

Data Mining Introduction



Organization

▣ Lectures

- ▣ Mondays from 14:00 to 16:00 and Tuesdays from 10:45 to 12:45
- ▣ Lecturer: Mouna Kacimi
- ▣ Office hours: appointment by email

▣ Labs

- ▣ Thursdays from 14:00 to 16:00
- ▣ Teaching Assistant: Mouna Kacimi

▣ Course Webpage: <http://www.inf.unibz.it/~mkacimi/teaching.shtml>

▣ Textbooks

- ▣ Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", Second Edition, 2006
- ▣ Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, "Introduction to Data Mining", Pearson Addison Wesley, 2008, ISBN: 0-32-134136-7

Project

- During Lab hours
- The project will be divided into small tasks, a new task every week
- The project can be done individually
- Groups of no more than 2 students are allowed
- You need to know how to program. If you do not know, team up with someone who knows
- You have the option to do a free project on your own: **your proposal needs to be approved by the teacher**

Exam Procedure

- Requirement: obtain 18 credit points in each of the following:
 - Project
 - Exam

$$\text{Final Grade} = 0.5 \times \text{Project Grade} + 0.5 \times \text{Exam Grade}$$

- Exams
 - Midterm Exam (optional) : 15 points
 - Final Exam
 - Full: 30 points
 - Partial: 15 points

$$\text{Exam Grade} = \left\{ \begin{array}{ll} \text{Midterm Grade} + \text{Partial Exam Grade} & \text{if student took the midterm exam} \\ \text{Full Exam Grade} & \text{if student did not take the midterm exam} \\ & \text{or decided not to consider the midterm exam} \end{array} \right\}$$

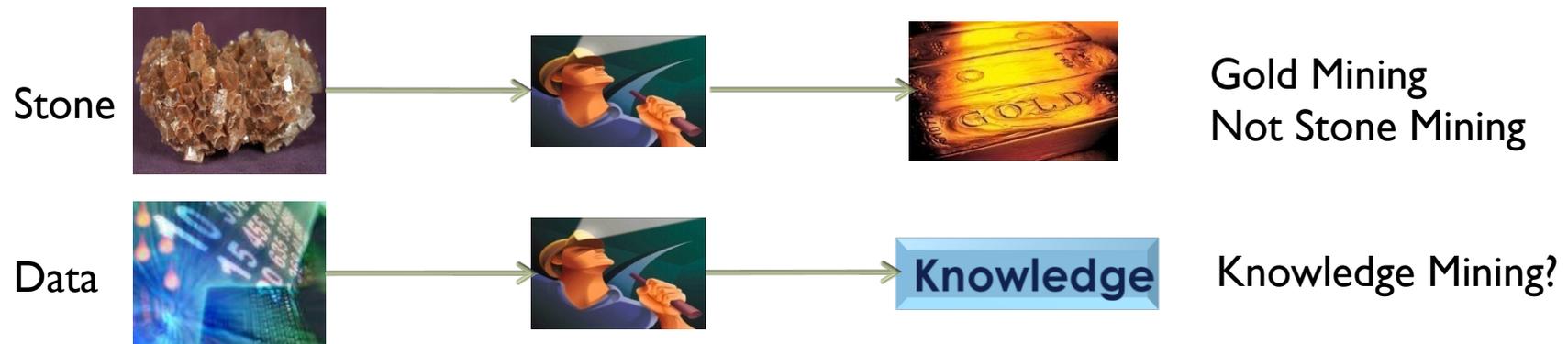
Exam Procedure

- Students must have a successful project to be able to take the final exam
- A successful project remains valid for three exam sessions
- If a project is unsuccessful until the day of the exam, its validity expires
- Students can do a new project until the next exam session. In this case, the teaching assistant does not guarantee support for supervising the students.

Road Map

1. Definitions & Motivations
2. Data to be mined
3. Knowledge to be discovered
4. Major Issues in Data Mining

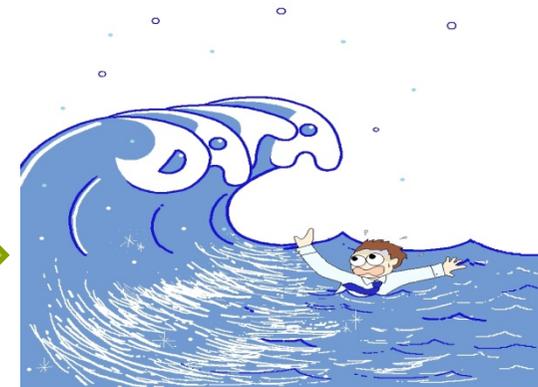
Data Mining: what is it?



- 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
- Alternative names
 - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Why Data Mining?

- Explosive Growth of Data: from terabytes to petabytes
- Data Collections and Data Availability
 - Crawlers, database systems, Web, etc.



- Sources

- Business: Web, e-commerce, transactions, etc.
- Science: Remote sensing, bioinformatics, etc.
- Society and everyone: news, YouTube, etc.

- **Problem:** We are drowning in data, but starving for knowledge!

- **Solution:** Use Data Mining tools for Automated Analysis of massive data sets

What Data Mining is Used For?

