

# ***Analyse des Données***

***Master 2 IMAFA***

---



**Andrea G. B. Tettamanzi**

Université Nice Sophia Antipolis

UFR Sciences - Département Informatique

[andrea.tettamanzi@unice.fr](mailto:andrea.tettamanzi@unice.fr)

# Agenda

- Introduction : le processus d'extraction de connaissances
- Pré-traitement
- Intégration et réduction
- OLAP et entrepôts de données
- Règles d'association
- Classification et prédiction
- Regroupement (clustering)
- Quelques outils : R, Weka
- Applications à la finance

*CM - Séance 1 – Partie A*

# **Introduction**

# *Plan*

- Motivation: Why data mining?
- What is data mining?
- Data Mining: On what kind of data?
- Data mining functionality
- Classification of data mining systems
- Top-10 most popular data mining algorithms
- Major issues in data mining

# *Why Data Mining?—Potential Applications*

- Data analysis and decision support
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)
- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining
  - Bioinformatics and bio-data analysis

# *Ex. 1: Market Analysis and Management*

- Where does the data come from?—Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies
- Target marketing
  - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
  - Determine customer purchasing patterns over time
- Cross-market analysis—Find associations/co-relations between product sales, & predict based on such association
- Customer profiling—What types of customers buy what products (clustering or classification)
- Customer requirement analysis
  - Identify the best products for different groups of customers
  - Predict what factors will attract new customers
- Provision of summary information
  - Multidimensional summary reports
  - Statistical summary information (data central tendency and variation)

## *Ex. 2: Corporate Analysis & 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
- Competition
  - monitor competitors and market directions
  - group customers into classes and a class-based pricing procedure
  - set pricing strategy in a highly competitive market

## Ex. 3: Fraud Detection & Mining Unusual Patterns

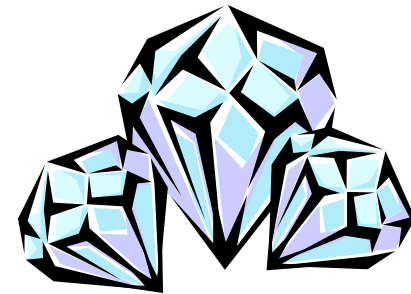
- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications: Health care, retail, credit card service, telecomm.
  - Auto insurance: ring of collisions
  - Money laundering: suspicious monetary transactions
  - Medical insurance
    - Professional patients, ring of doctors, and ring of references
    - Unnecessary or correlated screening tests
  - Telecommunications: phone-call fraud
    - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
  - Retail industry
    - Analysts estimate that 38% of retail shrink is due to dishonest employees



