

Data Mining: Concepts and Techniques

Outline

- Why & What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- What Kind of Data Can Be Mined?
- What Kinds of Patterns Can Be Mined?
- What Technology Are Used?
- What Kind of Applications Are Targeted?
- Major Issues in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary

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

Evolution of Sciences

- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
 - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
 - Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
 - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science**
 - The flood of data from new scientific instruments and simulations
 - The ability to economically store and manage petabytes of data online
 - The Internet and computing Grid that makes all these archives universally accessible
 - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. [Data mining](#) is a major new challenge!

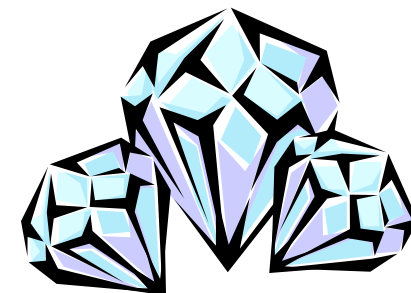
Evolution of Database Technology

- 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?

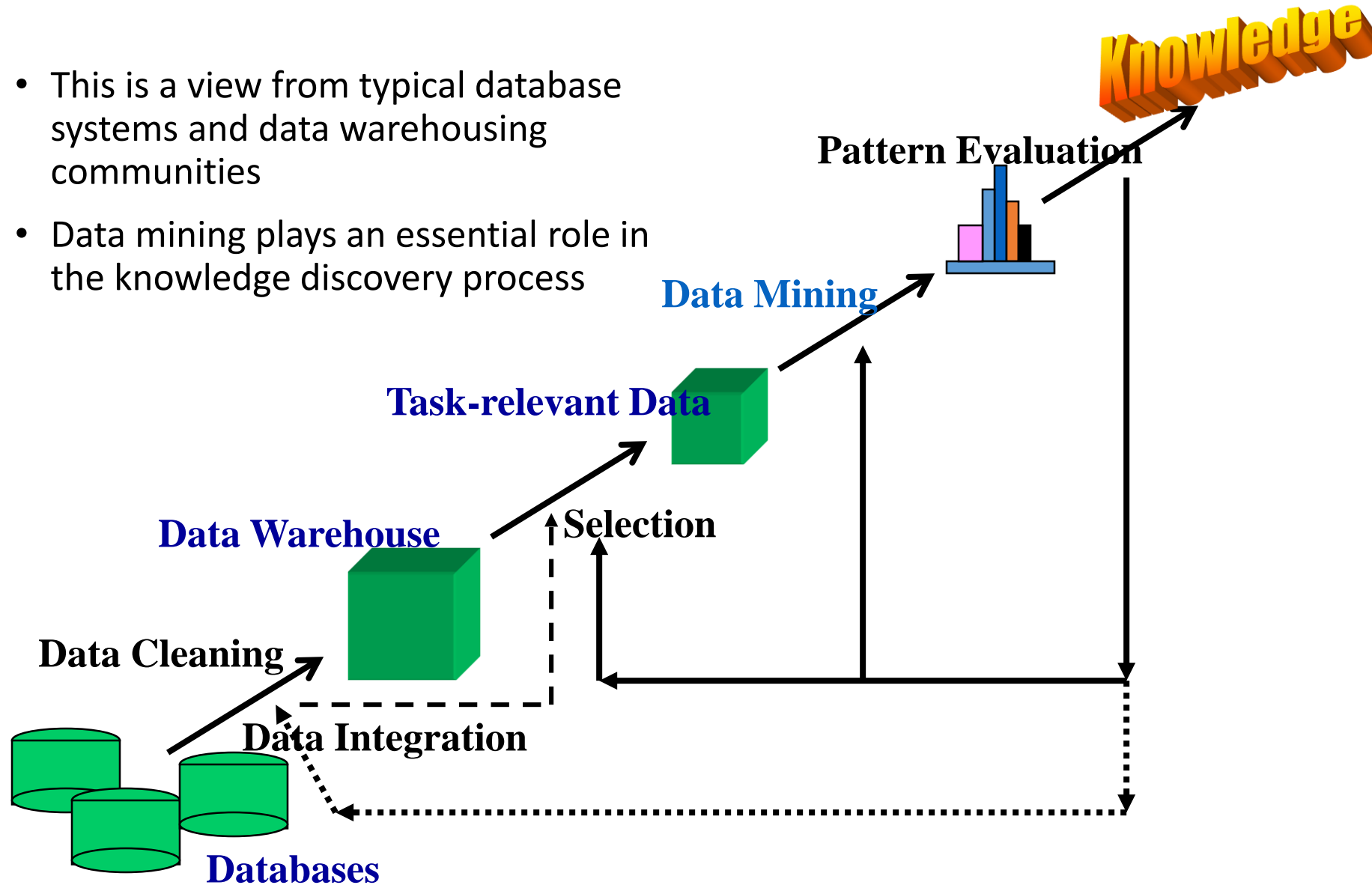


- Data mining (knowledge discovery from data)
 - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Alternative names
 - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis,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

