

Data Mining

cse392

Professor Anita Wasilewska
Computer Science Department
State University of New York
at Stony Brook

Data Mining

Course Plan

- **Part One:** Intuitive Introduction and DM, An Overview of DM concepts and techniques
- **Part Two:** Textbook chapters 1,2, 3 and 6-8
- **Part Three:** Students Presentations

- Course Textbook (on reserve in CS Library):
 - Jianwei Han, Micheline Kamber
DATA MINING Concepts and Techniques
Morgan Kaufmann, 2005

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

Data Mining

Main Objectives

- Identification of data as a source of useful information
- Use of discovered information for competitive advantages when working in business environment

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 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

Data Mining: extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

Data ware housing: new method of data organization and data management

What is Data Mining?

- There are many activities with the same name: hence a CONFUSSION
- **DM:** Huge volumes of data
- **DM:** Potential hidden knowledge
- **DM:** Process of discovery of hidden patterns in data

Data Mining Intuitive Definition

- **DM** is a process of extraction previously unknown knowledge from large volumes of data
- **DM** requires both new technologies and new methods

What Data Mining Does

- **DM** creates models (algorithms):
 - **classification** (chapter 5)
 - **association** (chapter 6)
 - **clustering** (chapter 8)
 - **statistical prediction** (chapter 7)

What Data Mining Does

- **DM** presents the knowledge as a set of rules of the form:

IF.... THEN...

DM presents the knowledge as a set of descriptions

DM finds other relationships in data (clusters, statistical patterns)

DM detects deviations

Data Mining: Some Applications

- competitive target marketing, customer relation management, market basket analysis, cross selling, market segmentation
- Forecasting, customer retention, improved underwriting, quality control analysis

Data Mining: Other Applications

- **Other Applications**

Text mining (news group, email, documents)

Web analysis

Intelligent query answering

Scientific Applications

DM: Business Advantages

- **Data Mining** uses gathered data to
- **Predicts** tendencies and waves
- **Classify** new data
- **Finds** previously unknown patterns
- **Discovers** unknown relationships

DM: Technologies

- There are many commercially available tools
- There are many methods (models, algorithms) for the same task
- **TOOLS ALONE ARE NOT THE SOLUTION**
- The user must be able to understand the chosen algorithms in order to prepare the data and interpret the results;

Data Mining Principle

- Remember that one of the principal requirements of DM is:
“the results must be easily comprehensible to the user”
- 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

Business Applications

- Buying patterns
- Fraud detection
- Customer Campaigns
- Decision support
- Medical applications
- 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 (machine learning techniques) 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 fraudulent 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 (1)

- **Which are 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 (2)

- Goal is to determine customer patterns over time
 - For example a 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 (3)

- Customer profiling and targeting
 - data mining is used to tell what types of customers buy what products (clustering or classification) and this information is then used to target new customers, i.e. is used for identifying the best products for different customers

Corporate Analysis and Risk Management (1)

- 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 (2)

- **Competition:**
monitor competitors and market directions by grouping customers into classes and use a class-based pricing procedure.
It allows to 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
- **Remember:** Different Data Mining methods and algorithms 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

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

- Data Mining find the patterns and once the patterns are found Data Mining process is finished
- The use of the patterns is not Data Mining it is an application of DM
- Monitoring is not analysis
- Queries to the database are not DM

Short History of Evolution of Database Technology

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

Evolution of Database Technology c.d.