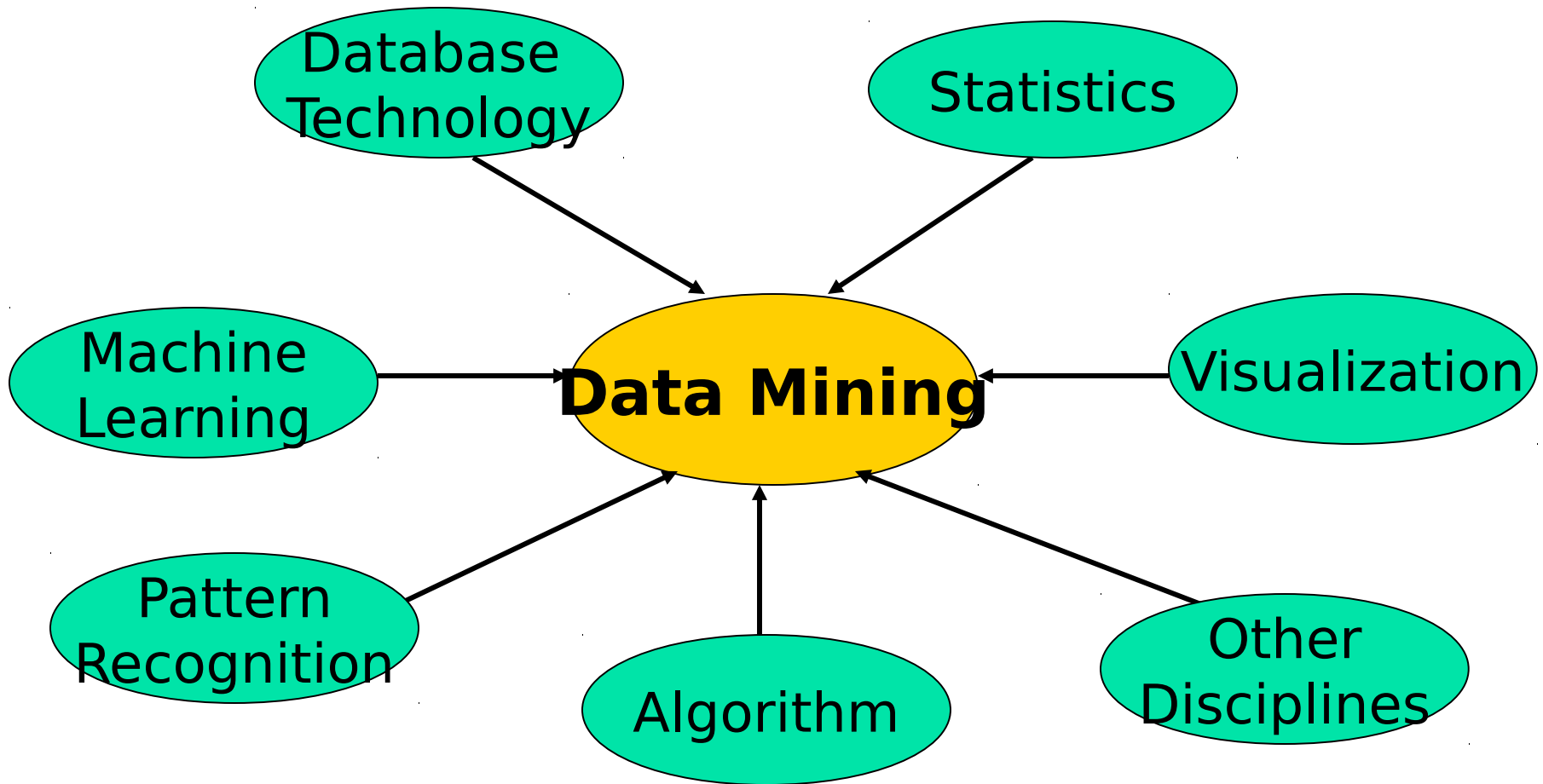
# *What Is Data Mining?*



- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything “data mining”?
  - Simple search and query processing
  - (Deductive) expert systems

