Knowledge Discovery in Databases

Overview of Data Warehouse and Data Mining

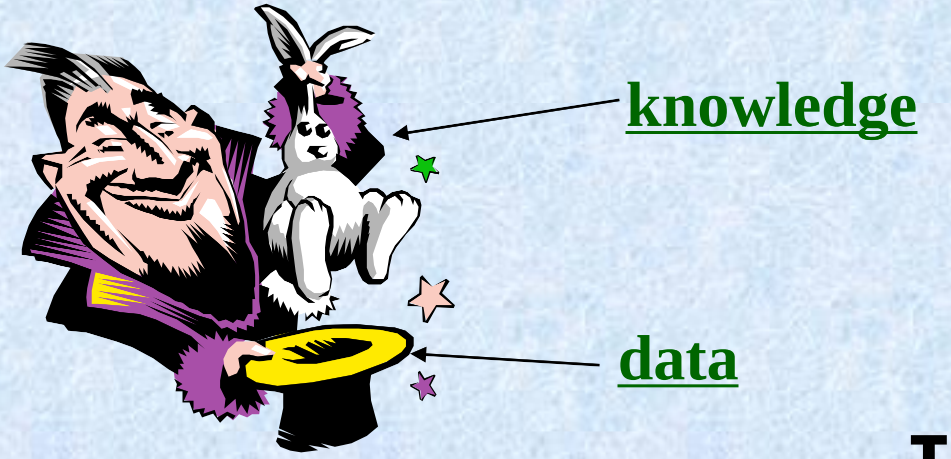
School of Software, Nanjing University

Introduction

- Motivation: Why data mining?
- What is data mining?
- Data Mining: On what kind of data?
- Data mining functionality
- Are all the patterns interesting?
- Classification of data mining systems
- Major issues in data mining

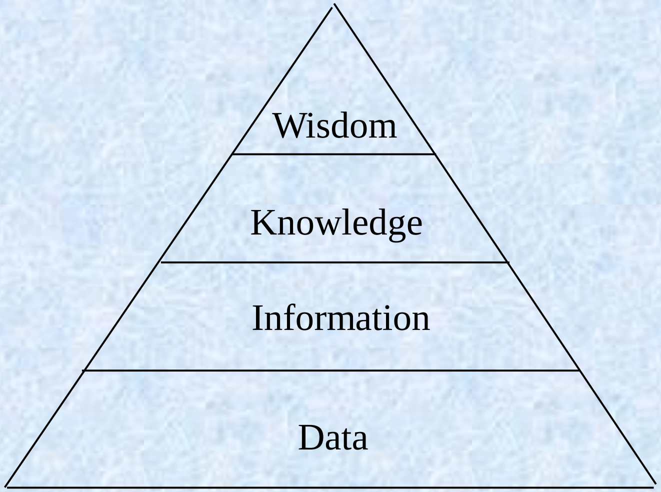
What are data? What is knowledge?

- **We can easily get a lot of data, while these data are meaningless to us**
- **Then what is the thing we really need?**
  - Knowledge is something meaningful drawn from data
  - Knowledge is just what is useful to you.



knowledge

data



Wisdom

Knowledge

Information

Data

The Knowledge Hierarchy

Motivation: "Necessity is the Mother of Invention"

- **Data explosion problem**
  - Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories
- **We are drowning in data, but starving for knowledge!**
- **Solution: Data warehousing and data mining**
  - Data warehousing and on-line analytical processing
  - Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

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.) and application-oriented DBMS (spatial, scientific, engineering, etc.)
- **1990s—2000s:**
  - Data mining and data warehousing, multimedia databases, and Web databases

What Is Data Mining?

