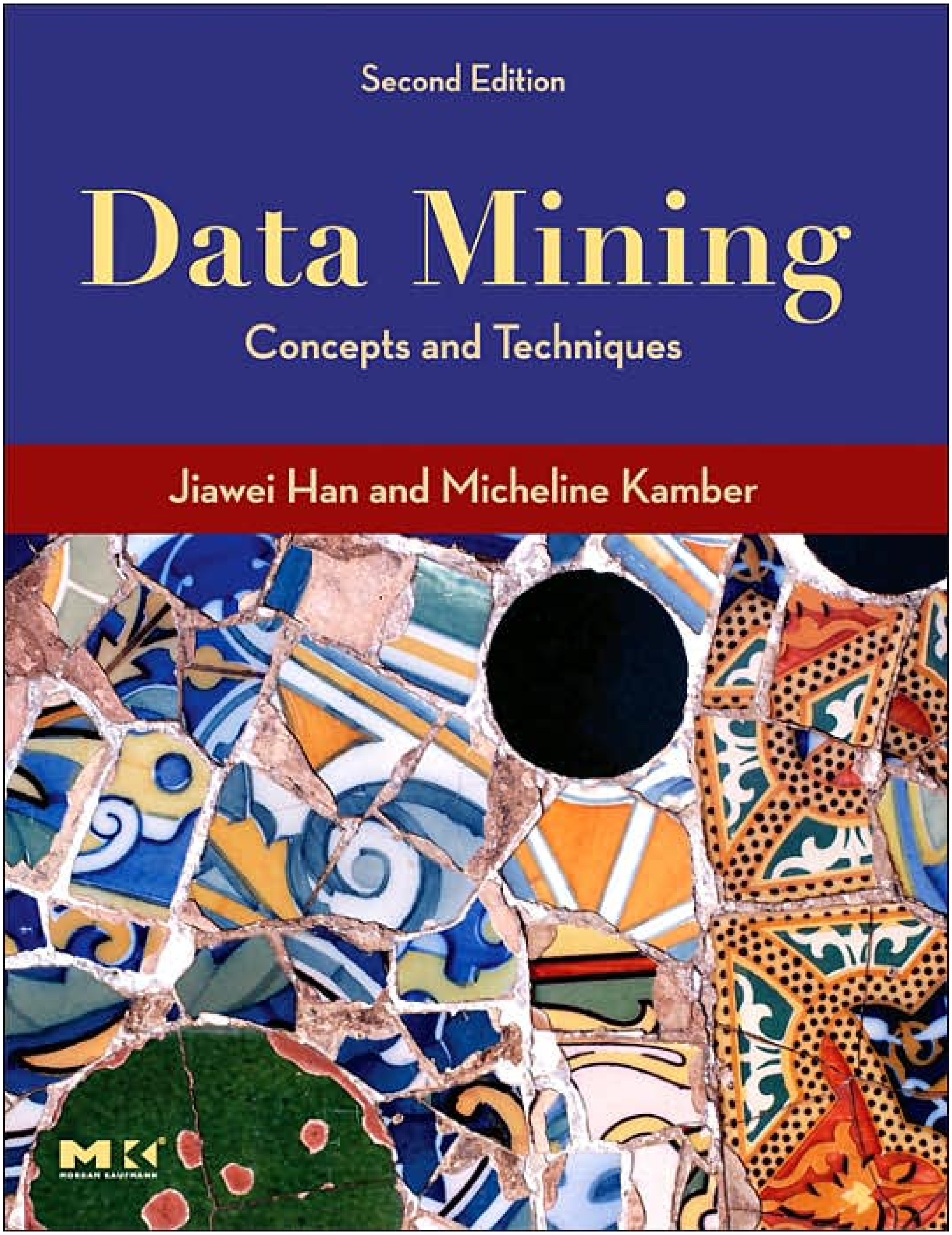
Introduction


Second Edition
Data Mining
Concepts and Techniques
Jiawei Han and Micheline Kamber
MK
MORGAN KAUFMANN

1. Objectives

- Large amount of data kept in data files, databases, and web servers:
    - Structured data
    - Unstructured data

- Users are expecting more information from these data

- Marketing managers are interested in customers' purchase behavior

- Simple structured/query language queries are not adequate to extract hidden information:

    - Traditional SQL statements only retrieve a subset of the database.