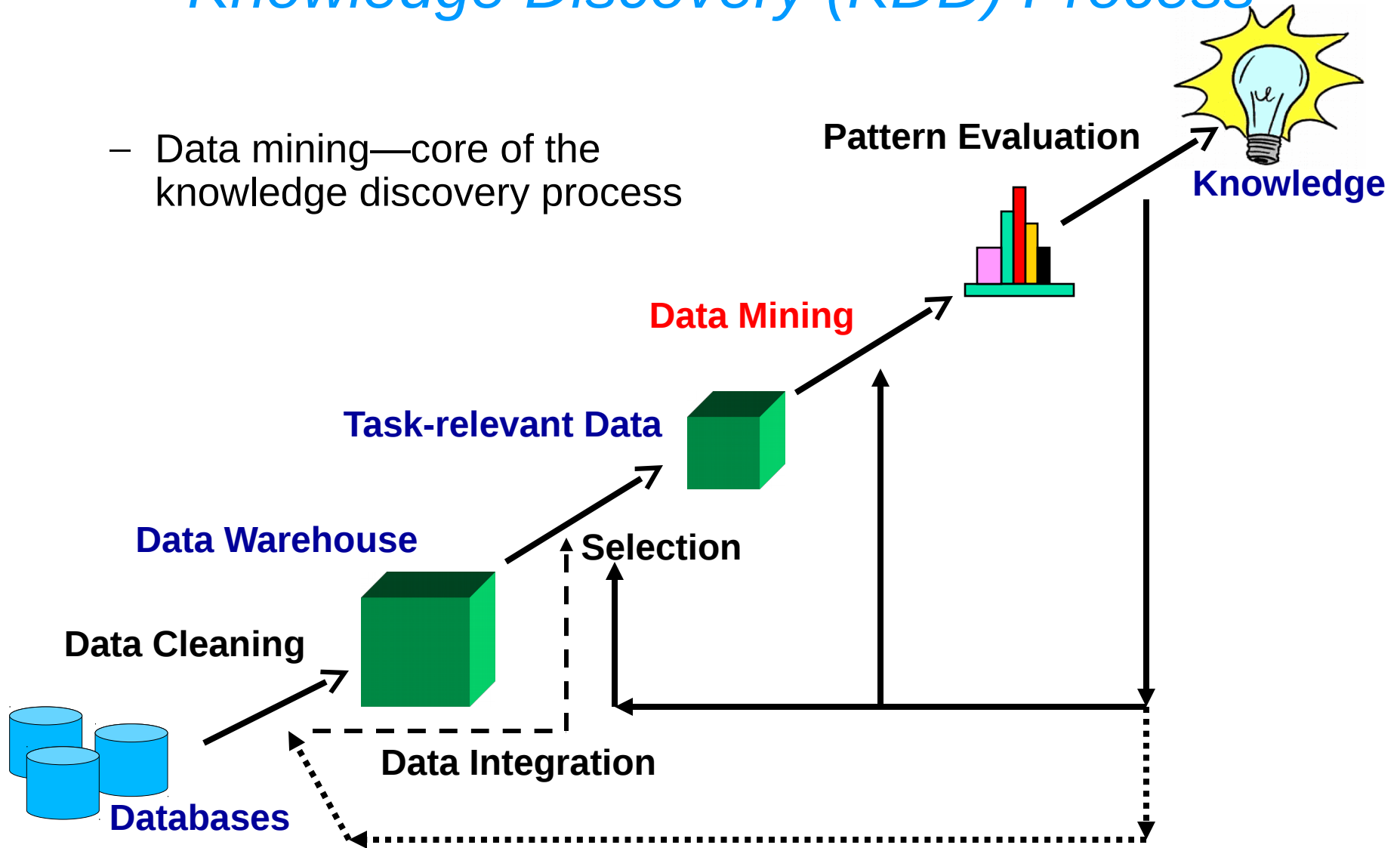


# *Data Mining: Confluence of Multiple Disciplines*



# Knowledge Discovery (KDD) Process

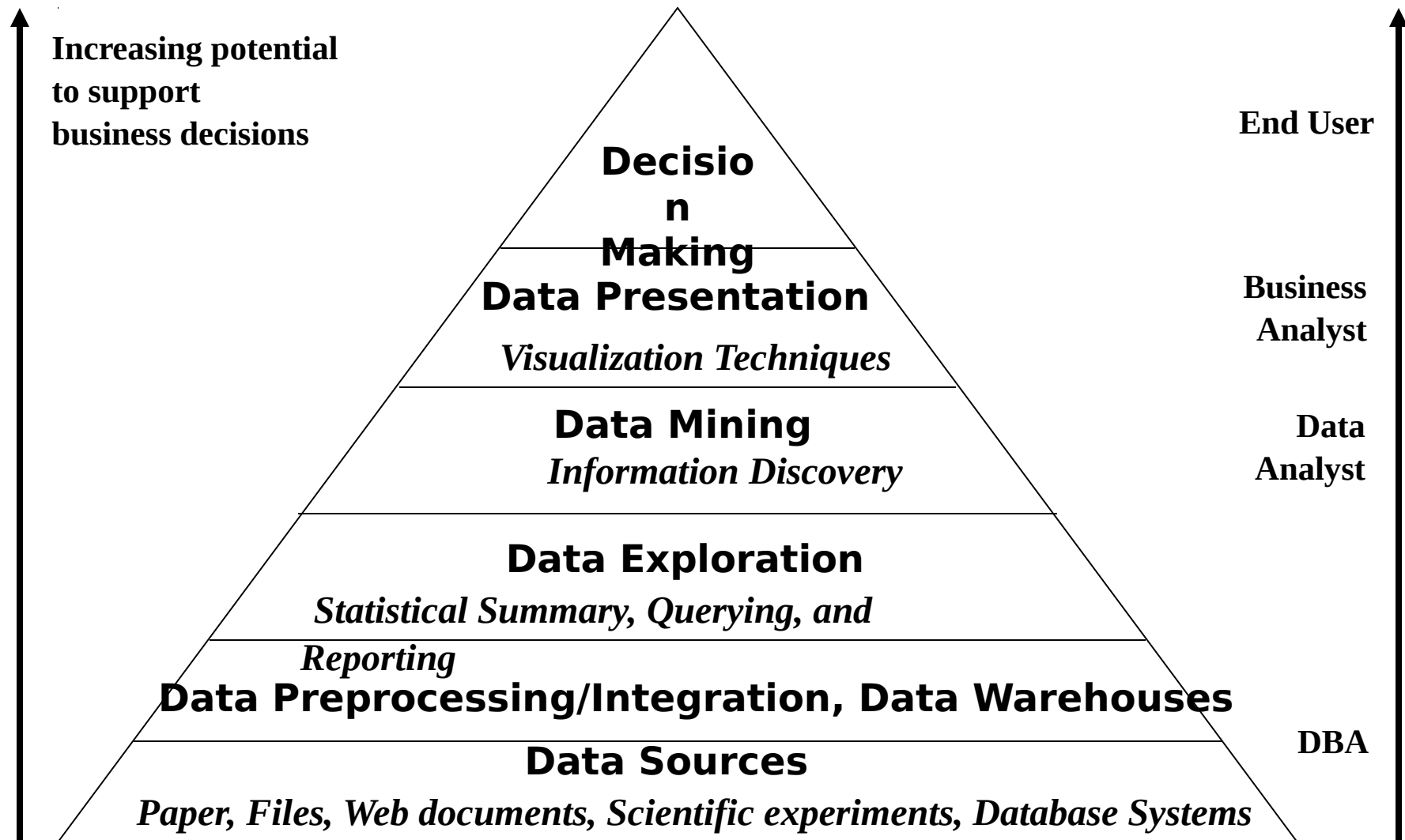
- Data mining—core of the knowledge discovery process



