Data Mining:

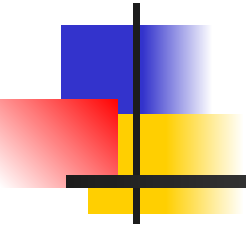# **Books : References**

Book 1: Introduction to Data Mining by
Pang-Ning Tan, Michale Steinbach, Anuj Karpathe, Vipin
Kumar.

Book 2: Data Mining Concepts and Techniques by
Jiawei Han, Micheline Kamber, Jian Pei.

Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes

  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube

Evolution of Database Technology

- 1960s:
    - Data collection, database creation, IMS and network DBMS
- 1970s:
    - Relational data model, relational DBMS implementation
- 1980s:
    - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
    - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s:
    - Data mining, data warehousing, multimedia databases, and Web databases
- 2000s
    - Stream data management and mining
    - Data mining and its applications
    - Web technology (XML, data integration) and global information systems

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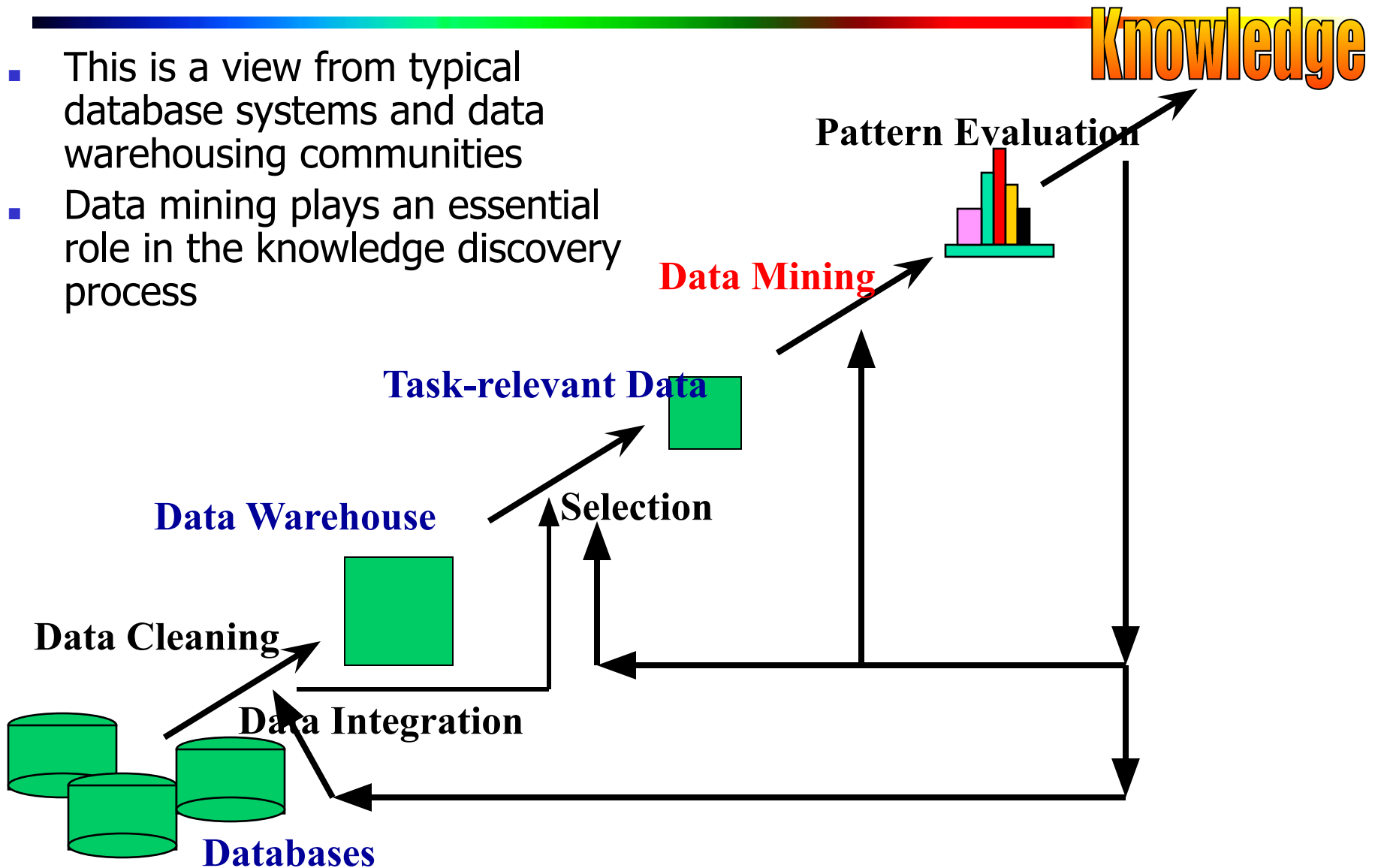
- Data mining (knowledge discovery from data)
  - Extraction of interesting (<u>non-trivial,</u> <u>implicit</u>, <u>previously</u> <u>unknown</u> and <u>potentially useful)</u> patterns or knowledge from huge amount of data
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge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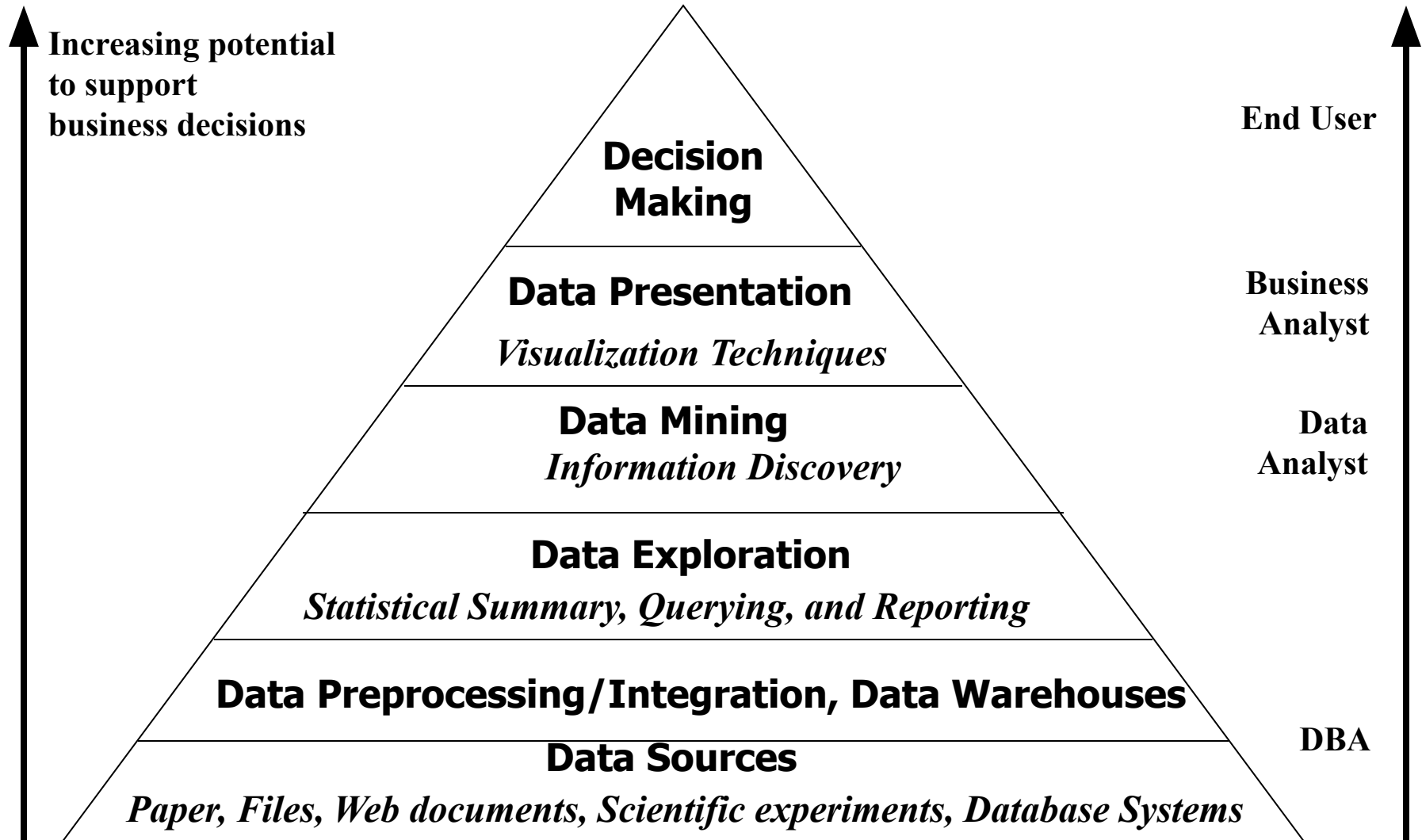
Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process

