Introduction to Data Mining

Why Data Mining?

- **Explosive Growth of Data**
  - Data collection and data availability
    - Automated data collection tools, Internet, smartphones, …
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, biotechnology, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- **We are drowning in data, but starving for knowledge!**

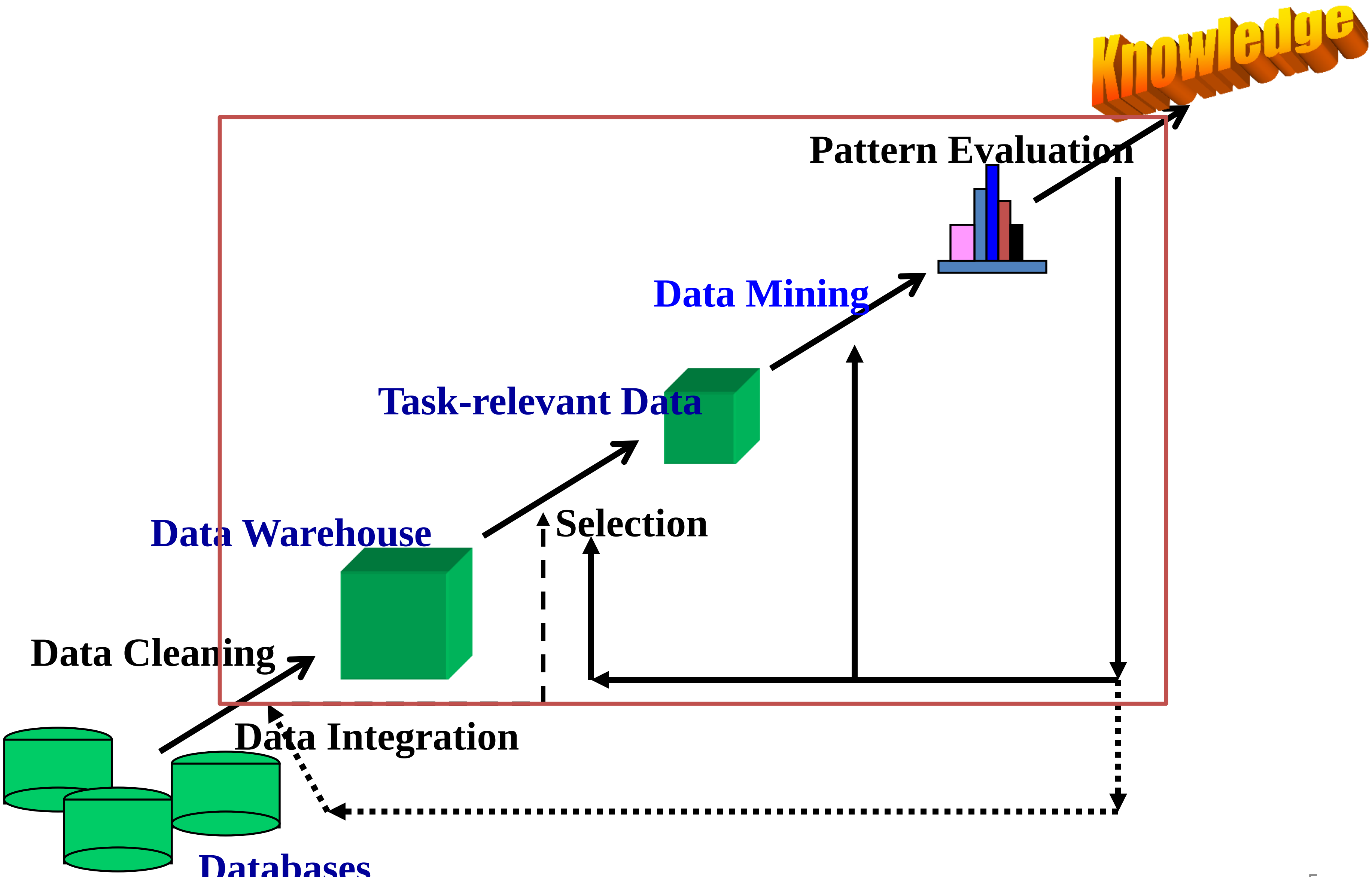
Decision Support

- **Typical procedure**
  - Data -> Knowledge -> Action/Decision -> Goal
- **Examples**
  - Netflix collects user ratings of movies (data) => What types of movies you will like (knowledge) => Recommend new movies to you (action) => Users stay with Netflix (goal)
  - Gene sequences of cancer patients (data) => Which genes lead to cancer? (knowledge) => Appropriate treatment (action) => Save life (goal)
  - Road traffic (data) => Which road is likely to be congested? (knowledge) => Suggest better routes to drivers (action) => Save time and energy (goal)