- **Data mining (knowledge discovery in databases):**
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases

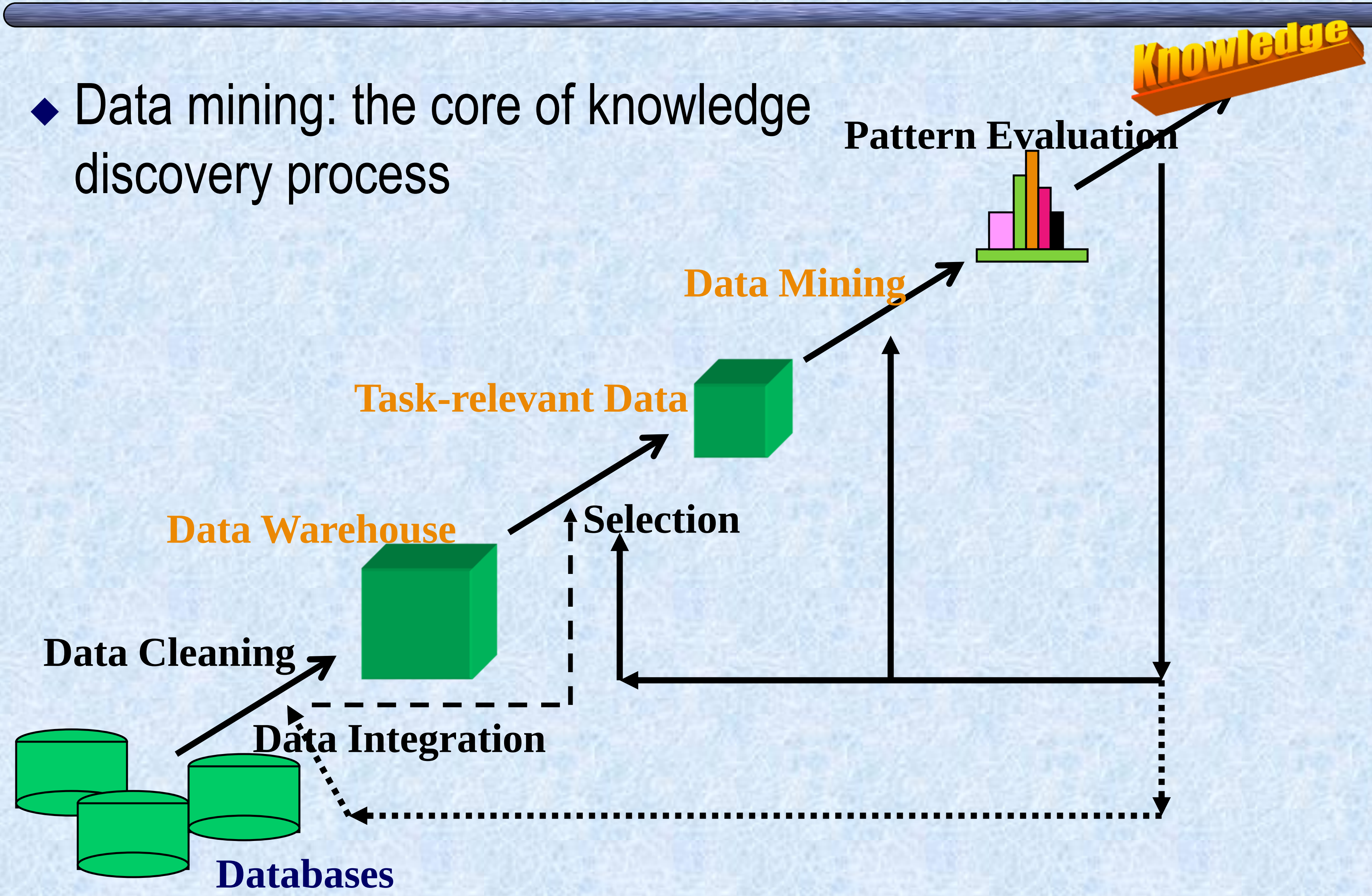
- **Alternative names and their “inside stories”:**
  - Knowledge discovery(mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Why Data Mining? — Potential Applications

- **Database analysis and decision support**
  - Market analysis and management
    - ✓ target marketing, customer relation management, market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - ✓ Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and management
- **Other Applications**
  - Text mining (news group, email, documents) and Web analysis.
  - Intelligent query answering