Financial Data Analysis

- Banks and Institutions offer a wide variety of banking services
 - Checking and saving accounts for business or individual customers
 - Credit business, mortgage, and automobile loans
 - Investment services (mutual funds)
 - Insurance services and stock investment services
- Financial data is relatively complete, reliable, and of high quality
- What to do with this data?



What Data Mining is Used For?

Financial Data Analysis

- Loan Payment Prediction and customer credit policy analysis
 - Attribute selection and attribute relevance ranking may help identifying important factors and eliminate irrelevant ones
 - Example of factors related to the risk of loan payment
 - Term of the loan
 - Debt ratio
 - Payment to income ratio
 - Customer level income
 - Education level
 - Residence region
- The bank can adjust its decisions according to the subset of factors selected

What Data Mining is Used For?

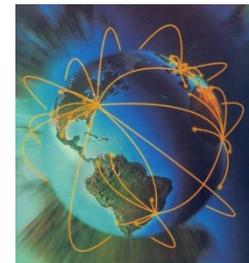
Retail Industry

- Collect huge amount of data on sales, customer shopping history, goods transportation, consumption and service, etc.
- Many stores have web sites where you can buy online. Some of them exist only online (e.g., Amazon)
- Data mining helps to
 - Identify customer buying behaviors
 - Discover customers shopping patterns and trends
 - Improve the quality of customer service
 - Achieve better customer satisfaction
 - Design more effective good transportation
 - Reduce the cost of business



What Data Mining is Used For?

- Many different ways of communicating
 - Fax, cellular phone, Internet messenger, images, e-mail, computer and Web data transmission, etc.
- Great demand of data mining to help
 - Understanding the business involved
 - Identifying telecommunication patterns
 - Catching fraudulent activities
 - Making better use of resources
 - Improve the quality of service

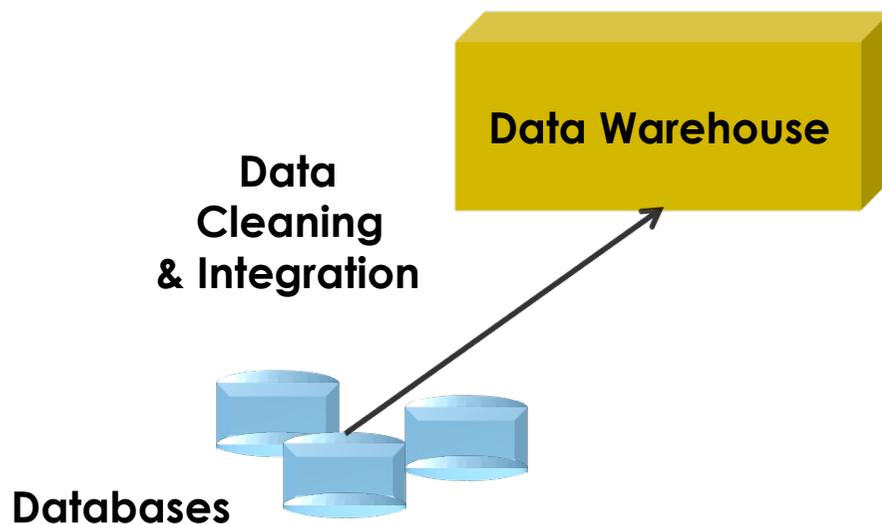


Example of a Data Mining Problem

- You want to do advertisement of sport activities for a set of new users on Facebook
- These people do not have explicit information about what they like and no history of past activities
- What you know is:
 - Their age, gender, and location
 - Their friends (not necessarily new users)
 - Messages they exchange with friends

