Chapter 1

1. Why Data Mining?

- **The Explosive Growth of Data**: from terabytes to petabytes
  - Major sources of data
    - **Business**: Web, e-commerce, transactions, stocks, …
    - **Science**: Remote sensing, bioinformatics, scientific simulation, …
    - **Society** and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- "Necessity is mother of invention"—Data mining—Automated analysis of massive data sets

2. Evolution of Sciences

➔ **empirical science**

➔ **theoretical science** (Each discipline has a *theoretical* component.)

➔ **computational science = simulation**

➔ **data science**

- ability to economically store and manage petabytes of data online
- Scientific info. management and visualization scale linearly with data volumes.
- **Data mining** is a major new challenge!

3.Evolution of Database Technology

➔ **database creation**, network DBMS

➔ **Relational** data model, relational DBMS implementation

➔ **RDBMS**, **OO** data models.
Application-oriented DBMS (spatial, scientific, engineering)

➔ **Data mining**, data warehousing, **multimedia DB,** and **Web DB**

➔ **Data mining and its applications**.
Web technology (XML, data integration) and global IS

4. What Is Data Mining? (knowledge discovery from data)

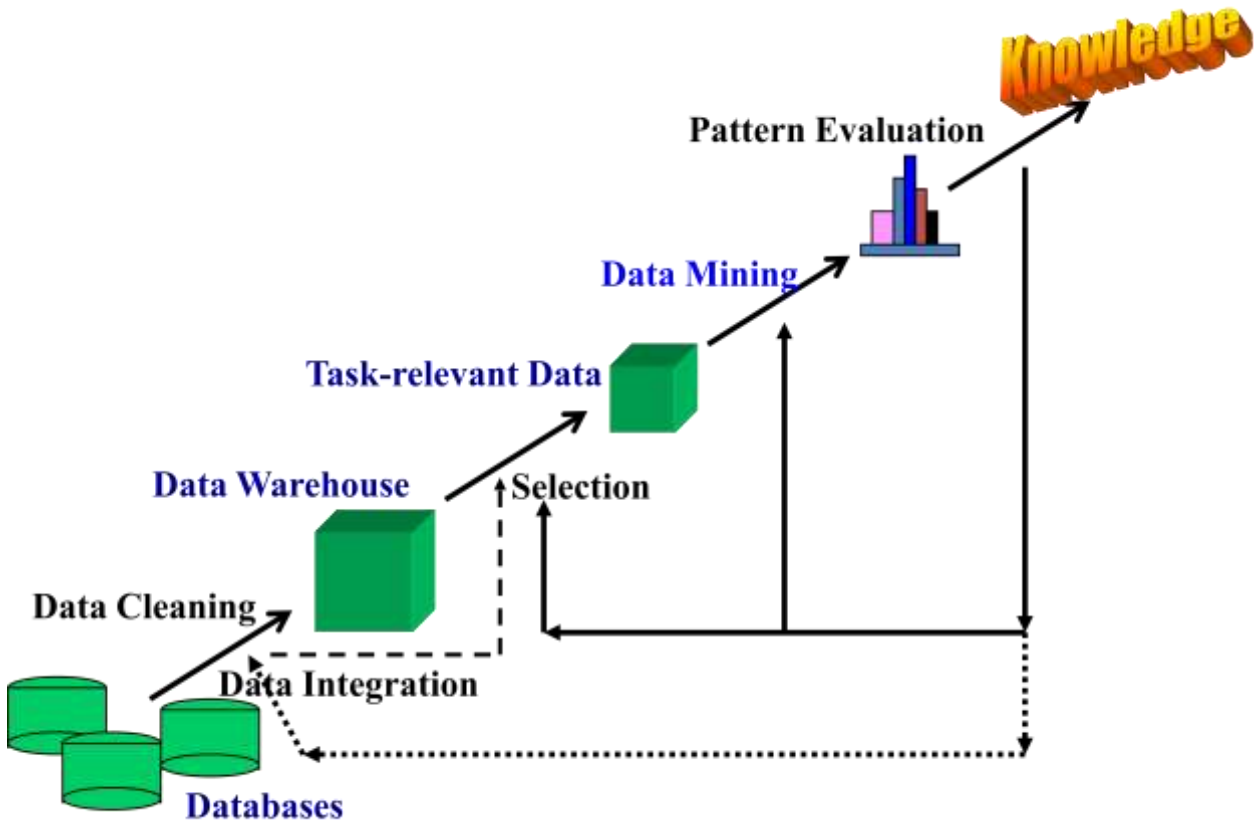
- **Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data**
- **Alternative names** : Knowledge discovery from databases (**KDD**), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, **business intelligence** (BI), etc.

