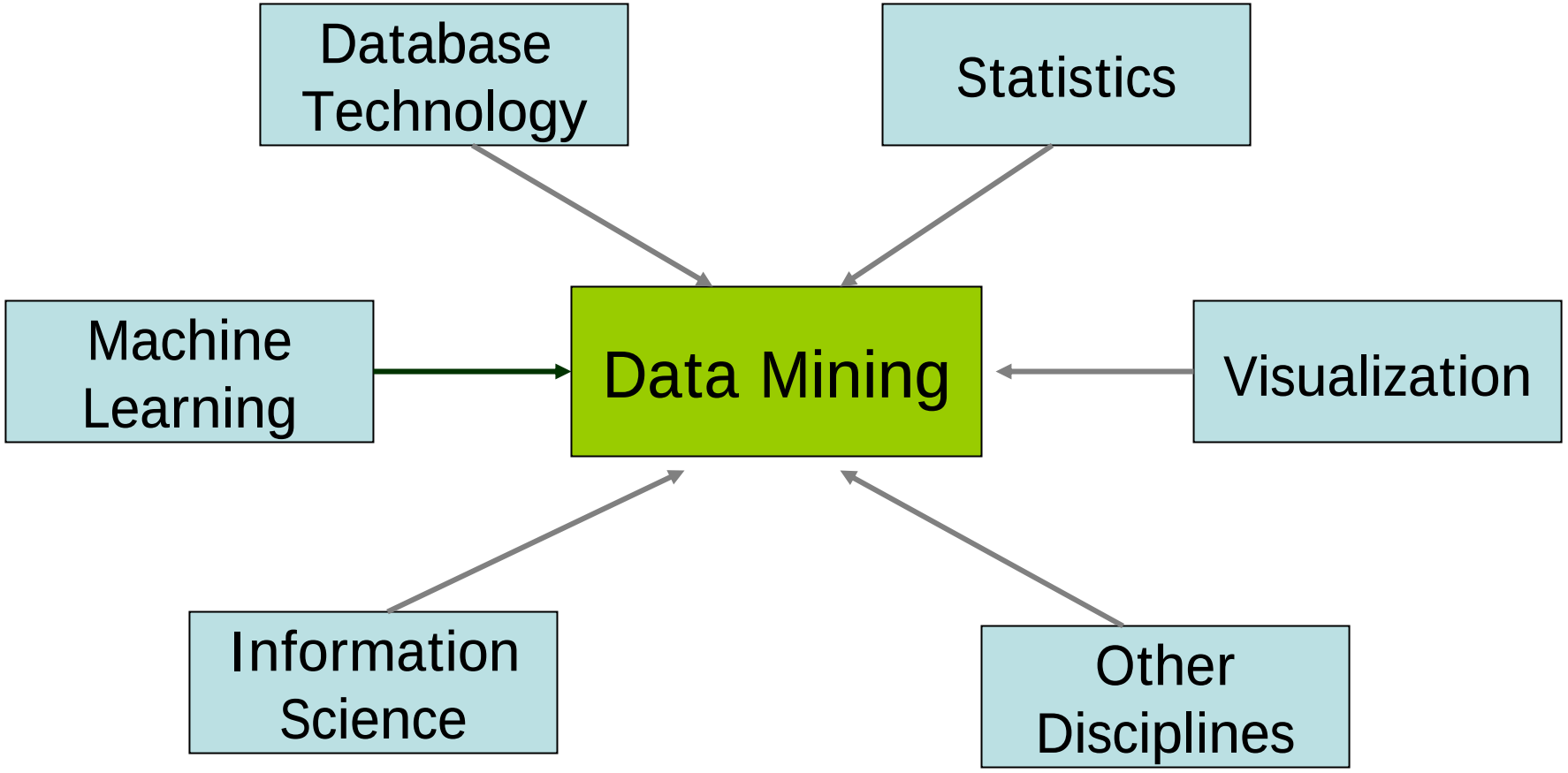
Evolution of Database Technology

- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.) and application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s—2000s:
  - Data mining and data warehousing, multimedia databases, and Web databases