- 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

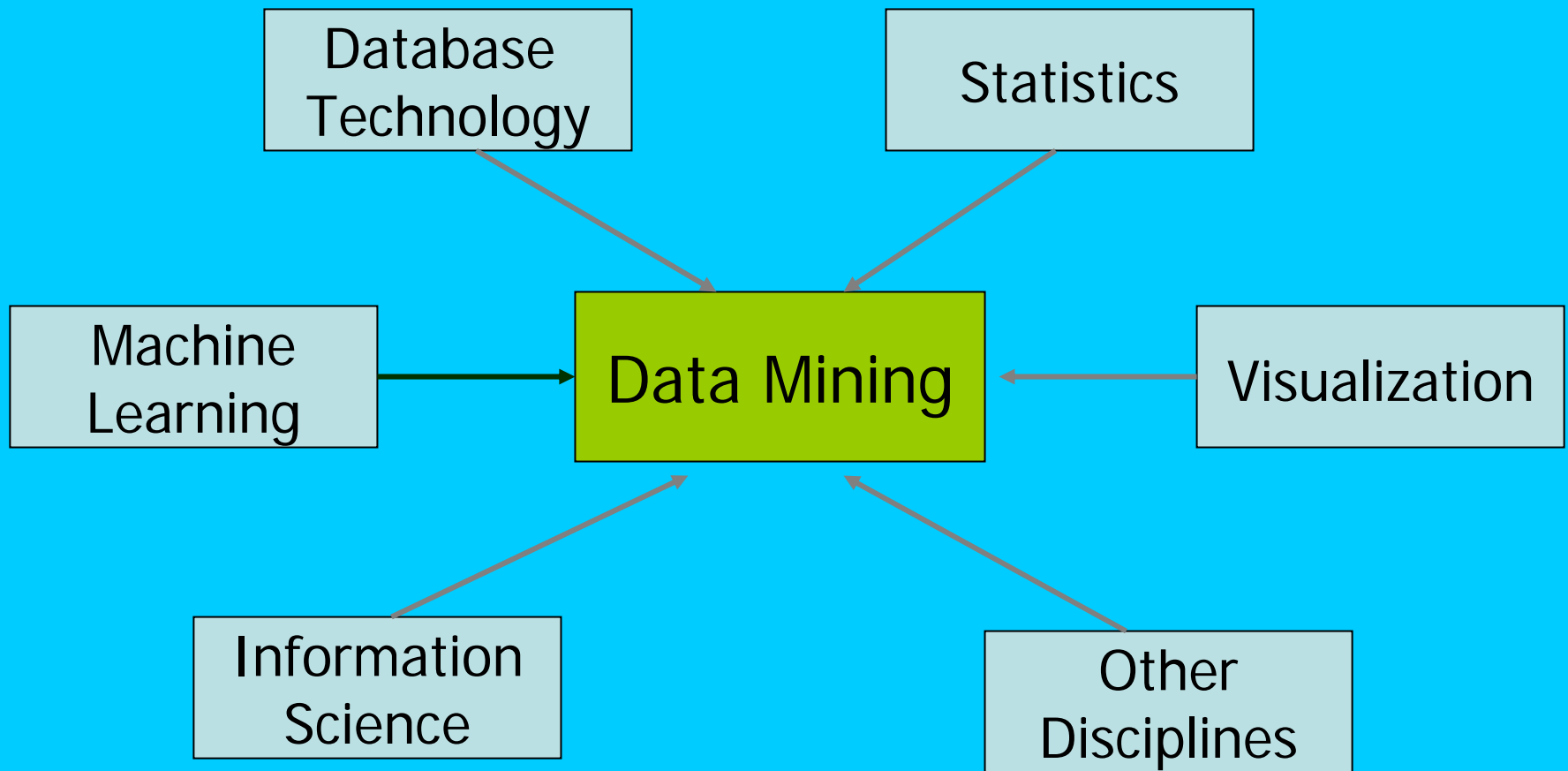
Short History of Data Mining

- **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.

Short History of Data Mining

- **1996 – present:**
- **KDD** evolves as a conjunction of different knowledge areas (data bases, machine learning, statistics, artificial intelligence) and the term **Data Mining** becomes popular