Knowledge Discovery (KDD) Process

- Data Mining as a step in the knowledge discovery process

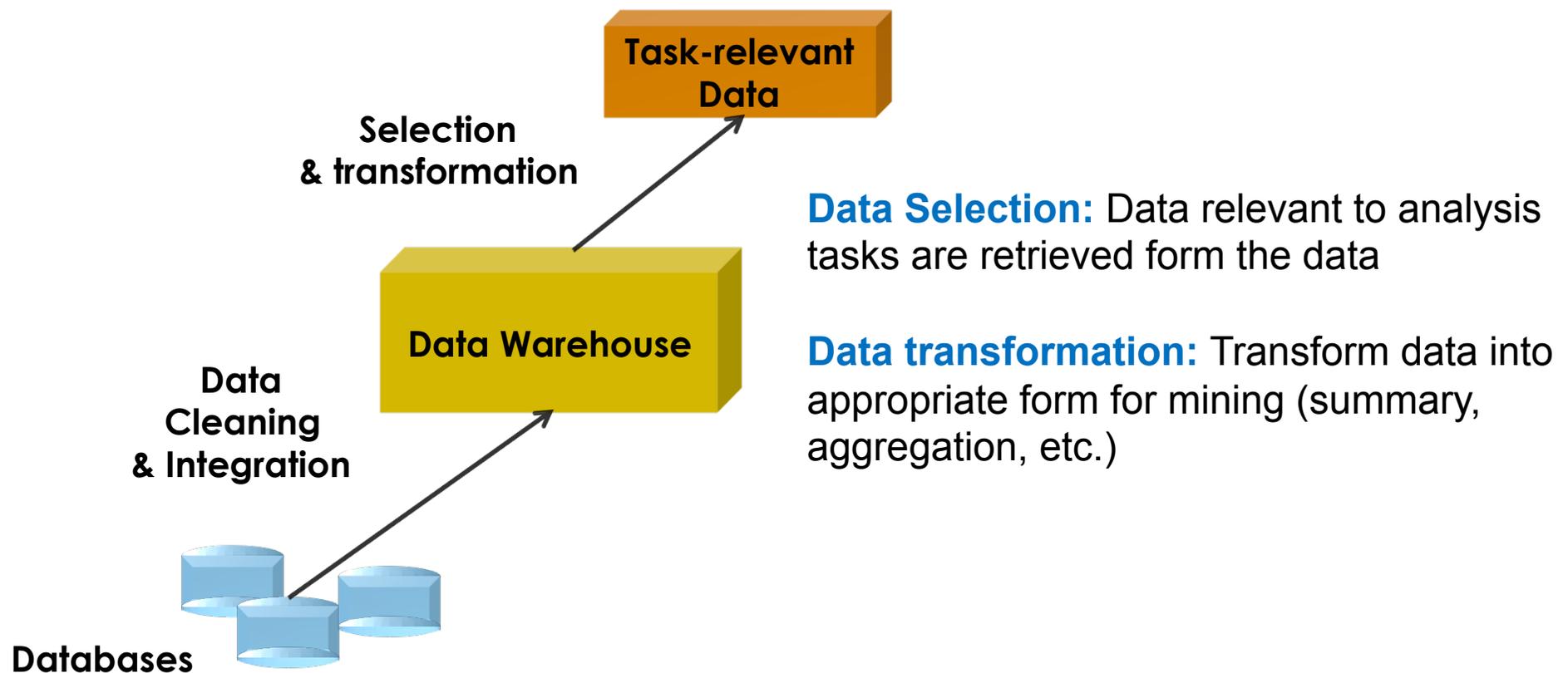


Data Cleaning: Remove noise and inconsistent data

Data Integration: Combine multiple data sources