Short History of Data Mining

- **1989 - KDD term (Knowledge Discovery in Databases)** appears in (IJCAI Workshop)
- **1991 -** a collection of research papers edited by Piatetsky-Shapiro and Frawley
- **1993 – Association Rule Mining Algorithm APRIORI** proposed by Agraval, Imielinski and Swami.
- **1996 – present: KDD** evolves as a conjuction of different knowledge areas (data bases, machine learning, statistics, artificial intelligence) and the term **Data Mining** becomes popular

Data Mining: Confluence of Multiple Disciplines

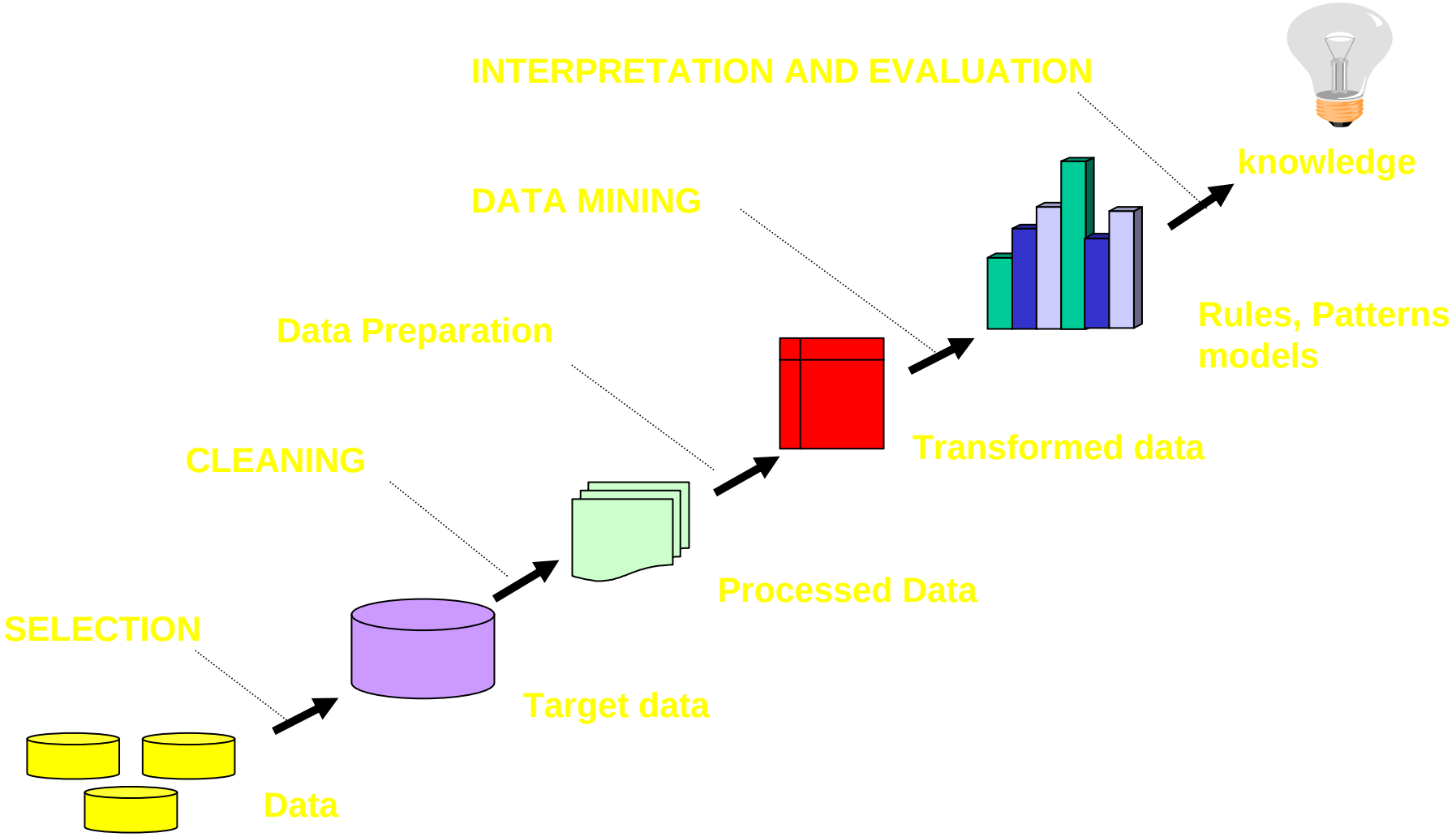


KDD process Definition

[Piatetsky-Shapiro 97]

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

The KDD process



DM: Data Mining

- **DM** is a step of the KDD process in which algorithms are applied to look for patterns in data
- We use term DATA MINING PROPER for DM the step in KDD Process
- We use DATA MINING process term for KDD process
- It is necessary to apply first the preprocessing operation to clean and preprocess the data in order to obtain significant patterns