Data Mining: A KDD Process

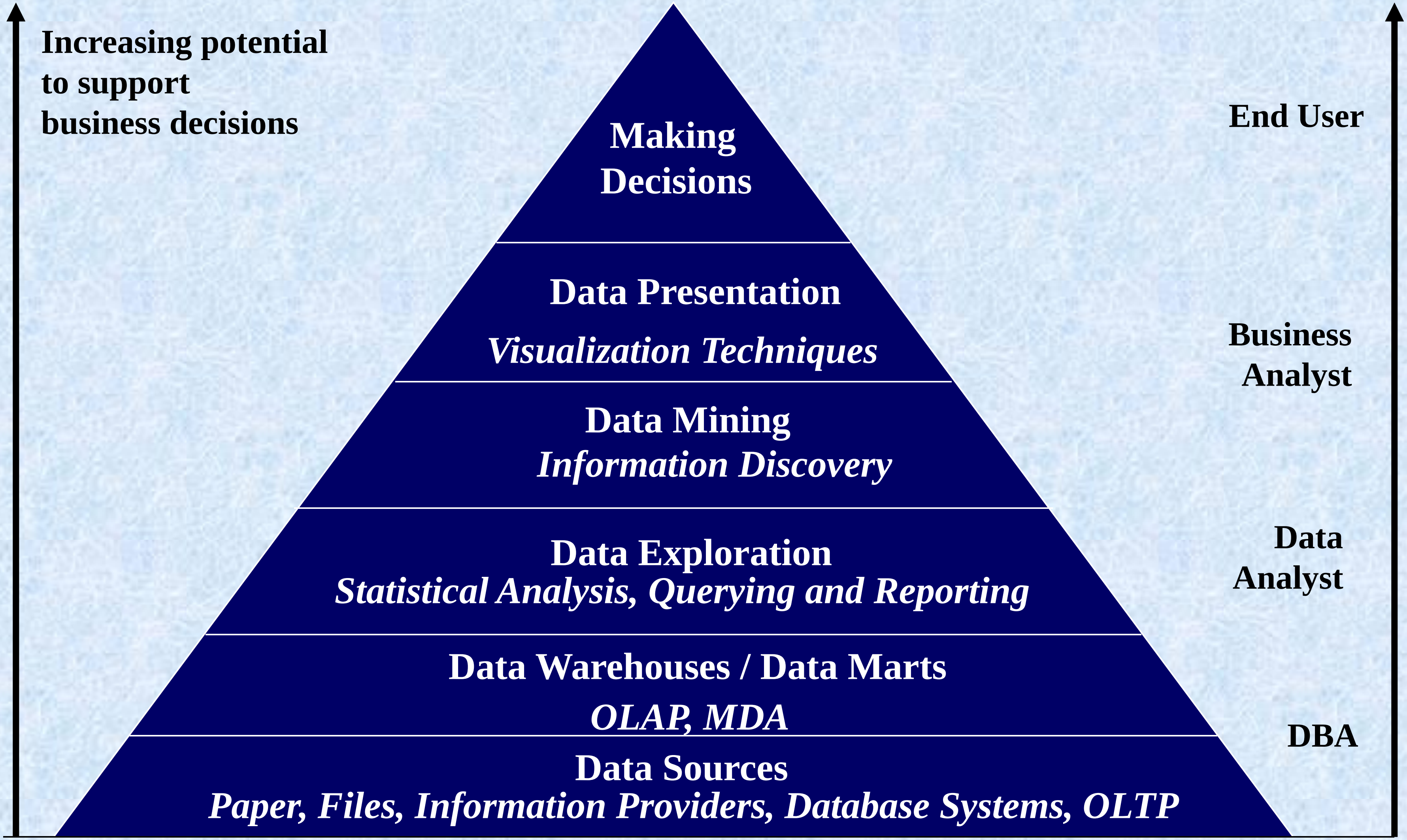
- Data mining: the core of knowledge discovery process



Steps of a KDD Process

- **Learning the application domain:**
  - relevant prior knowledge and goals of application
- **Creating a target data set: data selection**
- **Data cleaning and preprocessing: (may take 60% of effort!)**
- **Data reduction and transformation:**
  - Find useful features, dimensionality/variable reduction, invariant representation.
- **Choosing functions of data mining**
  - summarization, classification, regression, association, clustering.
- **Choosing the mining algorithm(s)**
- **Data mining: search for patterns of interest**
- **Pattern evaluation and knowledge presentation**
  - visualization, transformation, removing redundant patterns, etc.
- **Use of discovered knowledge**