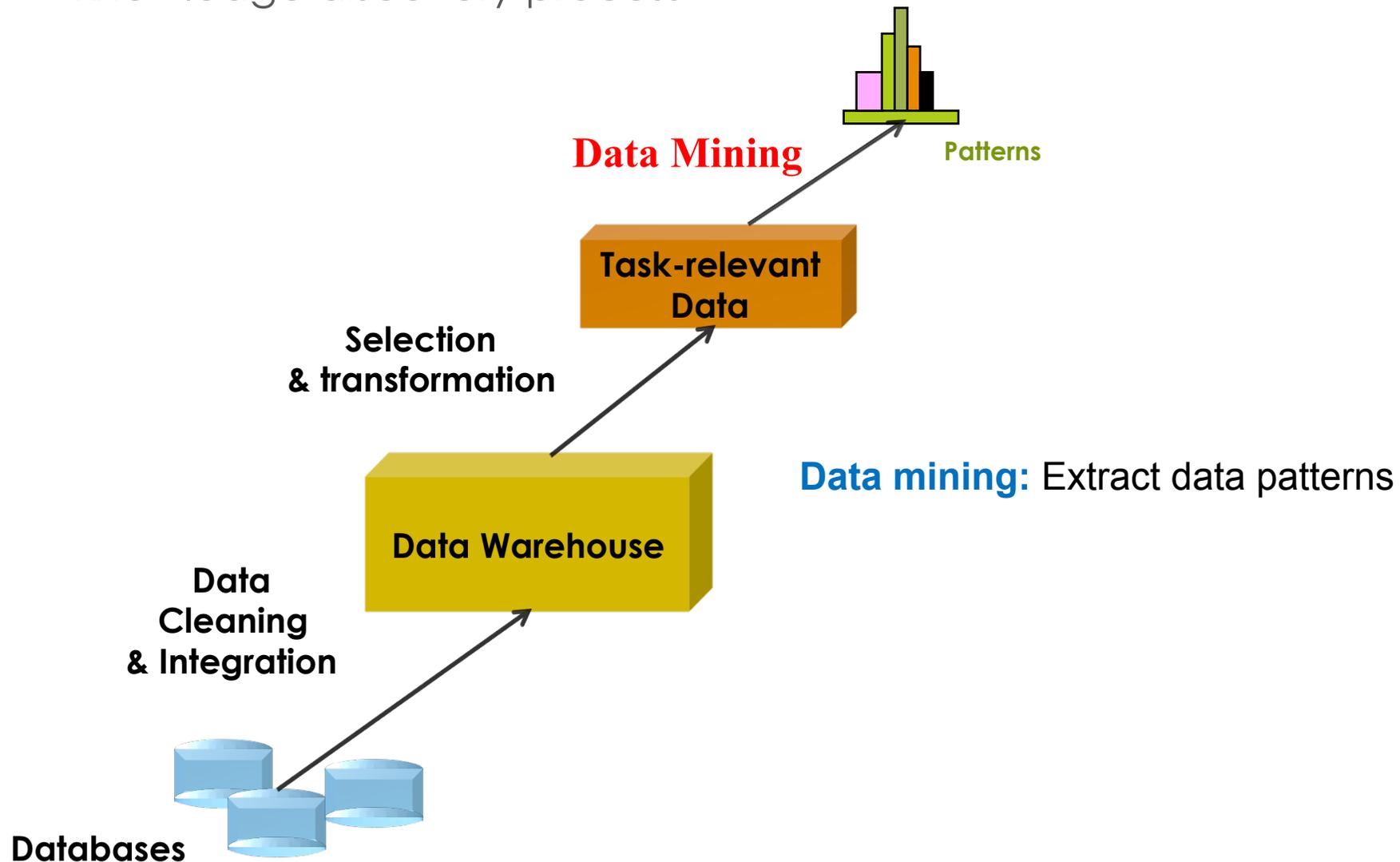
Knowledge Discovery (KDD) Process

- Data Mining as a step in the knowledge discovery process



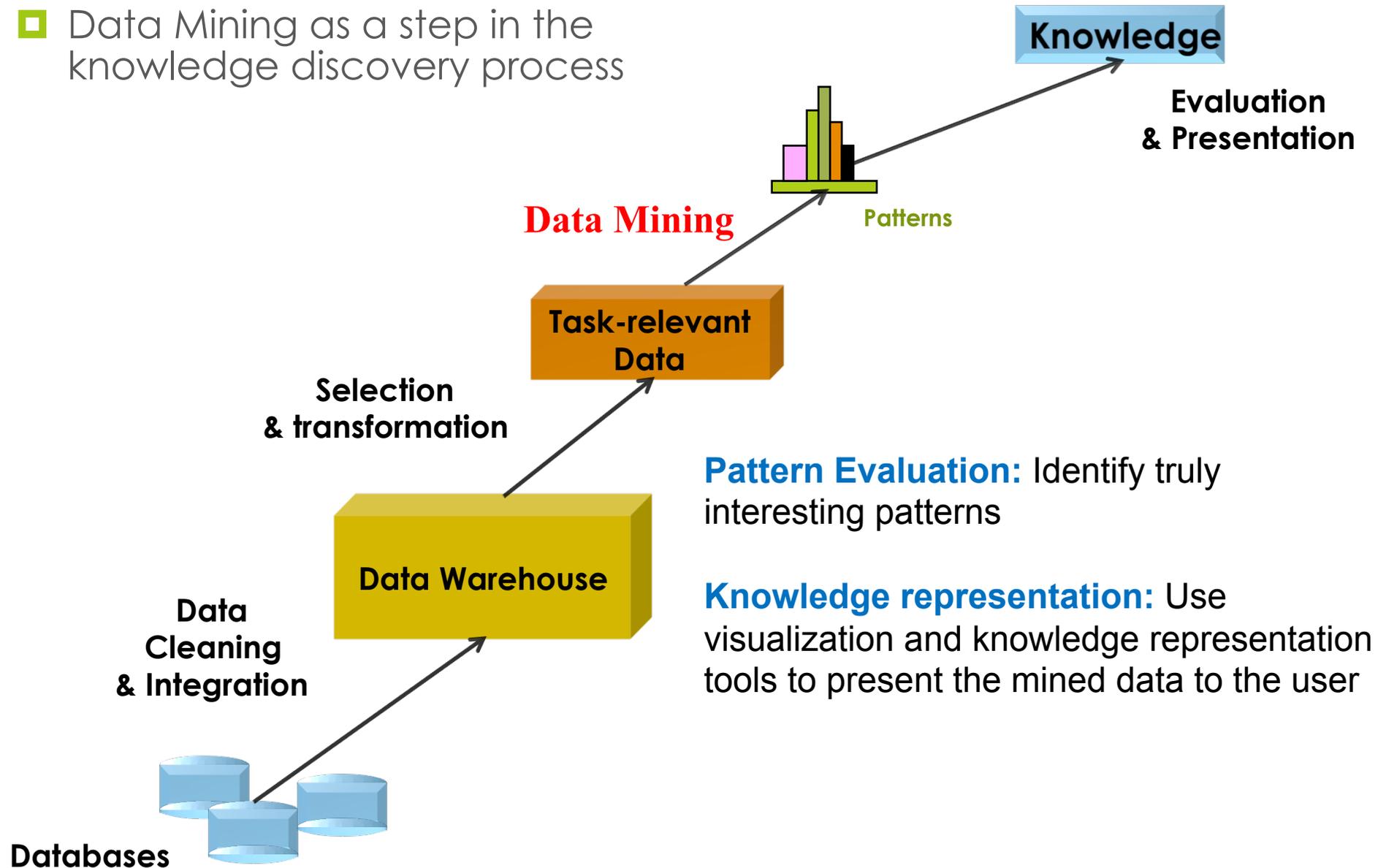
Knowledge Discovery (KDD) Process