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

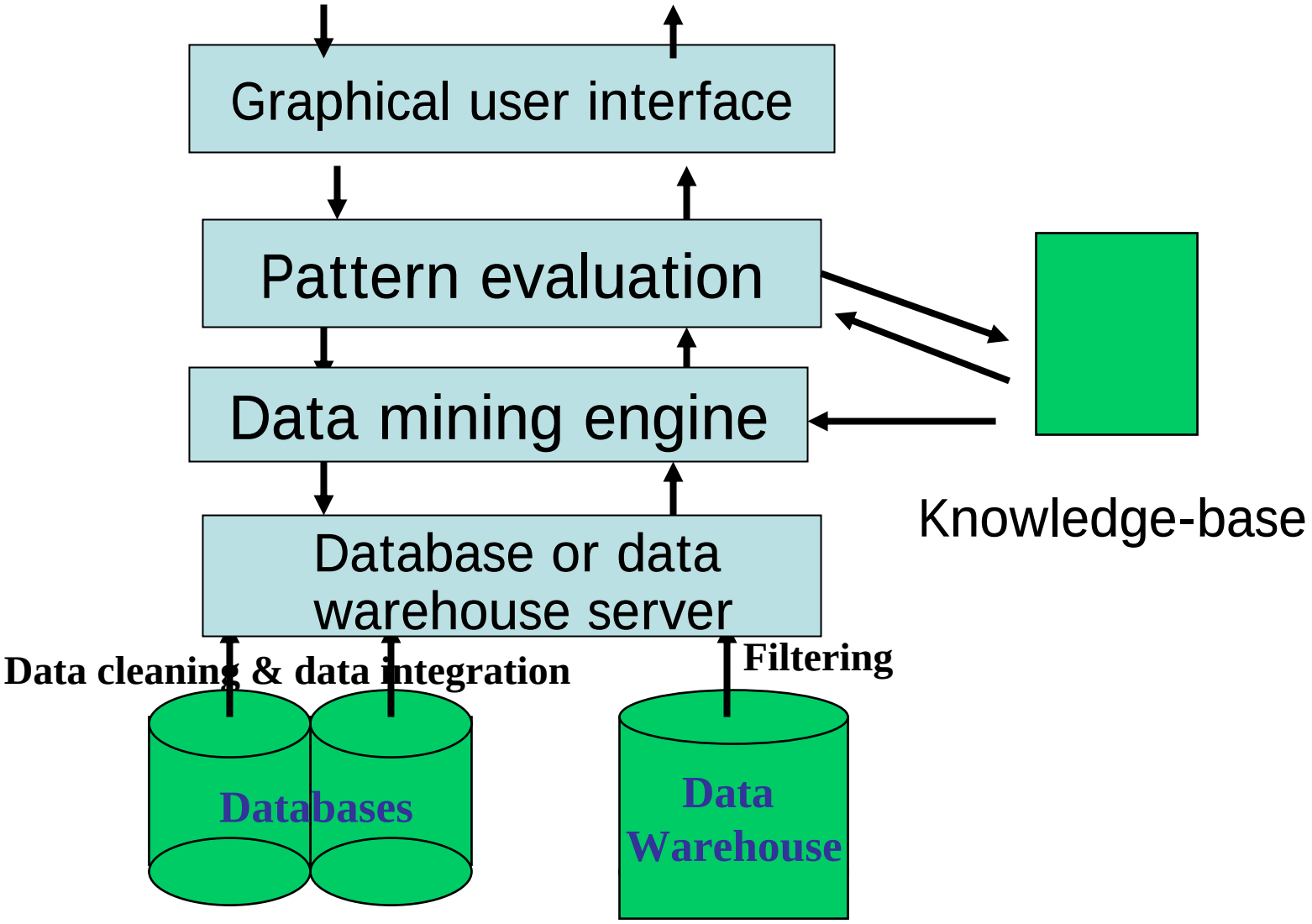
Steps of the KDD (or DM) process

- **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

Architecture of a Typical Data Mining System



Data Mining: On 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, class description

- **Concept – is defined semantically as any subset of records.**
- **We often define the concept by distinguishing a concept, or CLASS, decision, attribute c and its value v**
- **In this case the concept (class) description is syntactically written as : c=v and we define:**
- ***CONCEPT={records: c=v}***
- **For example: *climate=wet* *(description of the concept of WET CLIMATE)***
- ***CONCEPT={records: climate=wet}***
- **We use word: decision attribute, class attribute, concept attribute**
- **We talk about concept, class, or decision DESCRIPTION c=v,**
- **Hence descriptive data mining**

REMEMBER: all definitions are relative to the database we deal with.

Desctiptive DM
Concept characteristics

- ***Concept C characteristics* is a set of attributes *a1, a2, … ak,* and their respective values *v1, v2, …. vk* that are characteristic for a given concept *c* , such that**
- ***{**records: a1=v1 & a2=v2&…..ak=vk**}* /\ *C is a non empty set***
- ***Characteristics description of C** is then syntactically written as*

  a1=v1 & a2=v2&…..ak=vk

Characterization

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

If *concept* ***then*** *characteristics*

- C=1 → A=1 & B=3 **25%** (support: there are 25% o 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 (classes)***

–If characteristics then concept

- **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%**
- Discriminant rule can be good even if it has a low support (and high confidence)

Classification, Supervised learning

- **Classification and Classification Prediction is called Supervised learning**
  - **CLASSIFICATION (supervised learning):**
  - Finding models (rules) that describe (characterize) or/ and distinguish (discriminate) classes or concepts for future prediction
  - **Example:** classify countries based on climate (characteristics), or classify cars based on gas mileage and use it to predict classification of a new car
  - **Models, algorithms, methods:** decision-tree (descriptive), neural network, Bayes Network (statistical) , Rough Sets (descriptive), genetic algorithms
  - **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: Group data to form new classes- **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 (statistical)**
  - **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
Supervised Learning

- Given a set of objects (concept, class) described by a concept attribute or a set of attributes, a classification algorithms builds a set of discriminant and /or characterization rules (or other descriptions in case of Statistical methods) in order to be able,
- as the next step, to classify unknown sets of objects
- This is also called a supervised learning

Classification Methods, Models, Algorithms

(Chapter 6)

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

Association Model (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 taht 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 vs. single-dimensional association
  - **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%)**

Association Rules

- The problem of association rule discovery can be split into three sub-problems:

1. **Find the set of products (records) that have the minimum support required**
2. **Find frequent sets**
3. **Use the frequent sets to generate rules**

Clustering

- 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)

- 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)

- 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)

- 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, etc.
- 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.
- It depends on many factors
- Prepare data for algorithms
- Different algorithms accept differnt data format

Implementaion Preparation

- Identify the problem to be solved.
- Study 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
- Especify 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 usually 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
- Descriptive statistics is also often used in preprocessing steps to study the sample
- Hypothesis validation and regression analisys are used in data minign steps of the process