Knowledge

Pattern Evaluation

Data Mining

Task-relevant Data

Selection

Data Warehouse

Data Cleaning

Data Integration

Databases

Example: A Web Mining Framework

- Web mining usually involves
    - Data cleaning
    - Data integration from multiple sources
    - Warehousing the data
    - Data cube construction
    - Data selection for data mining
    - Data mining
    - Presentation of the mining results
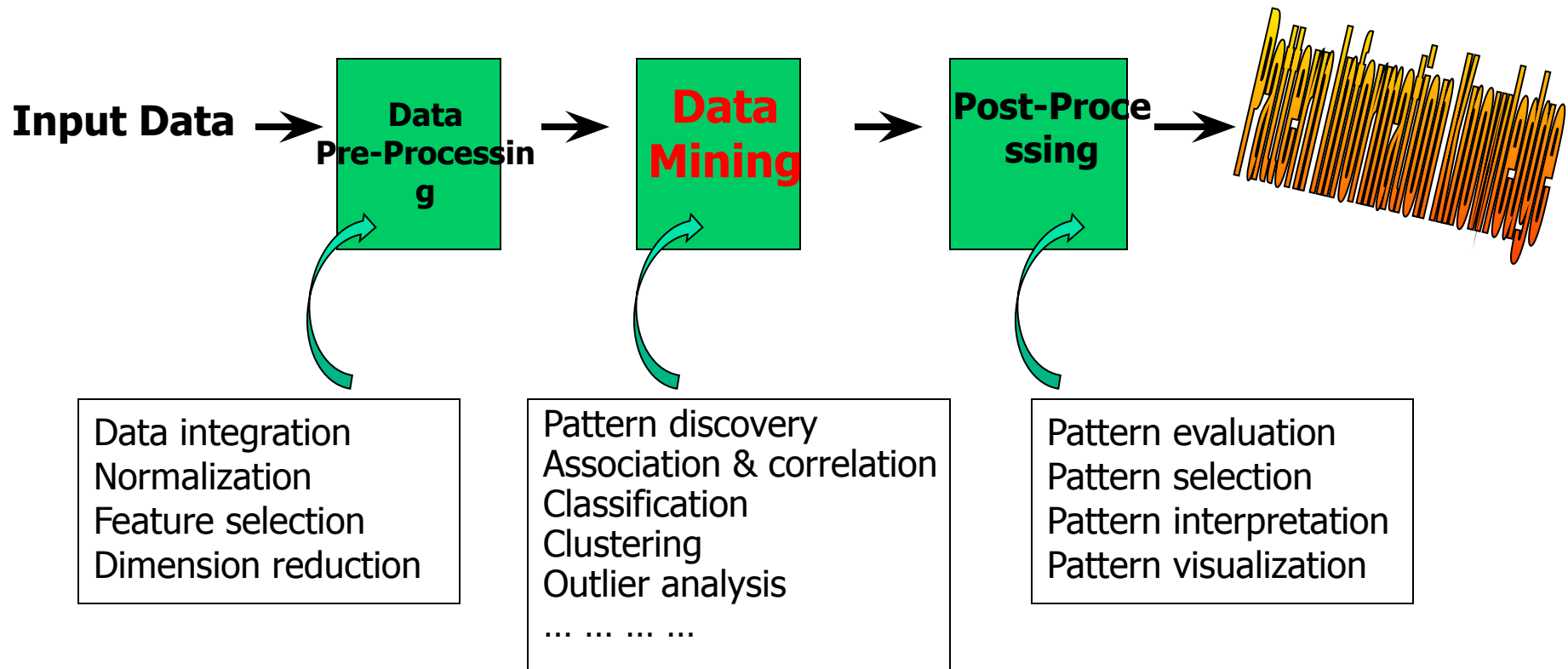   - Patterns and knowledge to be used or stored into knowledge-base