- Evolution of database technology:

  Hierarchical → Network→ Relational → Extended Relational → Semantic DB → (ORDBMS, OODBMS)

- Overall advancement of computing

2. What is Data Mining?

- Mining ‘Gold” from ‘Rocks”

- Simple Definition: Extract or “mine” knowledge from large amount of data.

- 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.

- Watch out: Is everything “data mining”?
    - (Deductive) query processing.
    - Expert systems or small ML/statistical programs

3. Knowledge Discovery Process

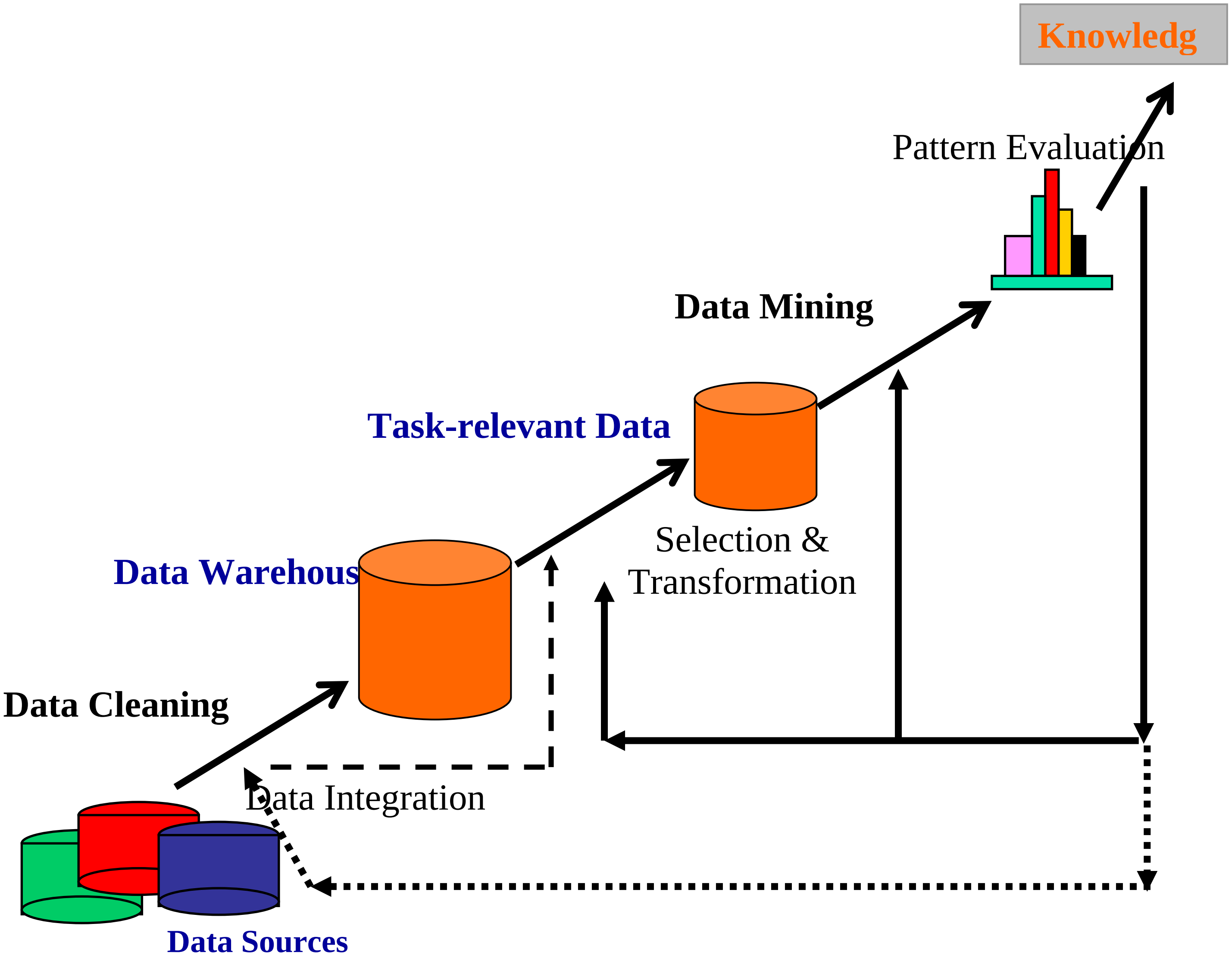


Figure 1.4 of the textbook (Modified)

- Learning the application domain
    - Relevant prior knowledge and goals of application