Decision trees

- Discovering discriminant rules
- DESCRIPTIVE Data Mining
- Succesive division of the set of data
- This is a CLASSIFICATION method
- They work better when attributes

have a small set of values

A priori Algorithm

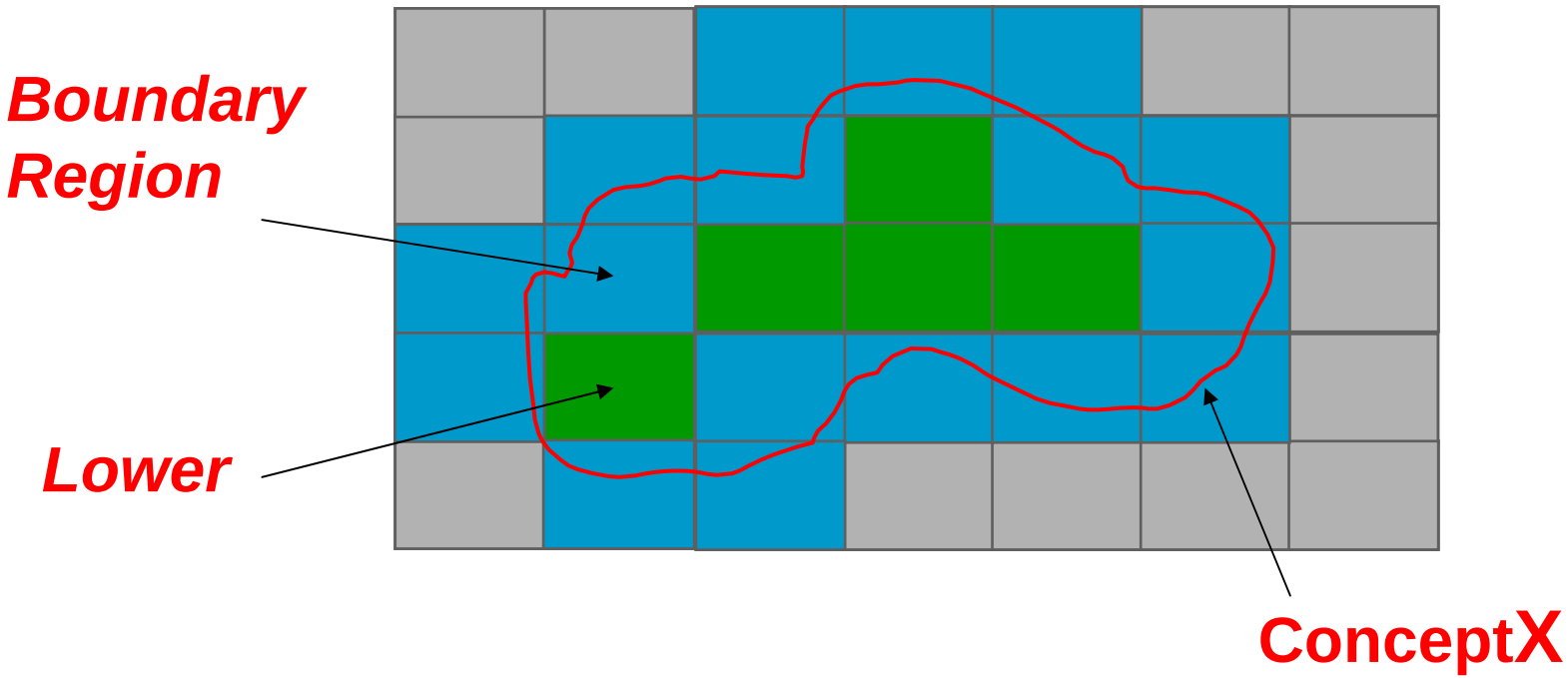
- Agrawal (IBM S. José. California).
- It is an intuitive and efficient algorithm to extract associations from transactions
- Also used as Classification (descriptive) algorithm (classification by association)