5. Is everything "data mining"?

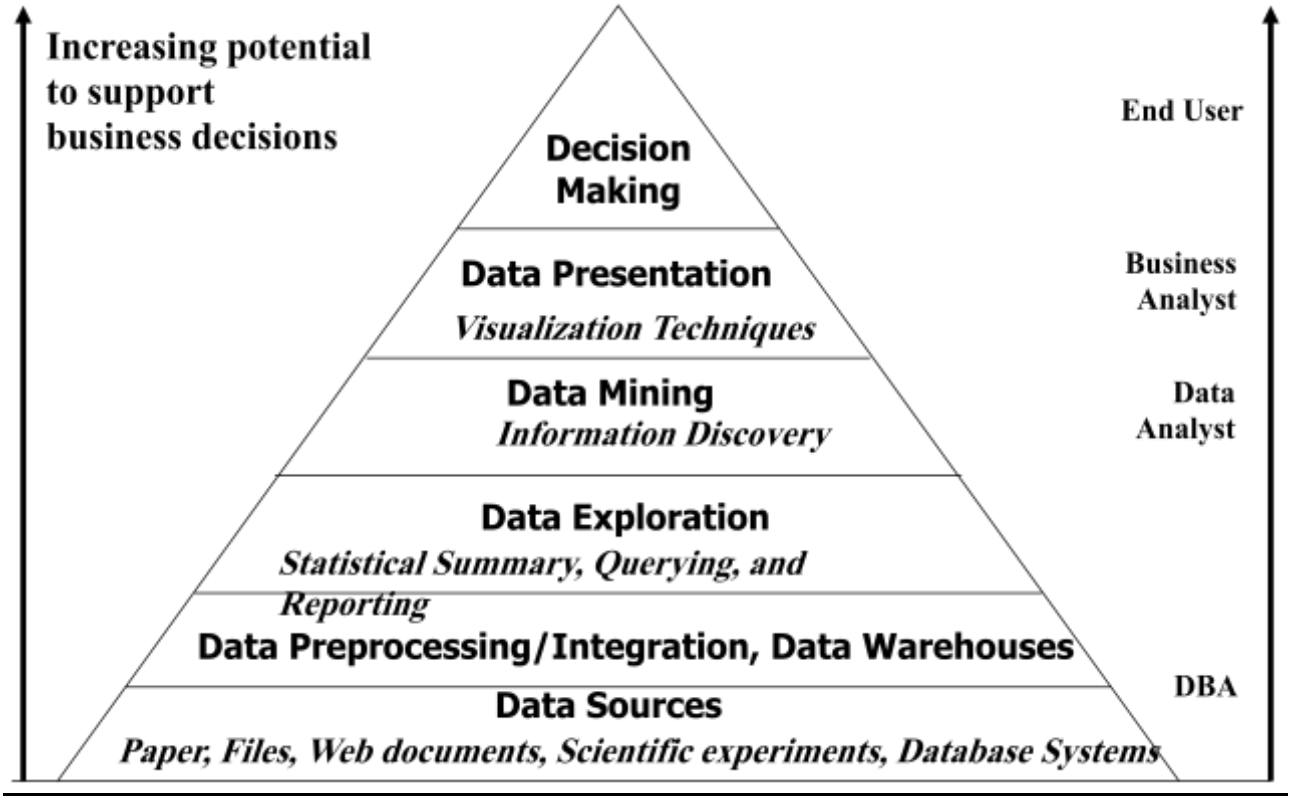
No : Simple search and query processing , (Deductive) expert systems