- Data Cleaning: Remove noise data and irrelevant data (stopwords in case of unstructured data)

- Data Integration: Combine multiple data sources

- Data Selection: Get data relevant to the task to be analyzed

- Data Reduction and Transformation: Prepare data in a form appropriate for mining:
    - Represent a text file as a vector
    - Find useful features
    - Reduce your space (dimensionality/variable).

- Data Mining: a process to extract data patterns, e.g., summarization, classification, regression, association, clustering.

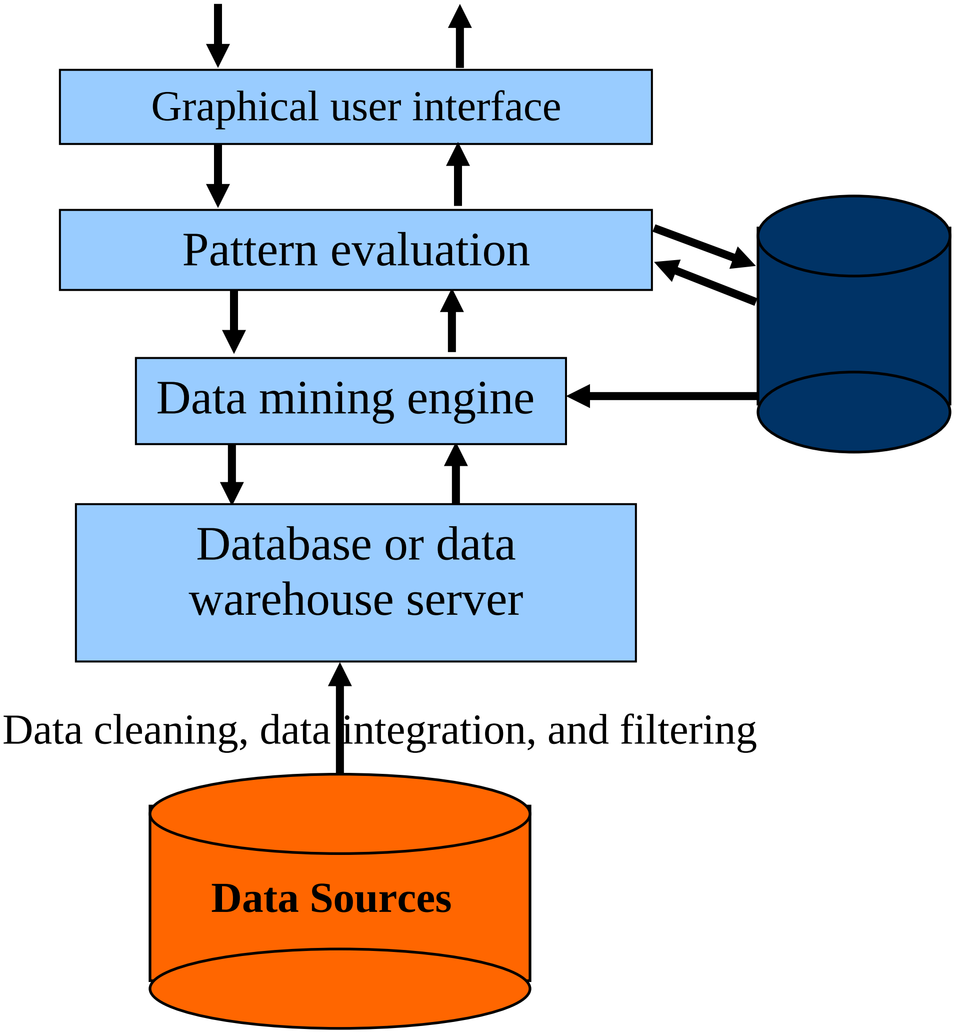
- Pattern Evaluation: Evaluate the output of the data mining process.