Data Mining and Business Intelligence



Increasing potential
to support
business decisions

Making Decisions

Data Presentation

Visualization Techniques

Data Mining

Information Discovery

Data Exploration

Statistical Analysis, Querying and Reporting

Data Warehouses / Data Marts

OLAP, MDA

Data Sources

Paper, Files, Information Providers, Database Systems, OLTP

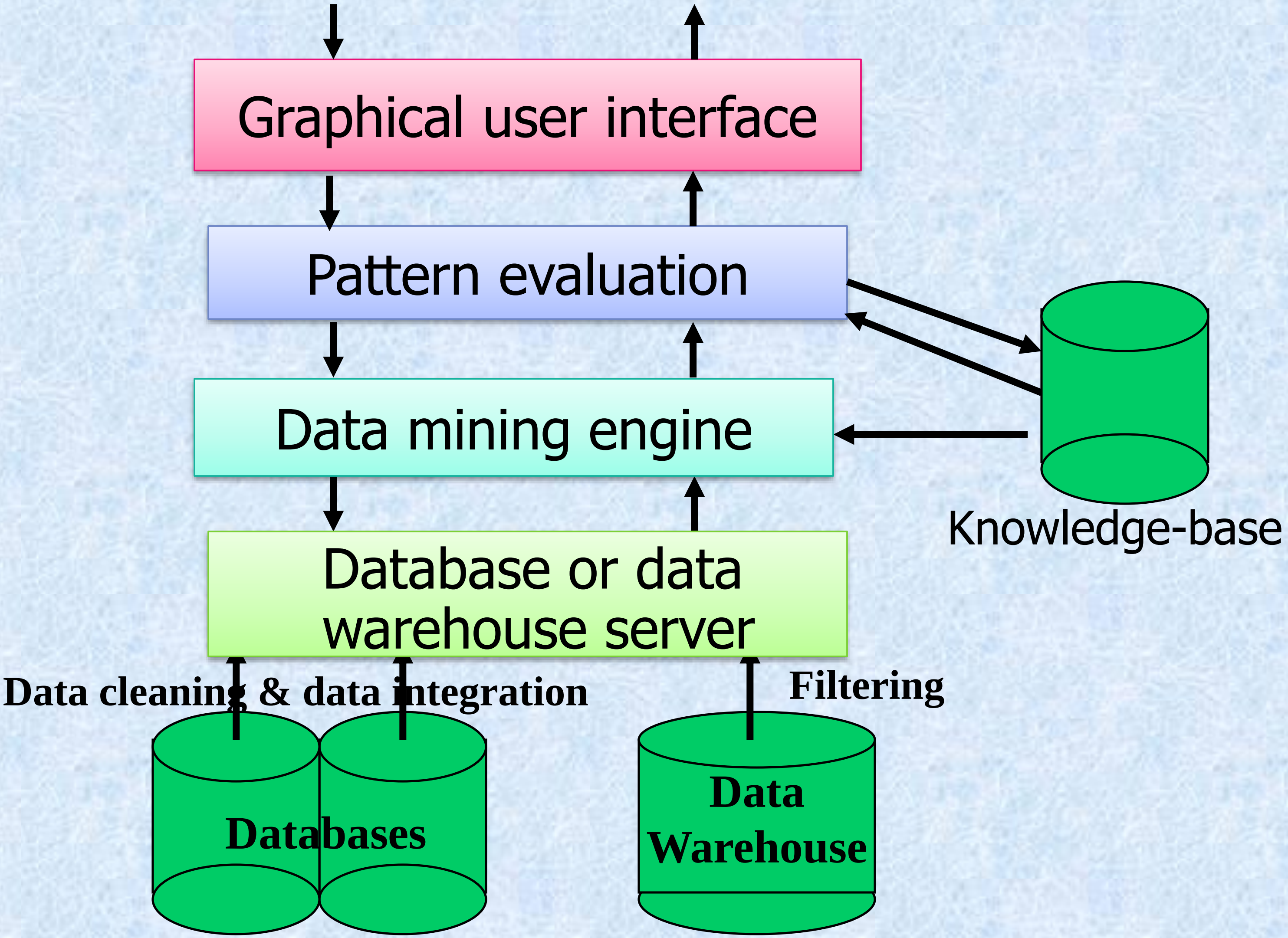
End User

Business Analyst

Data Analyst

DBA

Architecture of a Typical Data Mining System



Data Mining: On What Kind of Data?