Data Mining in Business Intelligence

Increasing potential to support business decisions

Decision Making — End User

Data Presentation
Visualization Techniques — Business Analyst

Data Mining
Information Discovery — Data Analyst

Data Exploration
Statistical Summary, Querying, and Reporting

Data Preprocessing/Integration, Data Warehouses

Data Sources — DBA
Paper, Files, Web documents, Scientific experiments, Database Systems

Example: Mining vs. Data Exploration

- Business intelligence view
  - Warehouse, data cube, reporting but not much mining
- Business objects vs. data mining tools
- Supply chain example: tools
- Data presentation
- Exploration

KDD Process: A Typical View from ML and Statistics

Input Data → **Data Pre-Processing** → **Data Mining** → **Post-Processing** → Pattern Information Knowledge

| Data integration
Normalization
Feature selection
Dimension reduction | Pattern discovery
Association & correlation
Classification
Clustering
Outlier analysis
… … … … | Pattern evaluation
Pattern selection
Pattern interpretation
Pattern visualization |

- This is a view from typical machine learning and statistics communities

Example: Medical Data Mining

- Health care & medical data mining – often adopted such a view in statistics and machine learning

- Preprocessing of the data (including feature extraction and dimension reduction)

- Classification or/and clustering processes

- Post-processing for presentation

Multi-Dimensional View of Data Mining

- **<u>Data to be mined</u>**
    - 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
- **<u>Knowledge to be mined (or: Data mining functions)</u>**
    - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
    - Descriptive vs. predictive data mining
    - Multiple/integrated functions and mining at multiple levels
- **<u>Techniques utilized</u>**
    - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.
- **<u>Applications adapted</u>**
    - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

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 Function: (1) Generalization

- Information integration and data warehouse construction
  - Data cleaning, transformation, integration, and multidimensional data model
- Data cube technology
  - Scalable methods for computing (i.e., materializing) multidimensional aggregates
  - OLAP (online analytical processing)
- Multidimensional concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet region

Data Mining Function: (2) Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association, correlation vs. causality
  - A typical association rule
    - Diaper ☐ Beer [0.5%, 75%] (support, confidence)
  - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

Data Mining Function: (3) Classification

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - 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 class labels
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …