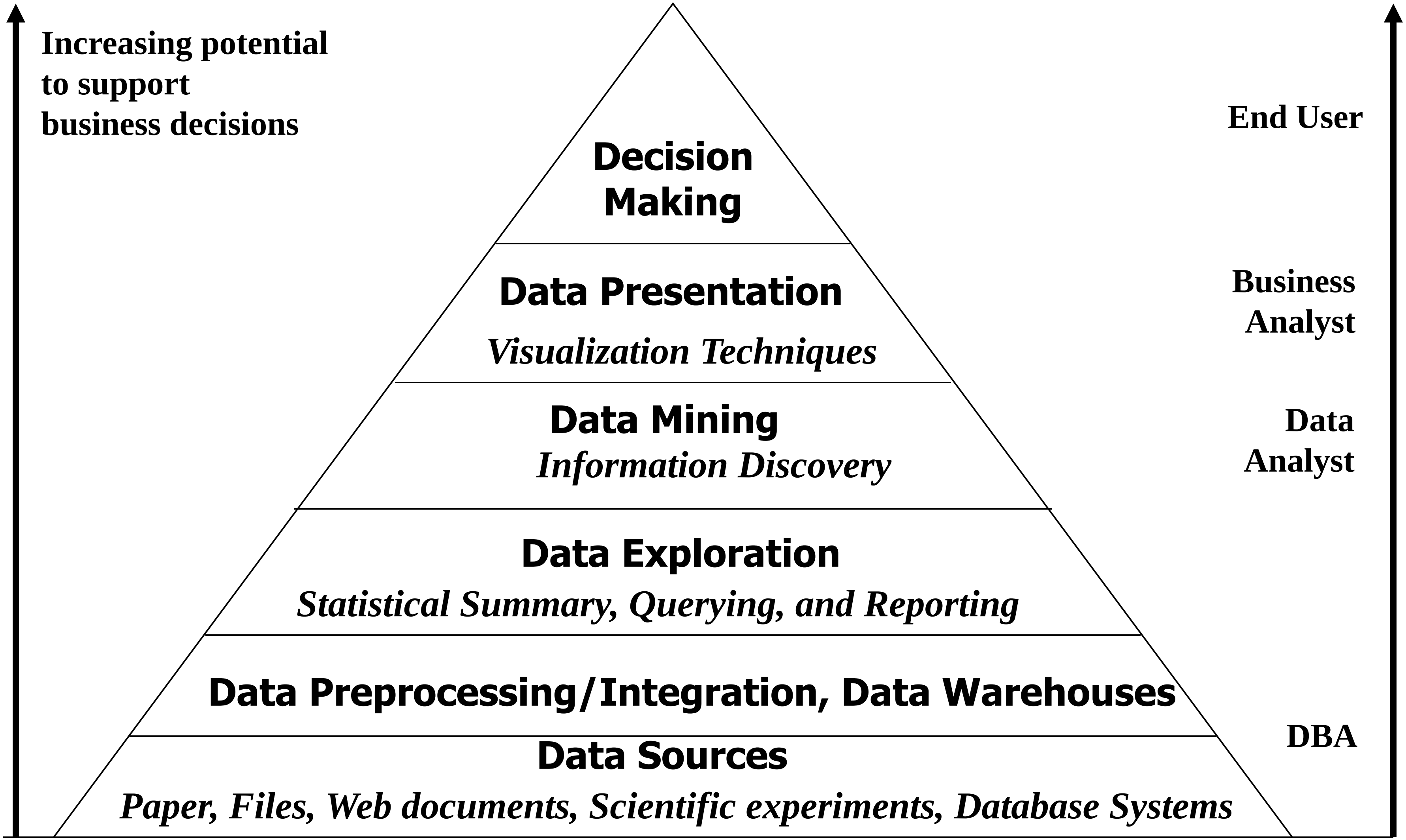
What Is Data Mining?