- Data Mining as a step in the knowledge discovery process



Knowledge Discovery (KDD) Process

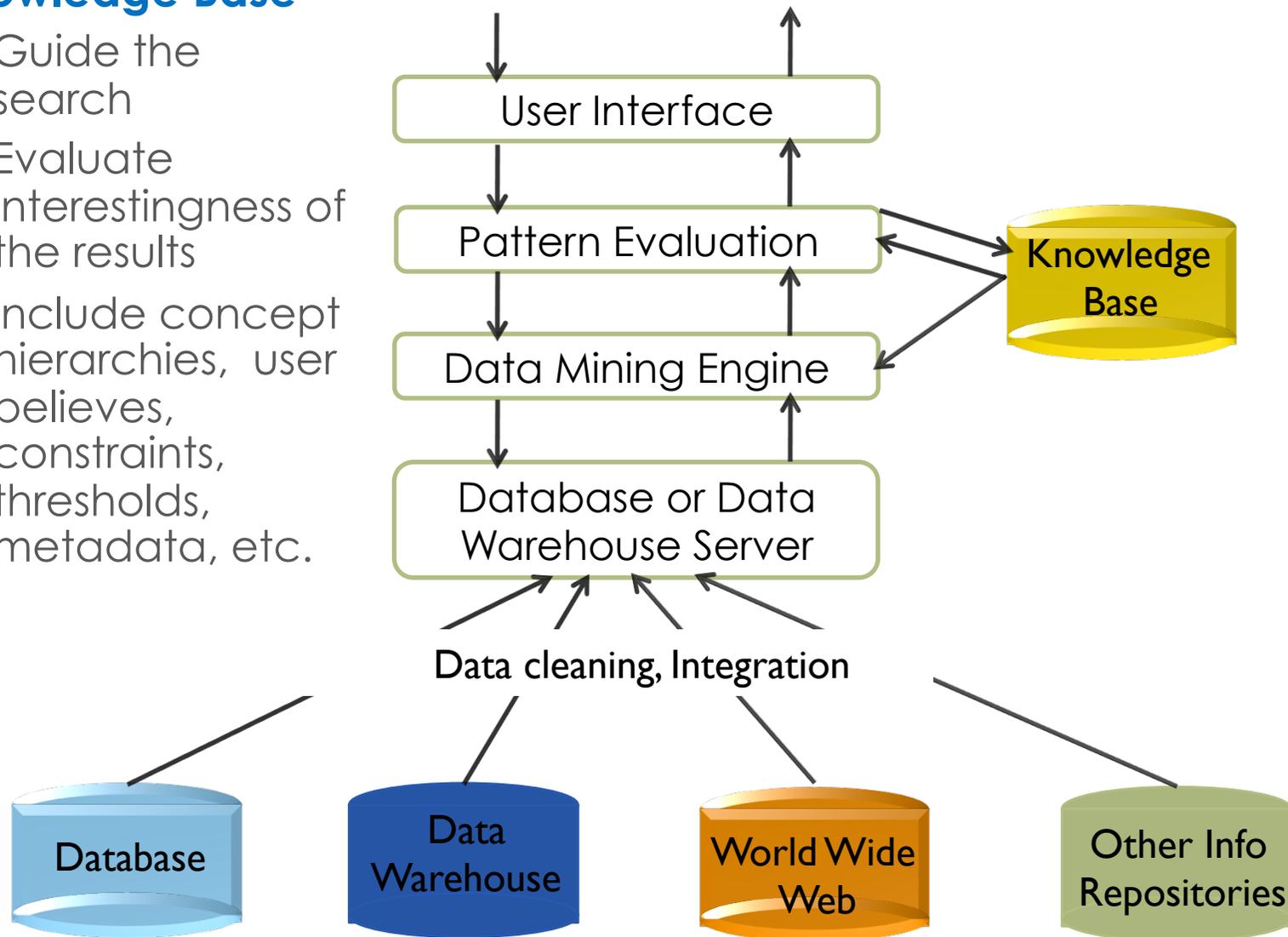
- Data Mining as a step in the knowledge discovery process