>>> 6. Draw view from typical database systems and data warehousing communities

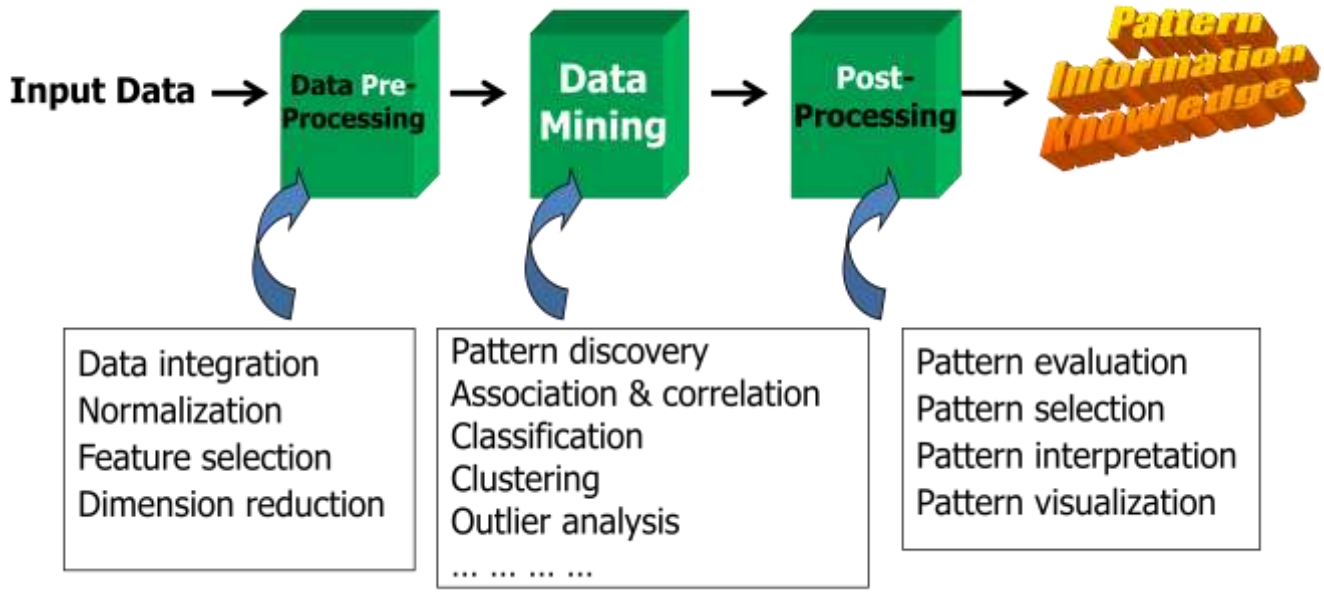
Or: draw Knowledge Discovery (KDD) Process



>>>7. Draw Data Mining in Business Intelligence



>>>8. KDD Process: A Typical View from ML and Statistics



9.Multi-Dimensional View of Data Mining

- **Data to be mined**
  - **Database data** (extended-relational, object-oriented, heterogeneous, legacy), **data warehouse**, transactional data, stream, spatiotemporal, time-series, **sequence**, **text and web**, **multi-media, graphs & social** and information networks
- **Knowledge to be mined (or: Data mining functions)**
  - **Characterization**, discrimination, **association**, **classification**, **clustering**, trend/deviation, **outlier** analysis, etc.
- **Techniques utilized**
  - **data warehouse (OLAP), machine learning, statistics**, pattern recognition, **visualization**, high-performance
- **Applications adapted**
  - Retail, **telecommunication**, **banking**, fraud analysis, bio-data mining, stock market analysis, **text mining, Web mining**, etc.

10. On What Kinds of Data?

11. Data Mining Function:

(1) Generalization

- Information **integration** and data **warehouse** construction
- Data **cube** technology
- **OLAP** (online analytical processing)
- **Generalize**, summarize, and contrast data **characteristics**, e.g., dry vs. wet region

(2) Association and Correlation Analysis

- **Frequent** patterns (or frequent itemsets)
- Association, correlation, causality
- A typical association rule: **Diaper → Beer**
- Are strongly associated items also strongly correlated? yes

(3) Classification

- Classification and label prediction
  - **Construct models (functions) based on some training examples**
  - **distinguish classes** or concepts for prediction
- **Typical methods** : Decision **trees**, naïve **Bayesian** classification, support vector machines, **neural networks**, **rule**-based classification, pattern-based classification,
- **Typical applications**: Credit card fraud detection, direct marketing, diseases,

(4) Cluster Analysis

- **Unsupervised** learning (i.e., **Class label is unknown**)
- Group data to form **new categories (i.e., clusters**),
- **Principle**: **Maximizing intra-class similarity & minimizing interclass similarity**
- Many methods and applications

(5) Outlier Analysis

- **Outlier**: A data object that **does not comply with the general** behavior of the data
- **Noise or exception?** — One person's garbage could be another person's treasure
- **Methods**: by product of **clustering** or regression analysis, …
- Useful in **fraud detection**, rare events analysis

12. Applications of Data Mining

- **Web page analysis**: from web page classification, clustering to PageRank
- **Collaborative analysis** & recommender systems
- **Biological & medical** data analysis: classification, cluster analysis, sequence analysis
- **Data mining and software engineering**
- From data mining systems/tools (e.g., SAS, MS SQL-Server Analysis Manager, **Oracle**) to invisible data mining