- **Data mining**
  - 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, etc.
- **Watch out: Is everything "data mining"?**
  - Simple search and query processing
  - (Deductive) expert systems

Data Mining Process



Data Mining Procedure



Increasing potential
to support
business decisions

Decision
Making

Data Presentation

Visualization Techniques

Data Mining

Information Discovery

Data Exploration

Statistical Summary, Querying, and Reporting

Data Preprocessing/Integration, Data Warehouses

Data Sources

Paper, Files, Web documents, Scientific experiments, Database Systems

End User

Business
Analyst

Data
Analyst

DBA

Multi-Dimensional View of Data Mining

- **Data to be mined**
  - Transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Knowledge to be mined**
  - Association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Descriptive vs. predictive data mining
- **Techniques utilized**
  - Data warehouse (OLAP), machine learning, statistics, pattern recognition, optimization, visualization, 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?

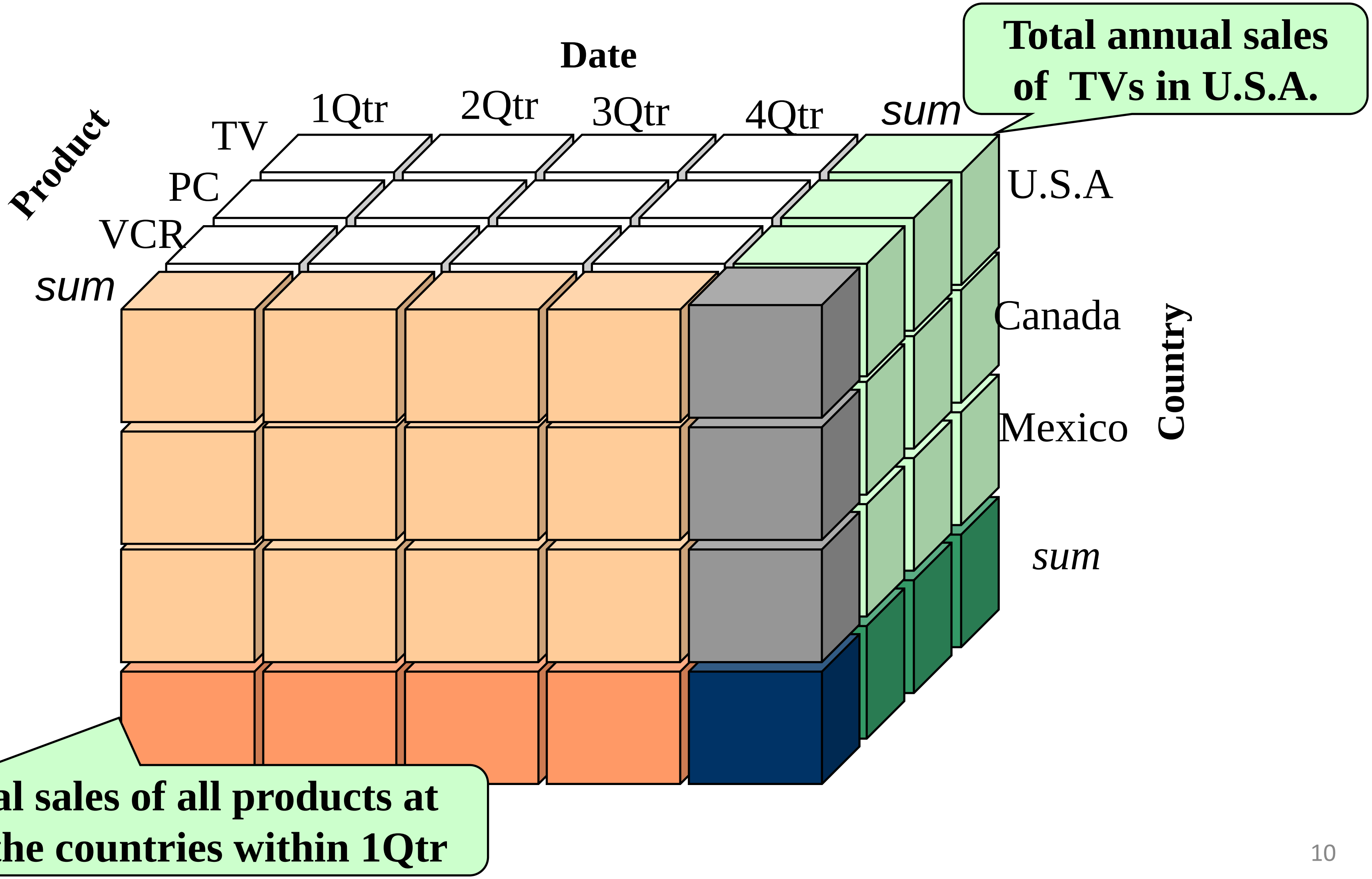
- Relational database, data warehouse, transactional database
- Data streams and sensor data
- Time-series data, temporal data, sequence data
- Structure data, graphs, social networks and multi-linked data
- Spatial data and spatiotemporal data
- Multimedia data
- Text data
- WWW data