Data Mining: Confluence of Multiple Disciplines

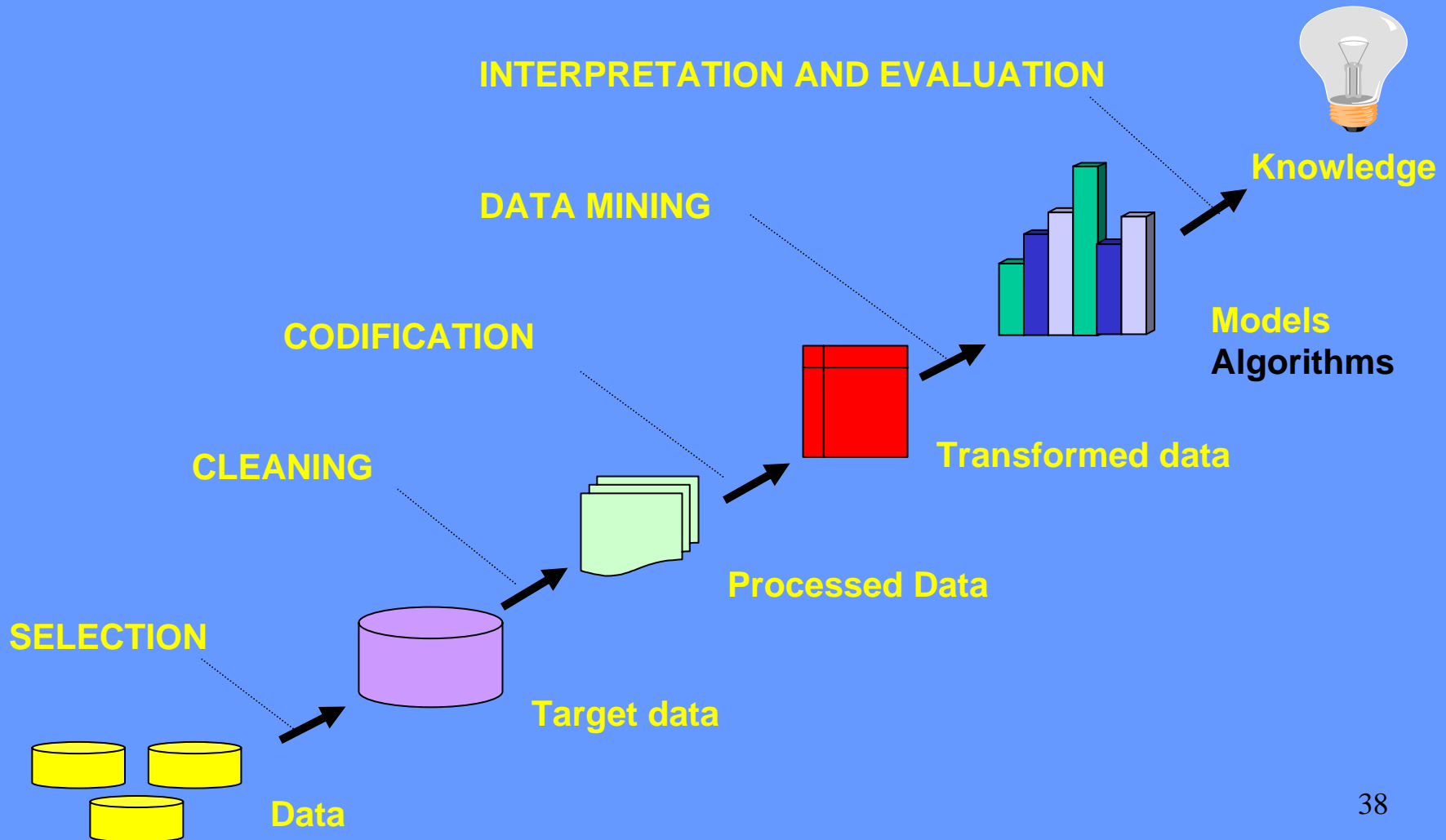


KDD process: Definition

[Piatetsky-Shapiro 97]

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

The KDD process



DM: Data Mining

- **DM** is a step of the **KDD process** in which special algorithms are applied to discover patterns in data
- **Remember:** It is necessary to apply first the preprocessing operation to clean and preprocess the data in order to obtain significant patterns