Data Mining Function: (4) 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
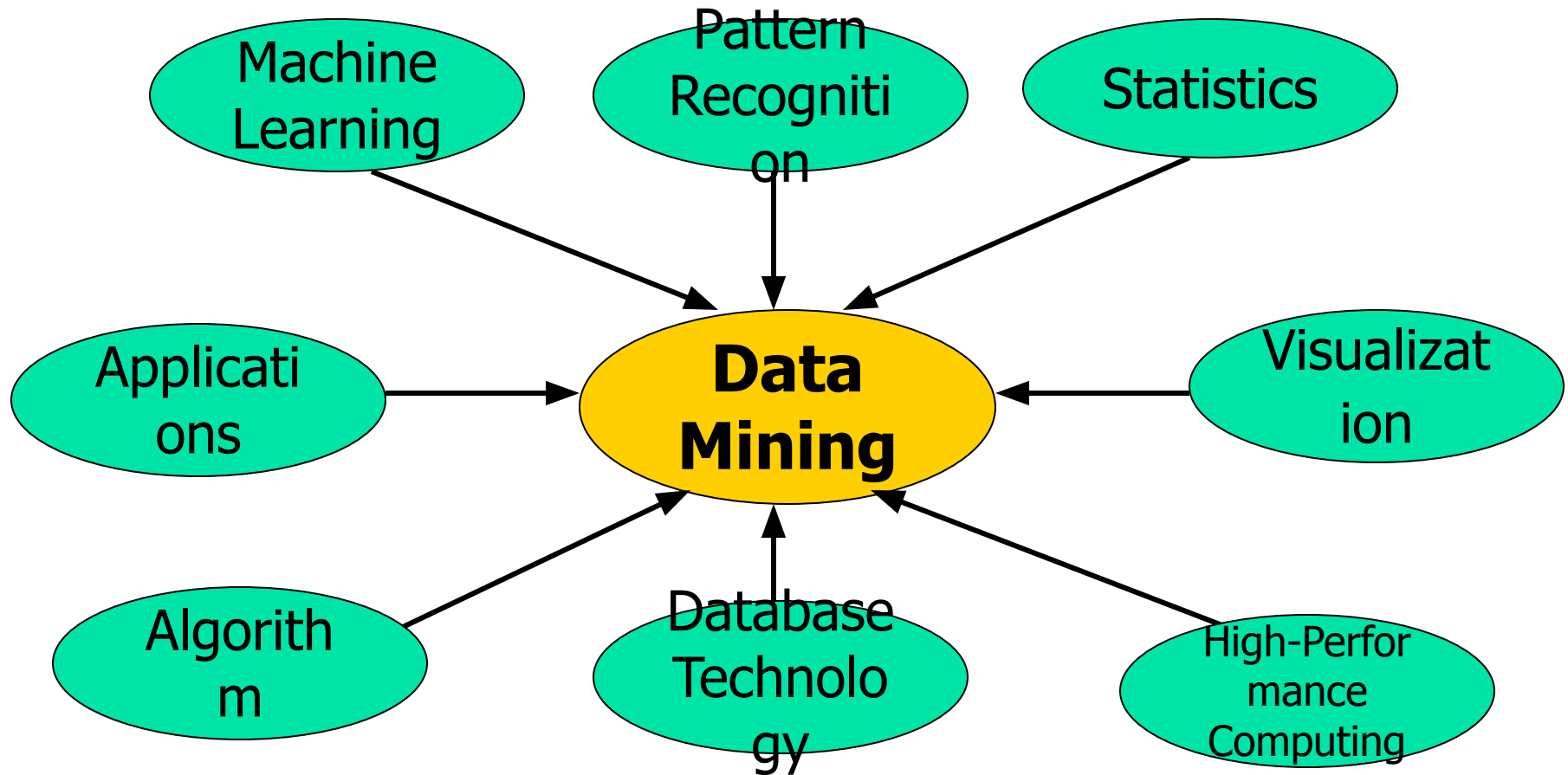- Many methods and applications

Data Mining Function: (5) Outlier Analysis

- 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

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 in CS, terrorist networks
  - Multiple heterogeneous networks
    - A person could be 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
    - Web community discovery, opinion mining, usage mining, …

Data Mining: Confluence of Multiple Disciplines

Machine Learning

Pattern Recognition

Statistics

Applications

Data Mining

Visualization

Algorithm

Database Technology

High-Performance Computing

Why Confluence of Multiple Disciplines?

- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as 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

Applications of Data Mining

- Web page analysis: from web page classification, clustering to PageRank & HITS algorithms

- Collaborative analysis & recommender systems

- Basket data analysis to targeted marketing

- Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis

- Data mining and software engineering (e.g., IEEE Computer, Aug. 2009 issue)

- From major dedicated data mining systems/tools (e.g., SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools) to invisible data mining

A Brief History of Data Mining Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
    - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
    - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
    - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
    - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.
- ACM Transactions on KDD starting in 2007

Conferences and Journals on Data Mining

- KDD Conferences
  - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
  - SIAM Data Mining Conf. (SDM)
  - (IEEE) Int. Conf. on Data Mining (ICDM)
  - European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining (ECML-PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
  - Int. Conf. on Web Search and Data Mining (WSDM)

- Other related conferences
  - DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, …
  - Web and IR conferences: WWW, SIGIR, WSDM
  - ML conferences: ICML, NIPS
  - PR conferences: CVPR,
- Journals
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD

Summary

- Data mining: Discovering interesting patterns and knowledge from massive amount of data

- A natural evolution of database technology, in great demand, with wide applications

- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation

- Mining can be performed in a variety of data

- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.

- Data mining technologies and applications

- Major issues in data mining

Typical data warehouse AllElectronics