- Knowledge Representation: Techniques to visualize mined knowledge.

4. KD Process Example

- Web Log Mining
    - Selection:
        - Select log data (dates and location) to use
    - Preprocessing:
        - Remove identifying URLs
        - Remove error logs
    - Transformation:
        - Sessionize (sort and group)
    - Data Mining:
        - Construct data structure
        - Create frequent sequences
    - Interpretation/Evaluation:
        - Cache prediction
        - Personalization

5. Typical Data Mining Architecture


Graphical user interface
Pattern evaluation
Data mining engine
Database or data warehouse server
Data cleaning, data integration, and filtering
Data Sources

6. Database vs. Data Mining

Query: - Well defined SQL	**Query**: - Poorly defined No precise query language
Data: - Operational data	**Data**: - Not operational data
Output: - Precise Subset of database	**Output**: - Fuzzy - Not a subset of database

- Query Example:
  - Database:
    - ✓ Find all credit applicants with last name of Smith.
    - ✓ Identify customers who have purchase more than $10,000 in last month.
    - ✓ Find all customers who have purchased milk

  - Data Mining:
    - ✓ Find all credit applicants who are poor credit risks. (Classification)
    - ✓ Identify customers with similar buying habits. (Clustering)
    - ✓ Find all items that are frequently purchased with milk. (Association rules)

7. Data Mining: On What kind of Data?

- Database
- Data warehouse
- Transactional database
- Object-oriented database
- Object-relational database
- Spatial data
- Temporal data and Time-series data
- Multimedia database
- Text Collections
- WWW

8. Potential Applications

8.1. 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
  - Associations/co-relations between product sales, & prediction based on such association

- Customer profiling
  - What types of customers buy what products (clustering or classification)