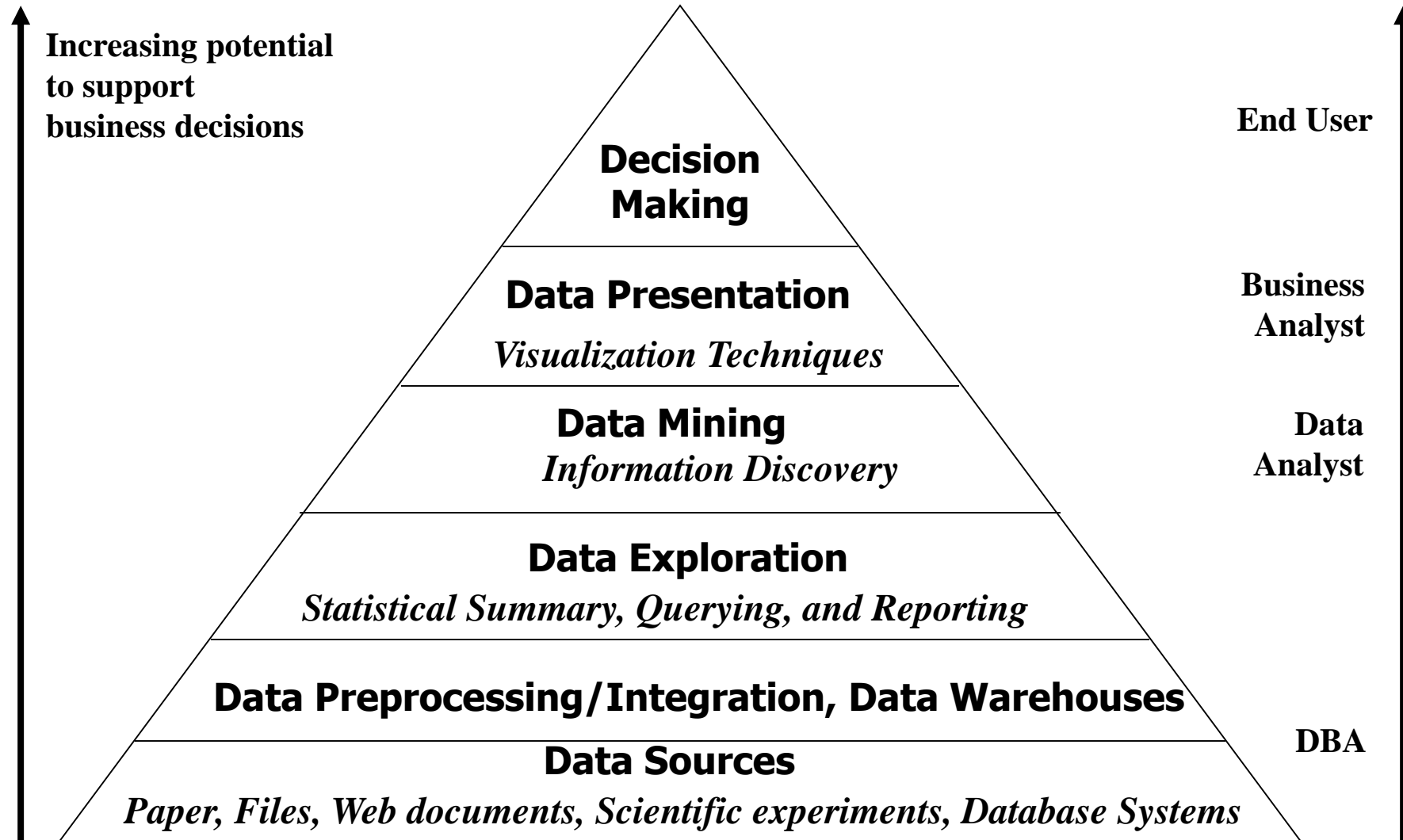


Knowledge Discovery (KDD) Process

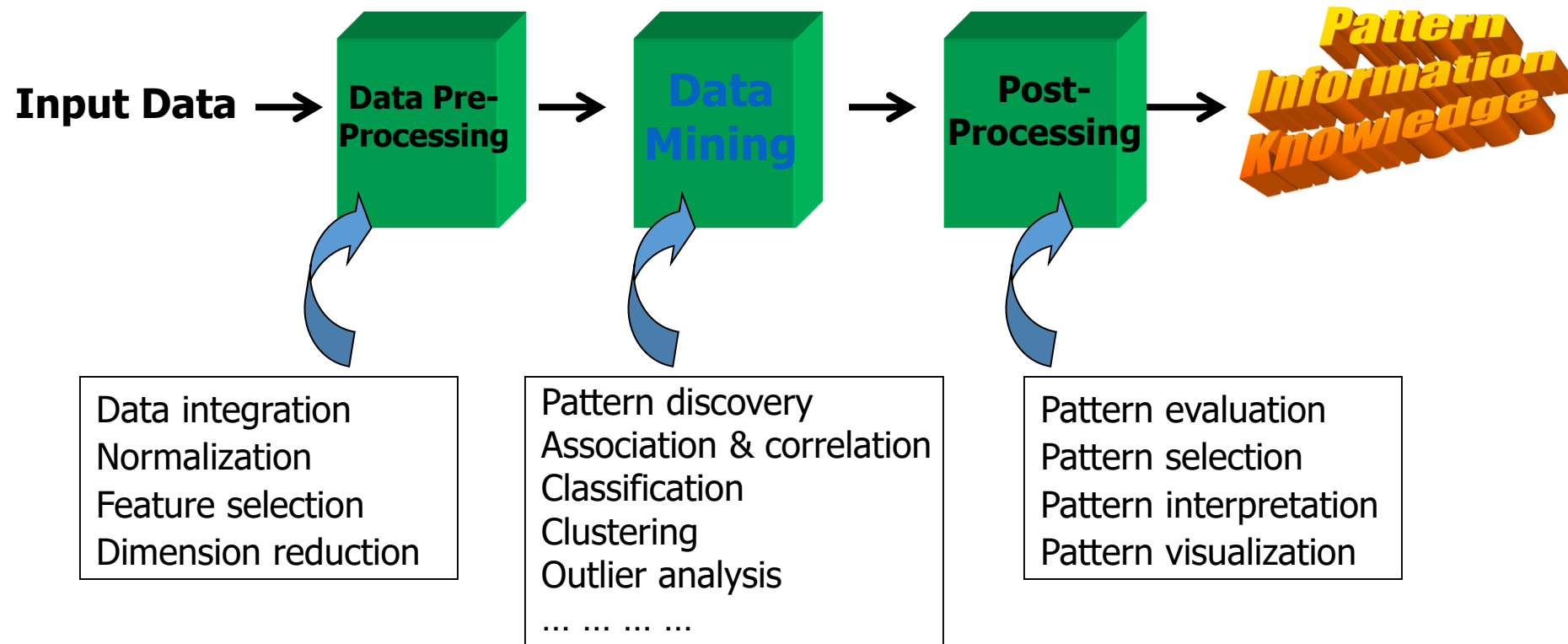
- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



Data Mining in Business Intelligence



KDD Process: A Typical View from ML and Statistics



- 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

- **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-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.

- **Applications adapted**

- 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 \rightarrow 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