- Iterates until the associations obtained don’t have the requiered support

Rough Sets
Descriptive Classification

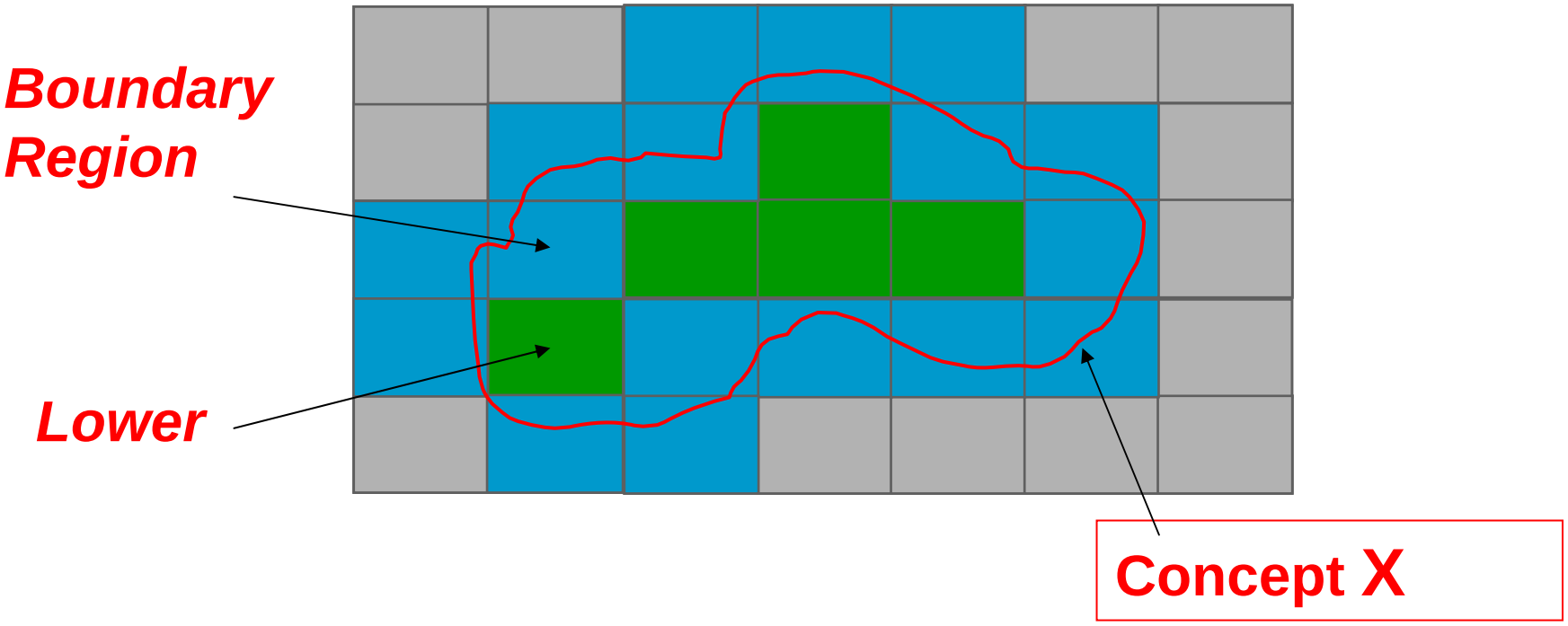
- Approximation space A=(U,IND(B)):
  - *Lower Approximation* $\underline{X}_B = \{o \in U / [o] \subseteq X\}$
  - *Upper Approximation* $\overline{X}_B = \{o \in U / [o] \cap X \neq \emptyset\}$
  - *Boundary* Region $\mathrm{Bnd(X)_B} = \overline{X}_B - \underline{X}_B$
  - Positive Region: $\mathrm{POS_B(D)}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

Rough Sets



Boundary + *Lower* = *Upper*

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 discrimoinant and/or characteristic rules for a given database and given class (set of classes)

Classification: requirements

- Decision attribute; called also class, 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

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

Clustering: requirements

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