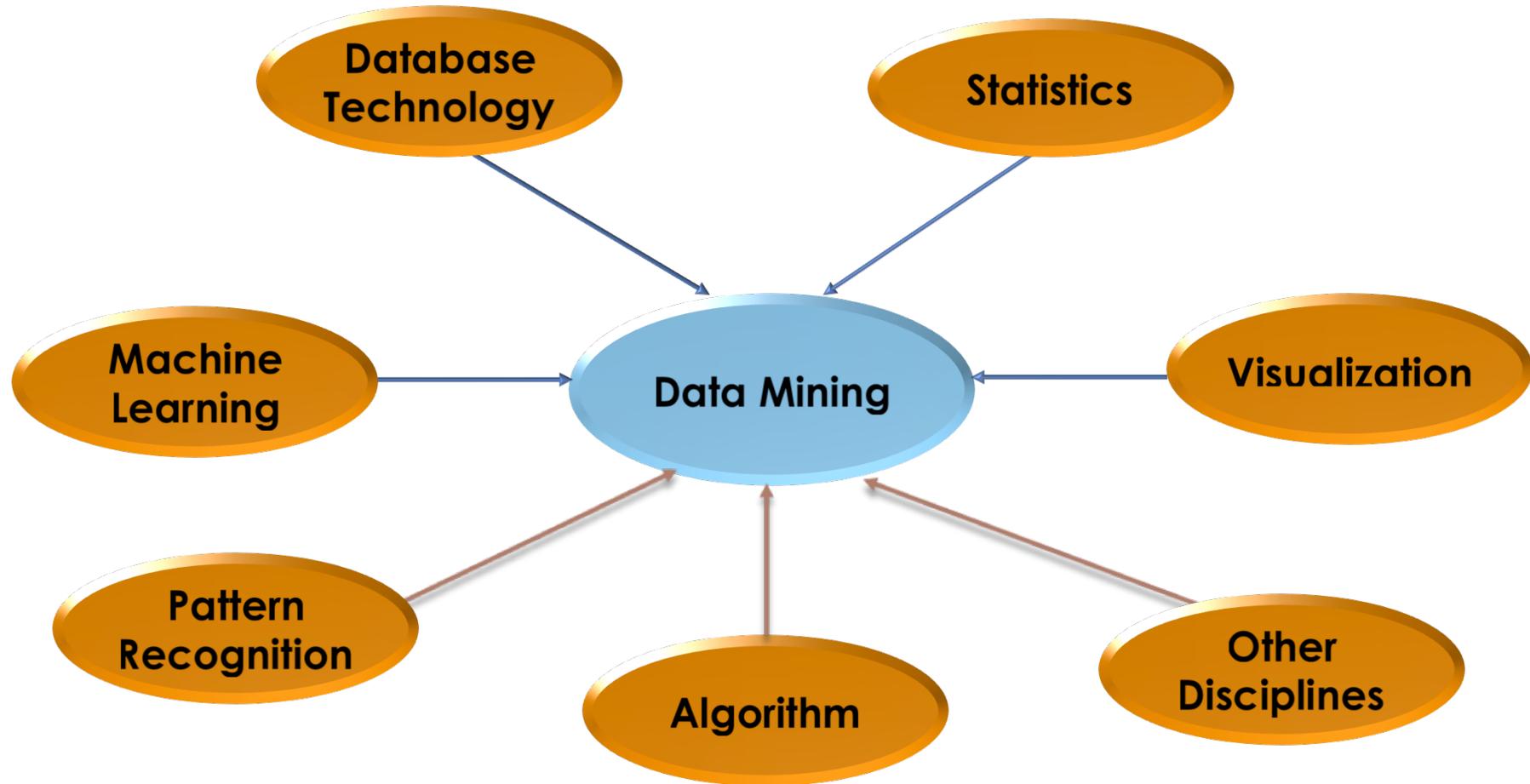
Typical Architecture of a DM System

Knowledge Base

- Guide the search
- Evaluate interestingness of the results
- Include concept hierarchies, user beliefs, constraints, thresholds, metadata, etc.



Confluence of Multiple Disciplines



Why Confluence of Multiple Disciplines?

August 2017: <https://zephoria.com/top-15-valuable-facebook-statistics/>

- ✓ 2.01 billion monthly active Facebook users
- ✓ 1.15 billion mobile daily active users
- ✓ Five new profiles are created every second
- ✓ 83 million fake profiles
- ✓ Photo uploads total 300 million per day
- ✓ Every 60 seconds on Facebook: 510,000 comments are posted, 293,000 statuses are updated, and 136,000 photos are uploaded
- ✓ 4.75 billion pieces of content shared daily

Different Views of Data Mining

▣ Data View

- ▣ Kinds of data to be mined

▣ Knowledge view

- ▣ Kinds of knowledge to be discovered

▣ Method view

- ▣ Kinds of techniques utilized

▣ Application view (seen before)

Road Map

1. Definitions & Motivations
2. Data to be mined
3. Knowledge to be discovered
4. Major Issues in Data Mining

Data to be Mined

- In principle, data mining should be applicable to any data repository

- This lecture includes examples about:
 - Relational databases
 - Data warehouses
 - Transactional databases
 - Advanced database systems

Relational Databases

Database System

- Collection of interrelated data, known as **database**
- A set of software programs that manage and access the data

Relational Databases (RD)

- A collection of tables. Each one has a unique name
- A table contains a set of attributes (columns) & tuples (rows).
- Each object in a relational table has a unique key and is described by a set of attribute values.
- Data are accessed using database queries (SQL): projection, join, etc.

Customers

cust_id	Name	age	income
152	Anna	27	24000 €
...