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)

Data Warehousing

- Aggregate data from different dimensions



Total annual sales of TVs in U.S.A.

Total sales of all products at all the countries within 1Qtr

Data Mining Function: (2) Association Analysis

- **Frequent patterns (or frequent itemsets)**
  - – What items are frequently purchased together in Walmart?
- **Association rules**
  - – A typical association rule
    - Diaper → Beer [0.5%, 75%] (support, confidence)
- **How to mine such patterns and rules efficiently in large datasets?**

Association Rule Mining

- Data: A set of transactions, and each transaction consists of a set of items
- Association rules: A set of rules that characterize associations between items

Market-Basket transactions

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

Rules Discovered:

{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

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
  - 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:**
  - Identifying spams, predicting treatment outcomes, categorizing articles, …

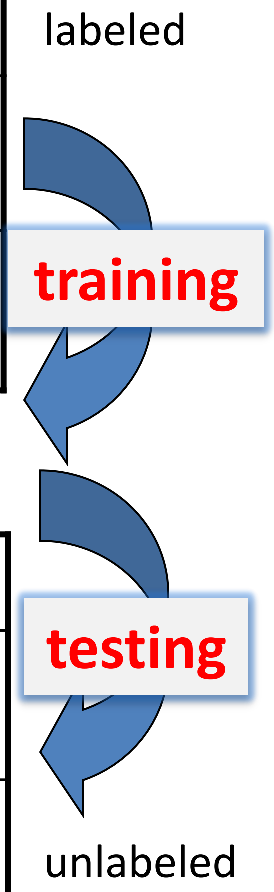
Classification

features class labels

user	age	gender	education	Ad?
	27	Female	Bachelor	Yes
	30	Male	PhD	Yes
	55	Male	Bachelor	No