- Relational databases
- Data warehouses
- Transactional databases
- Advanced DB and information repositories
  - Object-oriented and object-relational databases
  - Spatial databases
  - Time-series data and temporal data
  - Text databases and multimedia databases
  - Heterogeneous and legacy databases
  - WWW

Data Mining Functionalities (1)

- **Concept description: Characterization and discrimination**
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
- **Association (correlation and causality)**
  - Multi-dimensional vs. single-dimensional association
  - age(X, “20..29”) ^ income(X, “20..29K”) → buys(X, “PC”) [support = 2%, confidence = 60%]
  - contains(T, “computer”) → contains(x, “software”) [1%, 75%]

Data Mining Functionalities (2)

- **Classification and Prediction**
  - Finding 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
  - Prediction: Predict some unknown or missing numerical values
- **Cluster analysis**
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Clustering based on the principle: maximizing the intra-class similarity and minimizing the interclass similarity

Data Mining Functionalities (3)

- **Outlier analysis**
  - Outlier: a data object that does not comply with the general behavior of the data
  - It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis
- **Trend and evolution analysis**
  - Trend and deviation: regression analysis
  - Sequential pattern mining, periodicity analysis
  - Similarity-based analysis
- **Other pattern-directed or statistical analyses**

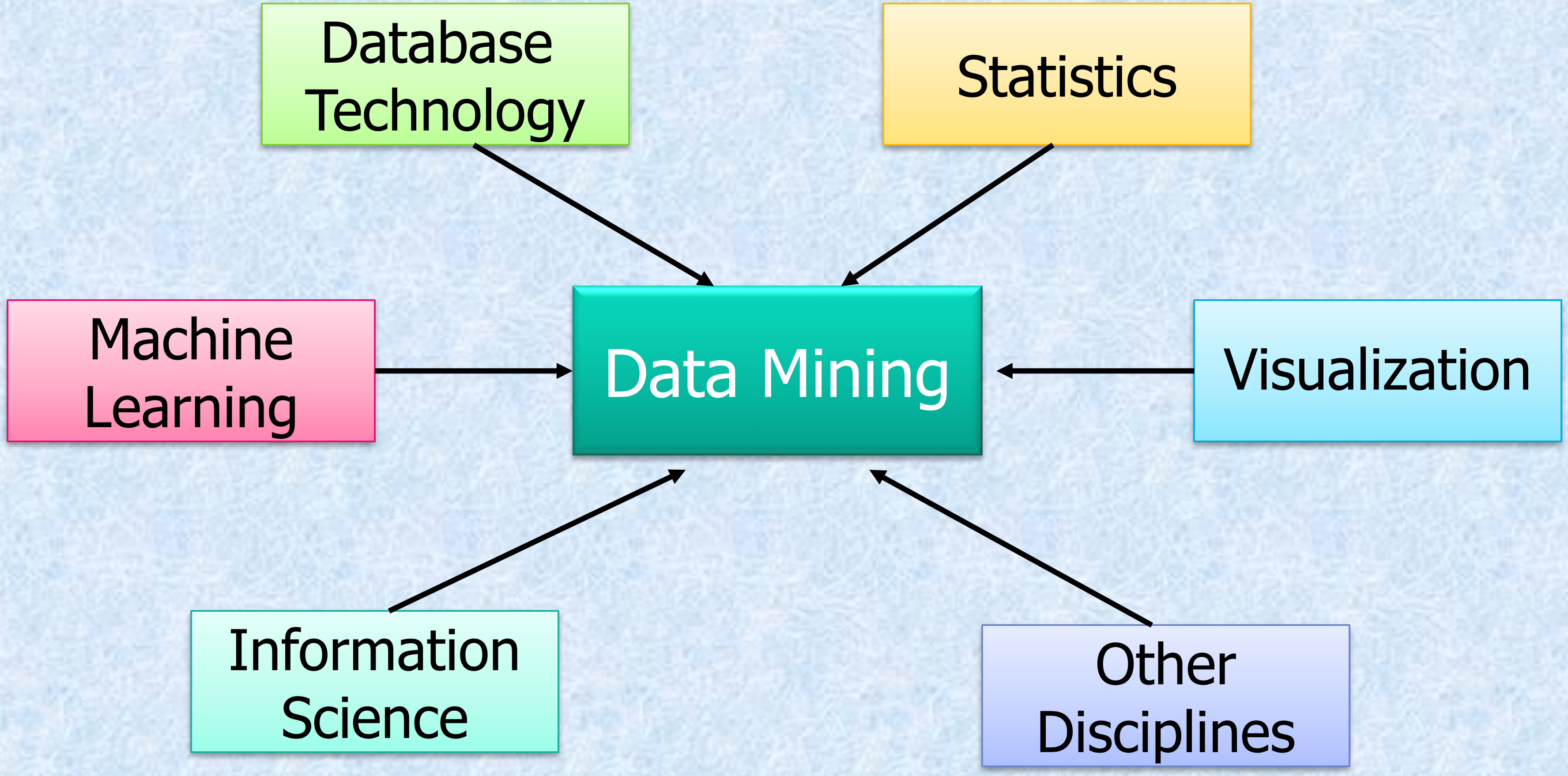
Are All the "Discovered" Patterns Interesting?