# *KDD Process: Several Key Steps*

- Learning the application domain
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- **Data cleaning** and preprocessing: (may take 60% of effort!)
- **Data reduction and transformation**
  - Find useful features, dimensionality/variable reduction, invariant representation
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering
- Choosing the mining algorithm(s)
- **Data mining**: search for patterns of interest
- **Pattern evaluation and knowledge presentation**
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

# Data Mining and Business Intelligence



# *Why Not Traditional Data Analysis?*

- Tremendous amount of data
  - Algorithms must be highly scalable to handle tera-bytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

# *Multi-Dimensional View of Data Mining*

- **Data to be mined**
  - Relational, data warehouse, transactional, stream, object-oriented/relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW
- **Knowledge to be mined**
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

# *Data Mining: Classification Schemes*

- General functionality
  - Descriptive data mining
  - Predictive data mining
- Different views lead to different classifications
  - **Data** view: Kinds of data to be mined
  - **Knowledge** view: Kinds of knowledge to be discovered
  - **Method** view: Kinds of techniques utilized
  - **Application** view: Kinds of applications adapted



# *Data Mining: On What Kinds of Data?*

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

# Data Mining Functionalities

- Multidimensional concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
- Frequent patterns, association, correlation vs. causality
  - Diaper → Beer [0.5%, 75%] (Correlation or causality?)
- Classification and prediction
  - Construct models (functions) that describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown or missing numerical values

# Data Mining Functionalities (2)

- Cluster analysis
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Maximizing intra-class similarity & minimizing interclass similarity
- Outlier analysis
  - Outlier: Data object that does not comply with the general behavior of the data
  - Noise or exception? Useful in fraud detection, rare events analysis
- Trend and evolution analysis
  - Trend and deviation: e.g., regression analysis
  - Sequential pattern mining: e.g., digital camera → large SD memory
  - Periodicity analysis
  - Similarity-based analysis
- ~~Other pattern-directed or statistical analyses~~

# Major Issues in Data Mining

- Mining methodology
  - Mining different kinds of knowledge from diverse data types, e.g., bio, stream, Web
  - Performance: efficiency, effectiveness, and scalability
  - Pattern evaluation: the interestingness problem
  - Incorporation of background knowledge
  - Handling noise and incomplete data
  - Parallel, distributed and incremental mining methods
  - Integration of the discovered knowledge with existing one: knowledge fusion
- User interaction
  - Data mining query languages and ad-hoc mining
  - Expression and visualization of data mining results
  - Interactive mining of knowledge at multiple levels of abstraction
- Applications and social impacts
  - Domain-specific data mining & invisible data mining
  - Protection of data security, integrity, and privacy