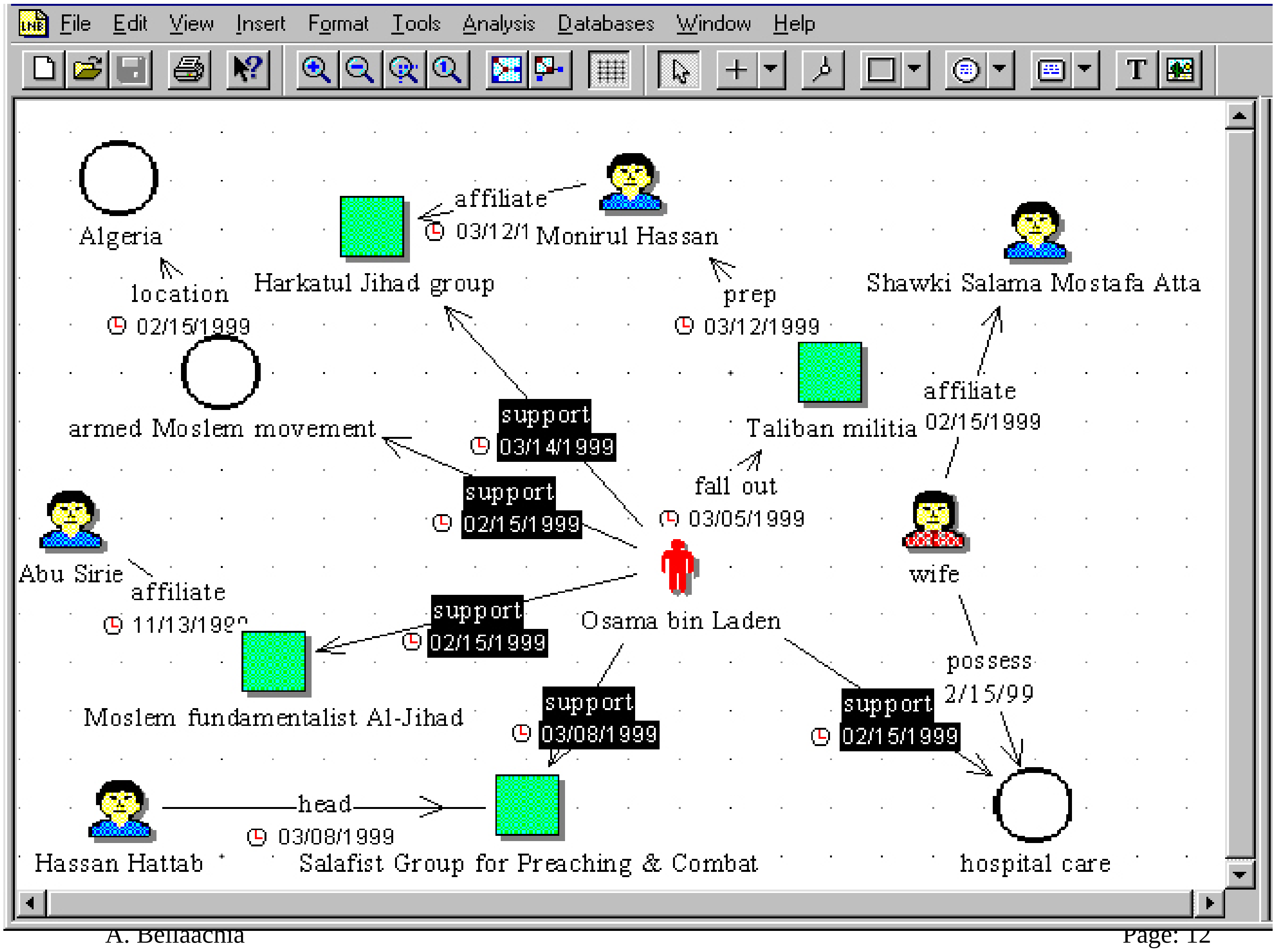
- Customer requirement analysis
  - Identifying the best products for different customers
  - Predict what factors will attract new customers

- Provision of summary information
  - Multidimensional summary reports
  - Statistical summary information (data central tendency and variation)

- Risk analysis and management

- Forecasting
- Customer retention
- Improved underwriting
- Competitive analysis