- **A data mining system/query may generate thousands of patterns, not all of them are interesting.**
  - Suggested approach: Human-centered, query-based, focused mining
- **Interestingness measures: A pattern is interesting 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**
- **Objective vs. subjective interestingness measures:**
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

Can We Find All and Only Interesting Patterns?

- **Find all the interesting patterns: Completeness**
  - Can a data mining system find all the interesting patterns?
- **Search for only interesting patterns: Optimization**
  - Can a data mining system find only the interesting patterns?
- **Approaches**
  - First general all the patterns and then filter out the uninteresting ones
  - Generate only the interesting patterns—mining query optimization

Data Mining: Confluence of Multiple Disciplines



Data Mining: Classification Schemes

- **General functionality**
  - Descriptive data mining
  - Predictive data mining
- **Different views, different classifications**
  - Kinds of databases to be mined
  - Kinds of knowledge to be discovered
  - Kinds of techniques utilized
  - Kinds of applications adapted

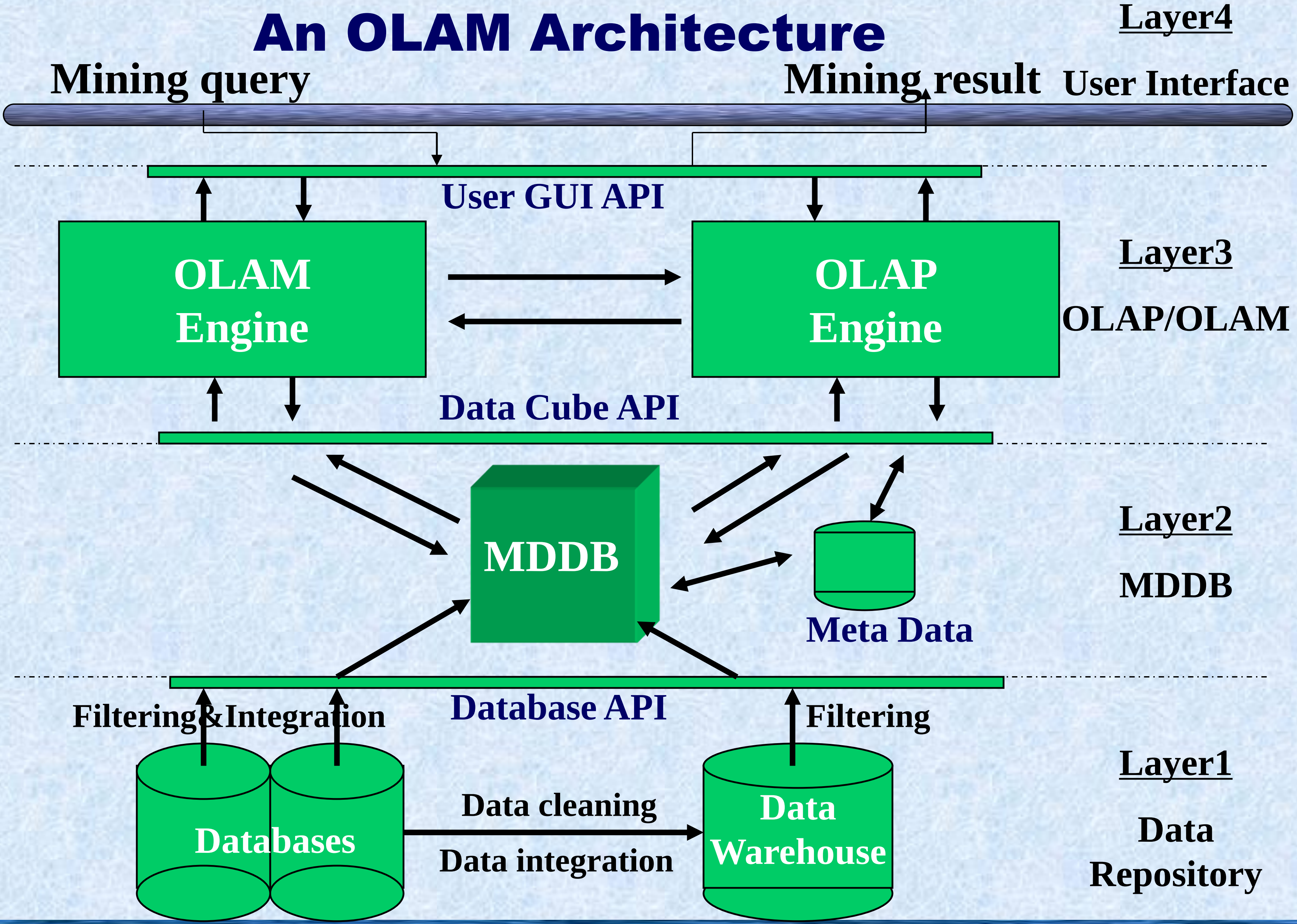
A Multi-Dimensional View of DM Classification