Purchases

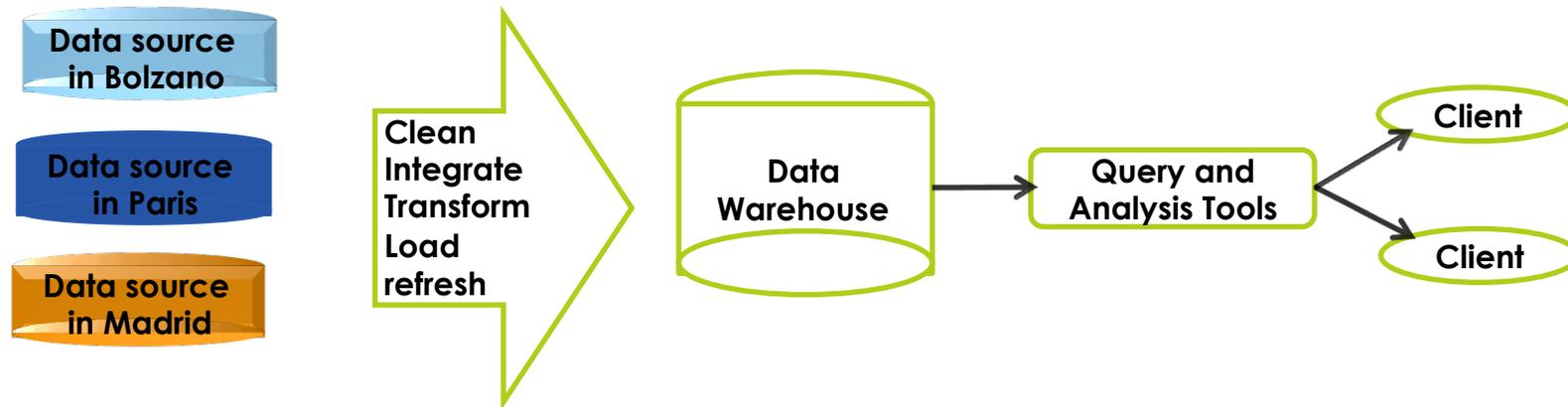
trans_id	cust_id	method	Amount
T156	152	Visa	1357 €
...

Data Mining applied to RD

- Search for trends or data patterns
- Example:** predict the credit risk of costumers based on their income, age and expenses.

Data Warehouses

- A data warehouse (DW) is a repository of information collected from multiple sources, stored under a unified schema.



- Data organized around major subjects (using summarization)
- Multidimensional database structure (e.g., data Cube)
 - Dimension = one attribute or a set of attributes
 - Cell = stores the value of some aggregated measures.
- **Data Mining applied to DW**
 - Data warehouse tools help data analysis
 - Data Mining tools are required to allow more in-depth and automated analysis

Transactional Databases

- A transactional database (TD) consists of a file where each record represents a transaction.
- A transaction includes a unique transaction identifier (trans_id) and a list of the items making the transaction.
- A transaction database may include other tables containing other information regarding the sale (customer_id, location, etc.)
- Basic analysis (examples)
 - Show me all the items purchased by David Winston?
 - How many transactions include item number 5?

■ Data Mining on TD

- Perform a deeper analysis
- Example: Which items sold well together?
- Basically, data mining systems can identify frequent sets in transactional databases and perform *market basket data analysis*.

trans_id	List of items_IDs
T100	I1,I3,I8,I16
T200	I2,I8
...	...

Advanced Database Systems (1)

- Advanced database systems provide tools for handling **complex** data
 - Spatial data (e.g., maps)
 - Engineering design data (e.g., buildings, system components)
 - Hypertext and multimedia data (text, image, audio, and video)
 - Time-related data (e.g., historical records)
 - Stream data (e.g., video surveillance and sensor data)
 - World Wide Web, a huge, widely distributed information repository made available by Internet
- Require **efficient** data structures and **scalable** methods to handle
 - Complex object structures and variable length records
 - Semi structured or unstructured data
 - Multimedia and spatiotemporal data
 - Database schema with complex and dynamic structures

Advanced Database Systems (2)

Example: World Wide Web

- Provide rich, worldwide, online and distributed information services.
- Data objects are linked together
- Problems
 - Data can be highly unstructured
 - Understand the semantic of web pages

Data Mining on WWW

- Web usage Mining (user access pattern)
 - Improve efficiency and make better marketing decisions
- Authoritative Web page Analysis
 - Ranking web pages based on their importance
- Automated Web page clustering and classification
 - Group and arrange web pages based on their content
- Web community analysis
 - Identify hidden web social networks and observe their evolution



Road Map

1. Definitions & Motivations
2. Data to be mined
3. Knowledge to be discovered
4. Major Issues in Data Mining

Knowledge to be Discovered

- Data mining functionalities are used to specify the kind of patterns to be found in **data mining tasks**
- Data mining tasks can be classified into two categories
 - **Descriptive** : Characterize the general properties of the data
 - **Predictive** : Perform inference on the current data to make predictions
- **What to extract?**
 - Users may not have an idea about what kinds of patterns in their data can be interesting
- **What to do?**
 - Have a data mining system that can mine multiple types of patterns to handle different user and application needs.
 - Discover patterns at various granularities (levels of abstraction)
 - Allow users to guide the search for interesting patterns



Example of different granularities

Characterization and Discrimination (1)

- Data can be associated with classes or concepts

Example of data from a store



- **Class/Concept descriptions:** describe individual classes and concepts in summarized, concise, and precise way.
 - Data characterization
 - Summarize the data of the class under study (**target class**)
 - Data Discrimination
 - Compare the target class with a set of comparative classes (**contrasting classes**)
 - Data characterization & Discrimination
 - Perform both analysis

Characterization and Discrimination (2)