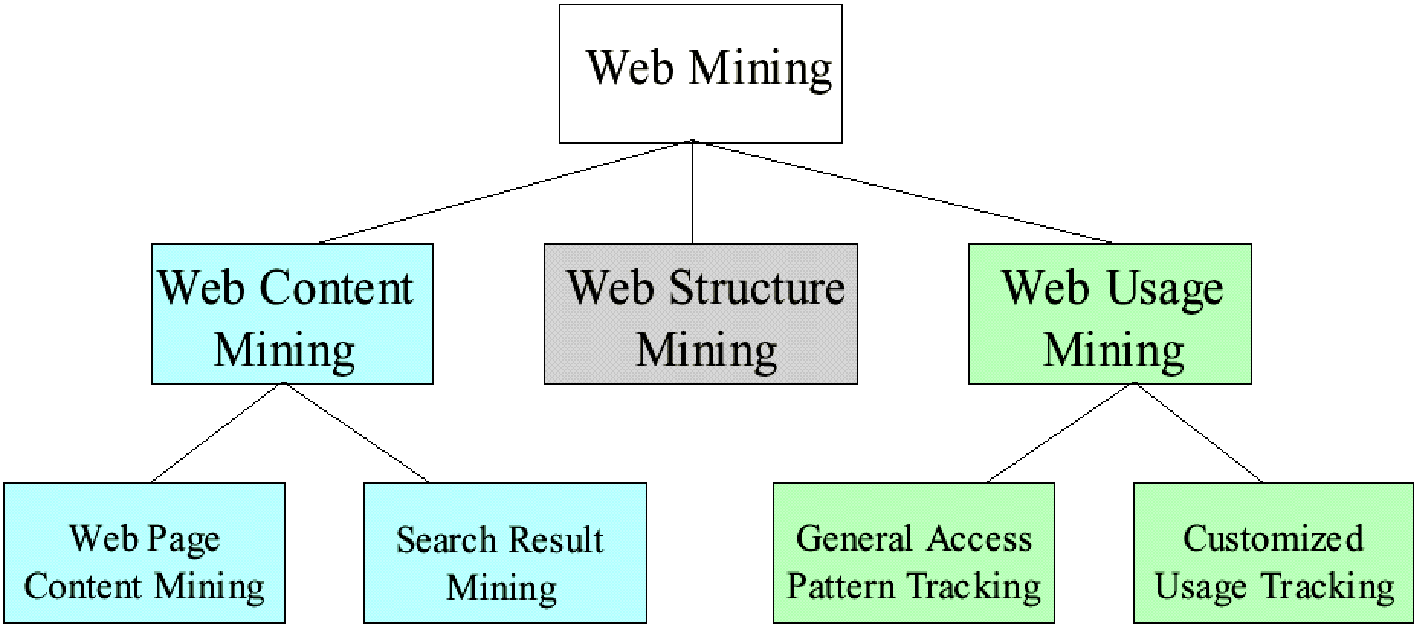
- Fraud detection and detection of unusual patterns (outliers)
  - Detect unusual patterns
  - Anti-Terrorism
  - Intrusion detection in network security.
  - Detection of credit card fraud.
  - Money Laundering: Detect suspicious money transactions.
  - Example: Terrorist Network [Ted Senator 2001]



8.2. Web Mining

- Web content: Text + Links
- Help web architects understand users needs
- User profiling
- Site structure

- Taxonomy:



- **Web Usage Mining**
  - Analyze web log to mine web users behavior (search engine, e-commerce, etc.)
  - Web personalization / collaborative filtering
  - Detection of new emerging research areas
  - Re-structure web sites based on users' needs
  - e-business intelligence, e-CRM, etc.
- **Web Content Mining**
  - Information filtering / knowledge extraction
  - Web document categorization
  - Detection of web categories and topics on the Web
- **Web Structure Mining**
  - Finding "Quality" or "authoritative" sites based on linkage and citation
    - ✓ IBM CLEVER project
    - ✓ Google

8.3. Text Mining

- Message filtering (e-mail, newsgroups, etc.)
- Newspaper articles analysis
- Text and document categorization

9. Data Mining Systems and Tools

- See www.kdnuggets.com
    - Oracle: Darwin
    - IBM: Intelligence Miner
    - SAS: Enterprise Miner
    - Business Objects
    - SPSS: Clementine
    - Xchange: e-CRM
    - Microsoft: SQL Server 2000
    - Weka
    - Etc.