# Are All the “Discovered” Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting
  - Suggested approach: Human-centered, query-based, focused mining
- **Interestingness measures**
  - A pattern is **interesting** if it is **easily understood** by humans, **valid** on new or test data with some degree of **certainty**, **potentially useful**, **novel**, or **validates some hypothesis** that a user seeks to confirm
- **Objective vs. subjective interestingness measures**
  - **Objective**: based on **statistics and structures of patterns**, e.g., support, confidence, etc.
  - **Subjective**: based on **user’s belief** in the data, e.g., unexpectedness, novelty, actionability, etc.

# *Recommended Reference Books*

- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2nd ed., 2006
- D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005

## *CM - Séance 1 – Partie B*

# **Pré-traitement**

# *Major Tasks in Data Preprocessing*

- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration
  - Integration of multiple databases, data cubes, or files
- Data transformation
  - Normalization and aggregation
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results
- Data discretization
  - Part of data reduction but with particular importance, especially for numerical data



# Data Cleaning

- Importance
  - “Data cleaning is one of the three biggest problems in data warehousing”—Ralph Kimball
  - “Data cleaning is the number one problem in data warehousing”—DCI survey
- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

# Missing Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred.

# How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably.
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - **the most probable value: inference-based such as Bayesian formula or decision tree**

# Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which requires data cleaning
  - duplicate records
  - incomplete data

# *How to Handle Noisy Data?*

- **Binning**
  - first sort data and partition into (equal-frequency) bins
  - then one can **smooth by bin means, smooth by bin median, smooth by bin boundaries**, etc.
- **Regression**
  - smooth by fitting the data into regression functions
- **Clustering**
  - detect and remove outliers
- **Combined computer and human inspection**
  - detect suspicious values and check by human (e.g., **deal with possible outliers**)

# Simple Discretization Methods: Binning

- **Equal-width** (distance) partitioning
  - Divides the range into  $N$  intervals of equal size: uniform grid
  - if  $A$  and  $B$  are the lowest and highest values of the attribute, the width of intervals will be:  $W = (B - A)/N$ .
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well
- **Equal-depth** (frequency) partitioning
  - Divides the range into  $N$  intervals, each containing approximately same number of samples
  - Good data scaling
  - ~~Managing categorical attributes can be tricky~~

# Data Cleaning as a Process

- Data discrepancy detection
  - Use metadata (e.g., domain, range, dependency, distribution)
  - Check field overloading
  - Check uniqueness rule, consecutive rule and null rule
  - Use commercial tools
    - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
    - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)
- Data migration and integration
  - Data migration tools: allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface
- Integration of the two processes
  - Iterative and interactive (e.g., Potter's Wheel [Raman 2001])

# Data Integration

- Data integration:
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id  $\equiv$  B.cust-#
  - Integrate metadata from different sources
- **Entity identification problem:**
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units



# Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data*: One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by *correlation analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

# *Data Transformation*

- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- Attribute/feature construction
  - New attributes constructed from the given ones

# Data Transformation: Normalization

- Min-max normalization: to  $[\text{new\_min}_A, \text{new\_max}_A]$

$$v' = \text{new\_min}_A + \frac{v - \text{min}_A}{\text{max}_A - \text{min}_A} (\text{new\_max}_A - \text{new\_min}_A)$$

- Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to  $v' = 0.0 + \frac{73,600 - 12,000}{98,000 - 12,000} (1.0 - 0.0) = 0.716$

- Z-score normalization ( $\mu$ : mean,  $\sigma$ : standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

- Ex. Let  $\mu = 54,000$ ,  $\sigma = 16,000$ . Then  $v' = \frac{73,600 - 54,000}{16,000} = 1.225$

- Normalization by decimal scaling

$$v' = \frac{v}{10^j} \quad \text{where } j \text{ is the smallest integer such that } \max(|v'|) < 1$$