- **Databases to be mined**
  - Relational, transactional, object-oriented, object-relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW, etc.
- **Knowledge to be mined**
  - Characterization, discrimination, association, classification, clustering, trend, deviation and outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, neural network, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, DNA mining, stock market analysis, Web mining, Weblog analysis, etc.

OLAP Mining: An Integration of Data Mining and Data Warehousing

- **Data mining systems, DBMS, Data warehouse systems coupling**
  - No coupling, loose-coupling, semi-tight-coupling, tight-coupling
- **On-line analytical mining data**
  - Integration of mining and OLAP technologies
- **Interactive mining multi-level knowledge**
  - Necessity of mining knowledge and patterns at different levels of abstraction by drilling/rolling, pivoting, slicing/dicing, etc.
- **Integration of multiple mining functions**
  - Characterized classification, first clustering and then association

An OLAM Architecture



Major Issues in Data Mining (1)

- **Mining methodology and user interaction**
  - Mining different kinds of knowledge in databases
  - Interactive mining of knowledge at multiple levels of abstraction
  - Incorporation of background knowledge
  - Data mining query languages and ad-hoc data mining
  - Expression and visualization of data mining results
  - Handling noise and incomplete data
  - Pattern evaluation: the interestingness problem
- **Performance and scalability**
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed and incremental mining methods

Major Issues in Data Mining (2)

- **Issues relating to the diversity of data types**
  - Handling relational and complex types of data
  - Mining information from heterogeneous databases and global information systems (WWW)
- **Issues related to applications and social impacts**
  - Application of discovered knowledge
    - ✓ Domain-specific data mining tools
    - ✓ Intelligent query answering
    - ✓ Process control and decision making
  - Integration of the discovered knowledge with existing knowledge: A knowledge fusion problem
  - Protection of data security, integrity, and privacy

Summary

- Data mining: discovering interesting patterns from large amounts 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 information repositories
- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.
- Classification of data mining systems
- Major issues in data mining

References

- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.(Including Course Materials)
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.
- T. Imielinski and H. Mannila. A database perspective on knowledge discovery. Communications of ACM, 39:58-64, 1996.
- G. Piatetsky-Shapiro, U. Fayyad, and P. Smith. From data mining to knowledge discovery: An overview. In U.M. Fayyad, et al. (eds.), Advances in Knowledge Discovery and Data Mining, 1-35. AAAI/MIT Press, 1996.

Thank you !!!