10. Data Mining Functionalities

- Concept description: Characterization and discrimination
    - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions

- Association (correlation and causality)
    - Diaper ➔ Beer [0.5%, 75%]

- 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
    - Presentation: decision-tree, classification rule, neural network
    - Predict some unknown or missing numerical values

- <u>Cluster analysis</u>

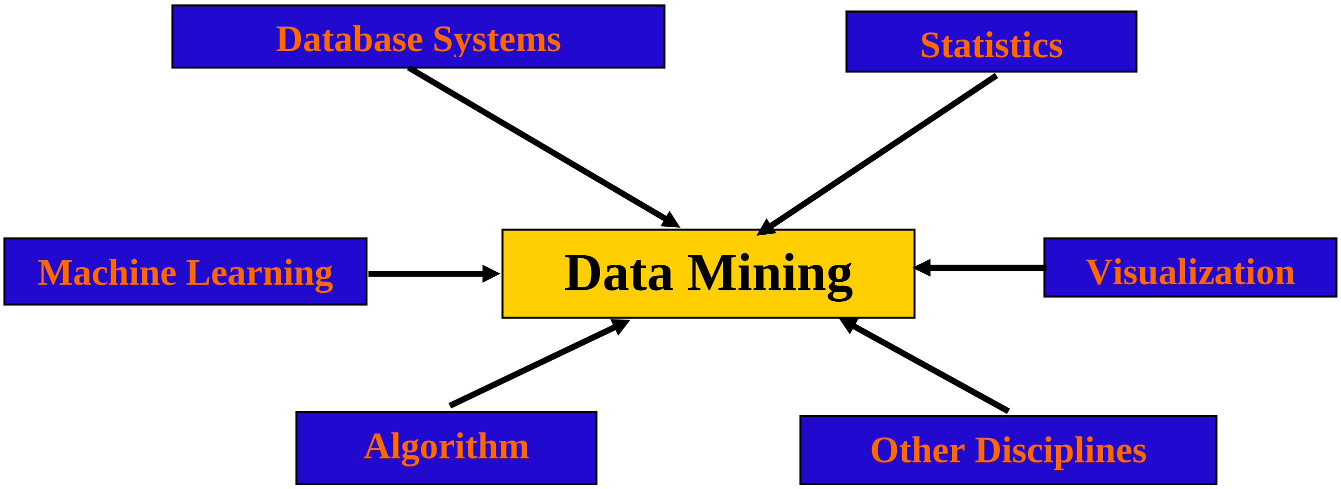
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Maximizing intra-class similarity & minimizing interclass similarity

- <u>Outlier analysis</u>
  - Outlier: a data object that does not comply with the general behavior of the data
  - Noise or exception? No! Useful in fraud detection, rare events analysis

- <u>Trend and evolution analysis</u>
  - Trend and deviation: regression analysis
  - Sequential pattern mining, periodicity analysis
  - Similarity-based analysis

- <u>Other pattern-directed or statistical analyses</u>

11. Data Mining: A multi-disciplinary area



12. Are All of the Patterns Interesting?

- Typically, thousands of patterns might be generated.

- How to get interesting patterns?

- What is an interesting pattern?

    - 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

- **<u>Objective vs. subjective interestingness measures</u>**

    - Objective: based on statistics and structures of patterns,
    - Example: support and confidence
        - Association rules:
            - Given an association rule: X ➔ Y
                - Rule support represents the percentage of transactions from a transaction database that the given rule satisfies.
                - Formally, it is the following probability:

                  $$P(XUY)$$

                  Where X U Y indicates a transaction that contains both X and Y.

            - Formally, it is denoted:

              $$\text{support}(X \rightarrow Y) = P(X \cup Y)$$

- Confidence rules:
    - Given an association rule: X ➔ Y
    - It assesses the degree of certainty of the detected association.
        - Formally, it is the conditional probability:

          P(Y|X) = The probability that a transaction containing X also contains Y.

    - Formally, it is denoted:

      confidence(X ➔ Y) = P(Y|X)

- Subjective: based on user’s belief in the data, e.g., unexpectedness, novelty, actionability, etc.

13. 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

14. Sample Datasets

- Document Collection: duc_nist_sample_text.txt
- 20 newsgroup
- Web Log: weblog_sample.txt
- Enron Data
- Intrusion data
- Medical data: Cancer Data