a classifier: f(*x*)=*y*: features → class labels

user	age	gender	education	Ad?
	60	Female	Bachelor	
	23	Male	Master	

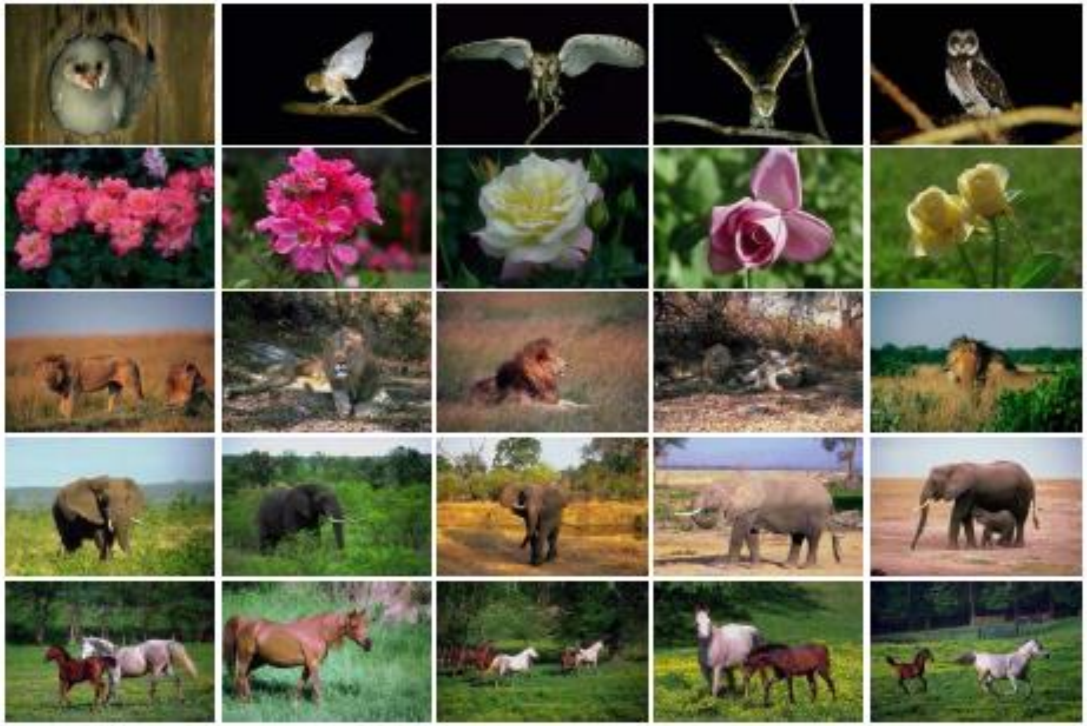


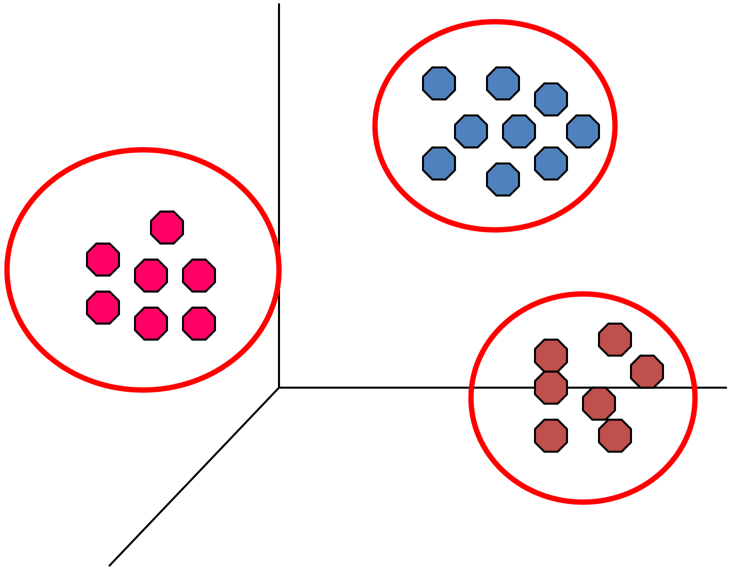
Data Mining Function: (4) Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Partition data into groups based on object similarity
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Methods: Partitional, hierarchical, density-based, mixture model, spectral methods
- Applications: document clustering, user log clustering, target marketing, climate modeling, …

Clustering

- Finding groups of objects such that the objects in a group will be similar to one another and different from the objects in other groups