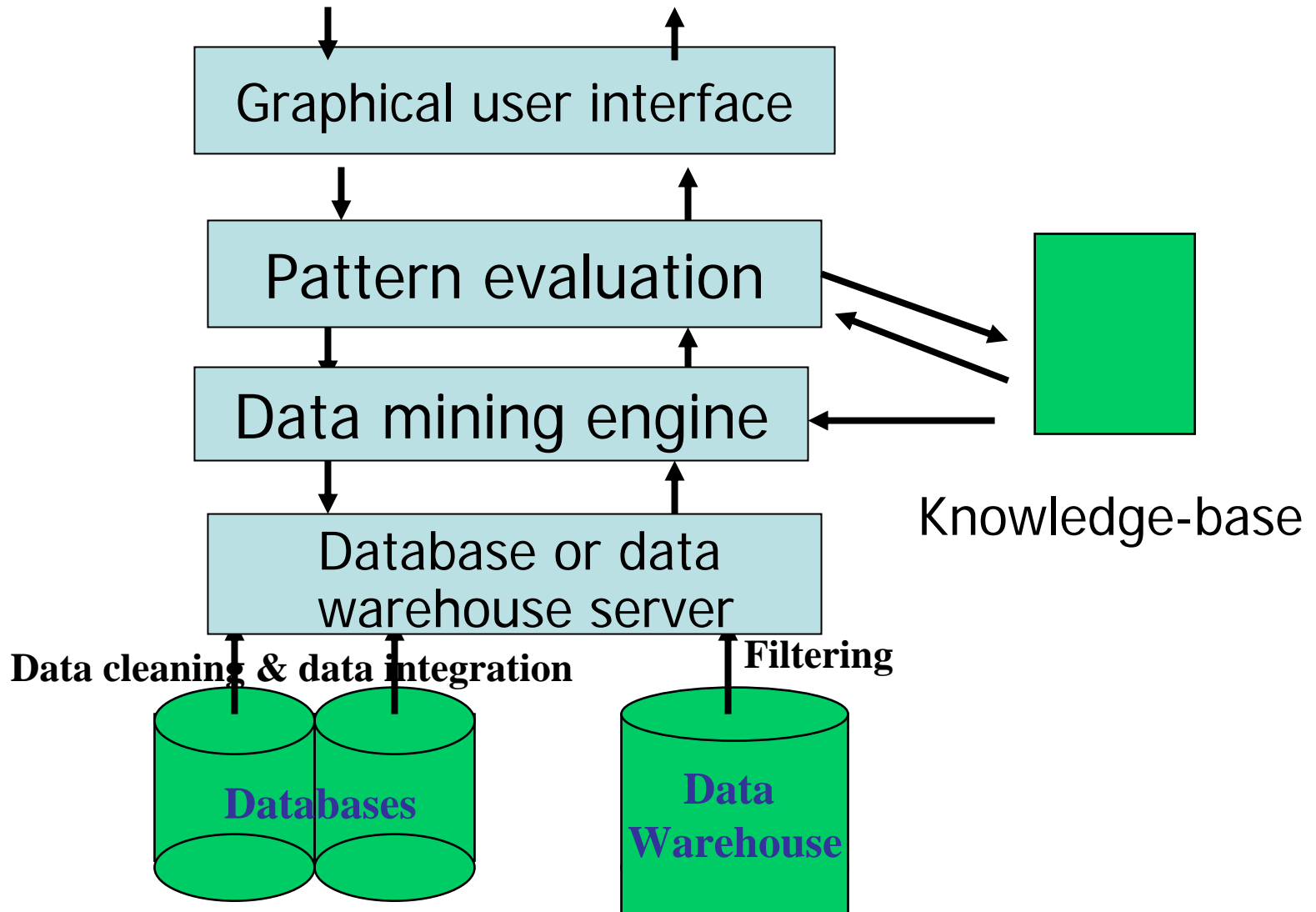
KDD vs DM

- **KDD** is a term used by Academia
- **DM** is a 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**

Steps of the KDD process

- **Preprocessing:** includes all the operations that have to be performed before a data mining algorithm is applied
- **Data Mining:** knowledge discovery algorithms are applied in order to obtain the patterns
- **Interpretation:** discovered patterns are presented in a proper format and the user, or implementer decides if it is necessary to re-iterate the algorithms

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

www