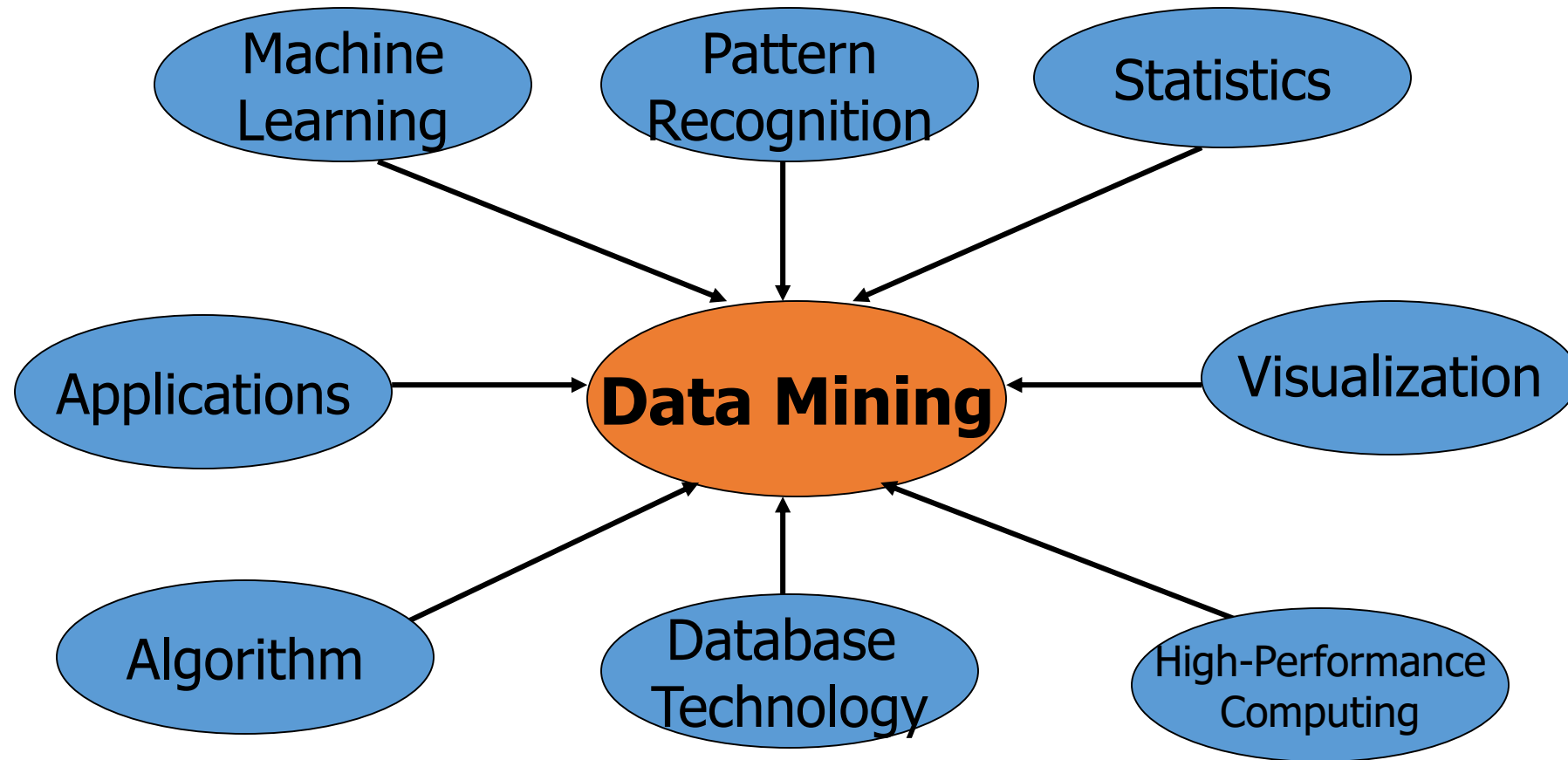
Data Mining Function: (5) Outlier Analysis

- 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

Evaluation of Knowledge

- Are all mined knowledge interesting?
 - 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, ...
- Evaluation of mined knowledge → directly mine only interesting knowledge?
 - Descriptive vs. predictive
 - Coverage
 - Typicality vs. novelty
 - Accuracy
 - Timeliness
 - ...

Data Mining: Multiple Disciplines



Why 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

Applications of Data Mining (2): Cyber Security

- Due to the availability of large amounts of data in cyberinfrastructure and the number of cyber criminals attempting to gain access to the data, data mining, machine learning, statistics, and other interdisciplinary capabilities are needed to address the challenges of cybersecurity.
- Data mining is an interdisciplinary field that employs the use of analysis tools from statistical models, mathematical algorithms, and machine learning methods to discover previously unknown, valid patterns and relationships in large data sets, which are useful for finding hackers and preserving privacy in cybersecurity.

Major Issues in Data Mining (1)

- Mining Methodology
 - Mining various and new kinds of knowledge
 - Mining knowledge in multi-dimensional space
 - Data mining: An interdisciplinary effort
 - Boosting the power of discovery in a networked environment
 - Handling noise, uncertainty, and incompleteness of data
 - Pattern evaluation and pattern- or constraint-guided mining
- User Interaction
 - Interactive mining
 - Incorporation of background knowledge
 - Presentation and visualization of data mining results

Major Issues in Data Mining (2)

- 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

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

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