Data Mining Function: (5) Anomaly Detection

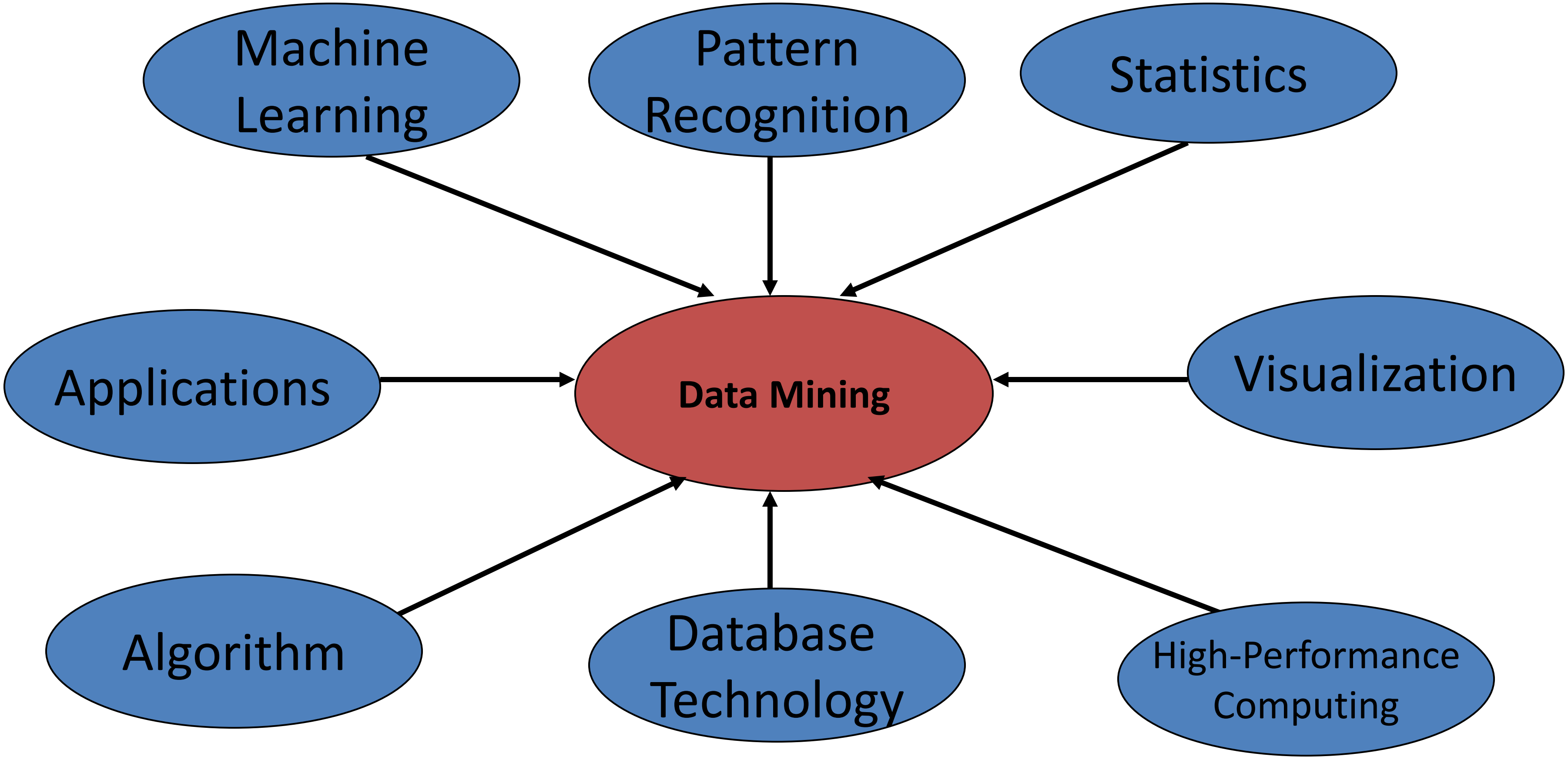
- **Anomalies**
  - the set of objects are considerably dissimilar from the remainder of the data
  - occur relatively infrequently
  - when they do occur, their consequences can be quite dramatic and quite often in a negative sense
- **Approaches**
  - Statistics-based, depth-based, model-based, by product of cluster analysis
- **Applications**
  - credit card frauds, network intrusions, system failures, water leak, ...

“Mining needle in a haystack.
So much hay and so little time”

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

Data Mining: Confluence of Multiple Disciplines



Challenges in Data Mining

- **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**
  - Noisy and unreliable
  - Dynamically evolving
  - High dimensionality
  - Multiple heterogeneous sources
- **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
- Social media analysis: mine user opinions and obtain insights from data collected from social networking platforms

Major Issues in Data Mining (1)

- **Mining Methodology**
  - Mining various and new kinds of knowledge
  - Mining knowledge from different perspectives
  - 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

Take-away Message

- Data Mining refers to non-trivial extraction of implicit, previously unknown and potentially useful knowledge from data
- Data Mining covers topics including warehousing, association analysis, clustering, classification, anomaly detection, etc. (based on the type of mined knowledge), as well as transaction data mining, stream data mining, sequence data mining, graph data mining, etc. (based on the type of data)
- Data Mining has wide applications in many different fields in business, science, engineering, education, and many more