13. Time and Ordering Or Sequential Pattern, Trend and Evolution Analysis

- **Sequence, trend and evolution analysis**
  - **Sequential** pattern mining : e.g., buy digital camera→buy memory cards
  - **Trend**, time-series, and deviation analysis: e.g., regression and value prediction
  - **Motifs** and biological sequence analysis
  - Periodicity analysis
  - Similarity-based analysis
- **Mining data streams :** Ordered, time-varying, infinite, data streams

14 Structure and Network Analysis

- **Graph mining** : Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- **Information network analysis**
  - **Social networks**: actors (objects, nodes) and relationships (edges)
    - e.g., author networks , terrorist networks
  - **Heterogeneous networks**
    - A person is multiple information networks: friends, family, classmates, …
  - **Links** carry a lot of semantic information: Link mining
- **Web mining**
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks : community discovery, opinion

15 Are all mined knowledge interesting? No

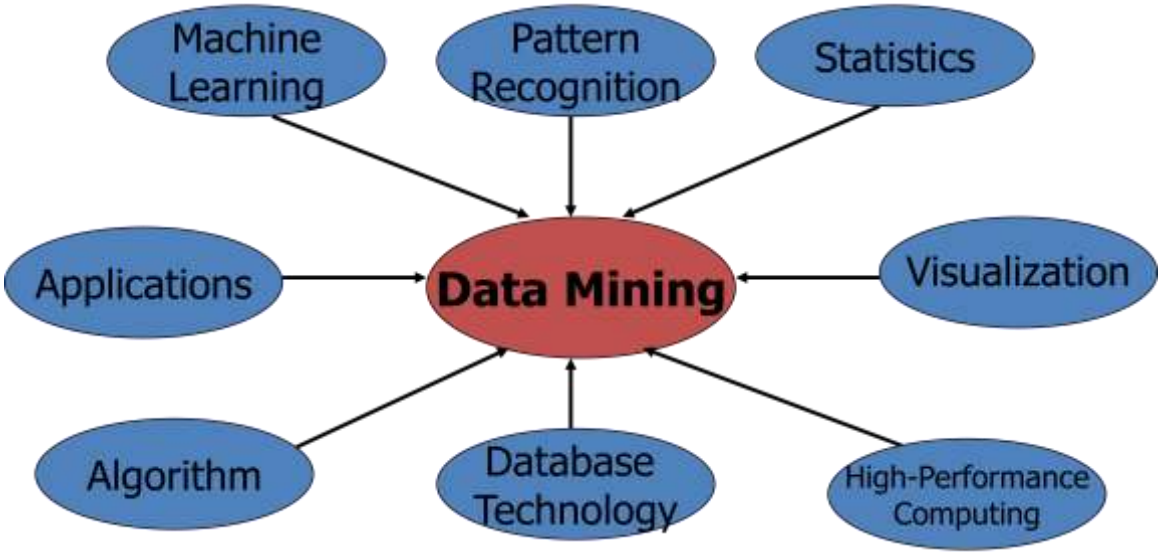
- One can mine tremendous **amount of "patterns"** and knowledge
- Some may **fit** only certain dimension space (**time, location**, …)
- Some may **not be representative**, may be transient, …

16 Evaluation of mined knowledge

directly mine only interesting knowledge?

- Descriptive vs. predictive
- Coverage
- Typicality vs. novelty
- Accuracy
- Timeliness

17 Data Mining: Confluence of Multiple Disciplines



18 Why Confluence of Multiple Disciplines?

- **Tremendous amount of data :** Algorithms must be highly scalable to tera-bytes
- **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**

>>>19. Major Issues in Data Mining

- **Mining Methodology**
  - Mining various and new **kinds of knowledge**
  - Mining knowledge in **multi-dimensional space**
  - Data mining is **interdisciplinary** effort
  - Boosting power of **discovery in a networked** environment
  - **Handling noise**, uncertainty, and incompleteness of data
- **User Interaction = Interactive mining**
  - Incorporation of **background knowledge**
  - **Presentation and visualization** of data mining results
- **Efficiency and Scalability**
  - Efficiency and scalability of data mining algorithms
  - **Parallel, distributed**, stream, and incremental mining methods
- **Diversity of data types**
  - Handling **complex types** of data
  - **Mining dynamic, networked**, and global data repositories
- **Data mining and society**
  - **Social impacts** of data mining
  - **Privacy-preserving** data mining
  - **Invisible data mining**