□ Data Characterization

- Output: charts, curves, multidimensional data cubes, etc.
- Example

Summarize the characteristics of costumers who spend more than 1000€



Costumers profile

- 40-50 years old
- Employed
- excellent credit ratings

□ Data Discrimination

- Output: similar to characterization + comparative measures
- Example

Compare costumers who shop for computer products regularly(more than 2 times a month) with those who rarely shop for such products(less than three times a year)



Comparative profile

Frequent costumers

80%
•Are between 20 and 40
•Have university education

Rare costumers

60%
•Are senior or youths
•Have no university degree

Frequent Patterns, Associations, Correlations

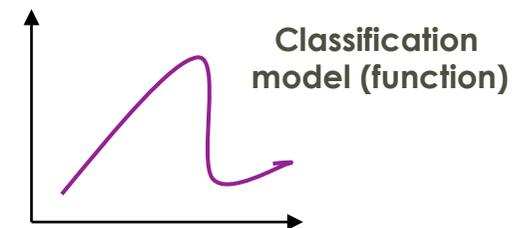
- **Frequent patterns** are patterns occurring frequently in the data (e.g., item-sets, sub-sequences, and substructures)
 - **Frequent item-sets**: items that frequently appear together
 - Example in a transactional data set: **bread** and **milk**
 - **Frequent Sequential pattern**: a frequently occurring subsequence
 - Example in a transactional data set: buy **first PC**, **second digital camera**, **third memory card**
- **Association Analysis**
 - Derive some association rules
 - $buys(X, \text{"computer"}) \Rightarrow buys(X, \text{"software"})$ [support =1%, confidence=50%]
 - $age(X, \text{"20...29"}) \wedge income(X, \text{"20K...29K"}) \Rightarrow buys(X, \text{"CD player"})$ [support =2%, confidence=60%]
- **Correlation Analysis**
 - Uncover interesting statistical correlations between associated attribute-value pairs

Classification & Prediction

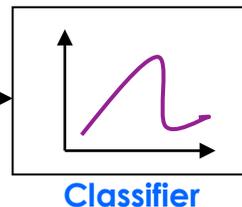
- Construct models (functions) based on some training examples
- Describe and distinguish classes or concepts for future prediction
- Predict some unknown class labels

	Age	Income	Class label
Training examples	27	28K	Budget-Spenders
	35	36K	Big-Spenders
	65	45K	Budget-Spenders

Supervised Learning



Unlabeled data Age Income
 29 25K



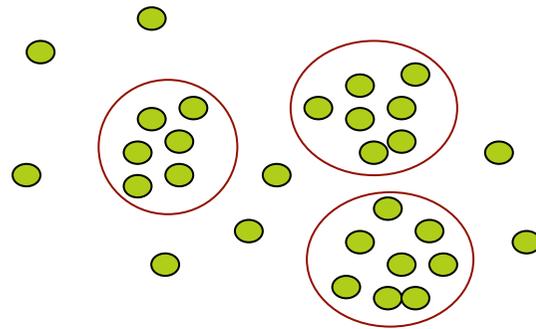
Class label [Budget Spender]

Numeric value [Budget Spender (0.8)]

- **Typical Models:** Decision trees, Bayesian classifiers, Regression, etc.
- **Typical Application:** Credit card fraud detection, classifying web pages, stars, diseases, etc

Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity



- **Typical methods:** Hierarchical, density-based, Grid-based, Model-Based, constraint-based , etc.
- **Typical Applications:** WWW, social networks, Marketing, Biology, Library, etc.

Outlier Analysis

- **Outlier:** A data object that does not comply with the general behavior of the data learning (i.e., Class label is unknown)
- Noise or exception? — One person's garbage could be another person's treasure



- **Typical methods:** Product of clustering or regression analysis, etc
- **Typical Applications:** Useful in fraud detection:
 - How to uncover fraudulent usage of credit card?
 - Detect purchases of extremely large amounts for a given account number in comparison to regular charges incurred by the same account
 - Outliers may also be detected with respect to the location and type of purchase, or the frequency.

Road Map

1. Definitions & Motivations
2. Data to be mined
3. Knowledge to be discovered
4. Major Issues in Data Mining

Major Challenges of Data Mining

- Efficiency and scalability of data mining algorithms
- Parallel, distributed, stream, and incremental mining methods
- Handling high-dimensionality, noise, uncertainty, and incompleteness of data
- Incorporation of constraints, expert knowledge, and background knowledge in data mining
- Pattern evaluation and knowledge integration
- Mining diverse and heterogeneous kinds of data
- Application-oriented and domain-specific data mining
- Protection of security, integrity, and privacy in data mining

Summary

- Data Mining is a process of extracting knowledge from data
- Data to be mined can be of any type
 - Relational Databases, Advanced databases, etc.
- Knowledge to be discovered
 - Frequent patterns, correlations, associations, classification, prediction, clustering
- Data Mining is interdisciplinary
 - Large amount of complex data and sophisticated applications
- Challenges of data Mining
 - Efficiency, scalability, parallel and distributed mining, handling high dimensionality, handling noisy data, mining heterogeneous data, etc.