# Data Reduction Strategies

- Why data reduction?
  - A database/data warehouse may store terabytes of data
  - Complex data analysis/mining may take a very long time to run on the complete data set
- Data reduction
  - Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results
- Data reduction strategies
  - Data cube aggregation:
  - Dimensionality reduction — e.g., remove unimportant attributes
  - Data Compression
  - Numerosity reduction — e.g., fit data into models
  - Discretization and concept hierarchy generation

# Data Cube Aggregation

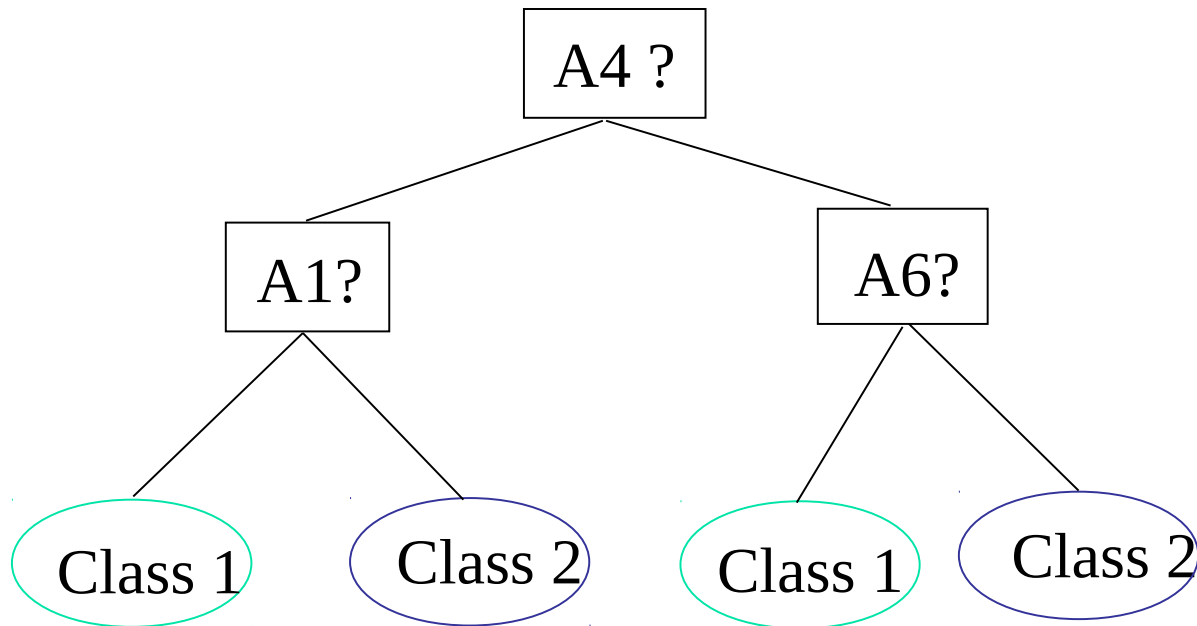
- The lowest level of a data cube (base cuboid)
  - The aggregated data for an **individual entity of interest**
  - E.g., a customer in a phone calling data warehouse
- Multiple levels of aggregation in data cubes
  - Further reduce the size of data to deal with
- Reference appropriate levels
  - Use the smallest representation which is enough to solve the task
- Queries regarding aggregated information should be answered using data cube, when possible

# Attribute Subset Selection

- Feature selection (i.e., attribute subset selection):
  - Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
  - reduce # of patterns in the patterns, easier to understand
- Heuristic methods (due to exponential # of choices):
  - Step-wise forward selection
  - Step-wise backward elimination
  - Combining forward selection and backward elimination
  - Decision-tree induction

# Example of Decision Tree Induction

Initial attribute set:  
{A1, A2, A3, A4, A5, A6}



Reduced attribute set: {A1, A4, A6}

# Heuristic Feature Selection Methods

- There are  $2^d$  possible sub-features of  $d$  features
- Several heuristic feature selection methods:
  - Best single features under the feature independence assumption: choose by significance tests
  - Best step-wise feature selection:
    - The best single-feature is picked first
    - Then next best feature condition to the first, ...
  - Step-wise feature elimination:
    - Repeatedly eliminate the worst feature
  - Best combined feature selection and elimination
  - Optimal branch and bound:
    - Use feature elimination and backtracking



*Merci de votre attention*

