

Data Mining and Warehousing

Chapter One

Introduction

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!
- “Necessity is the mother of invention”—Data mining—Automated analysis of massive data sets

Why Data Mining?

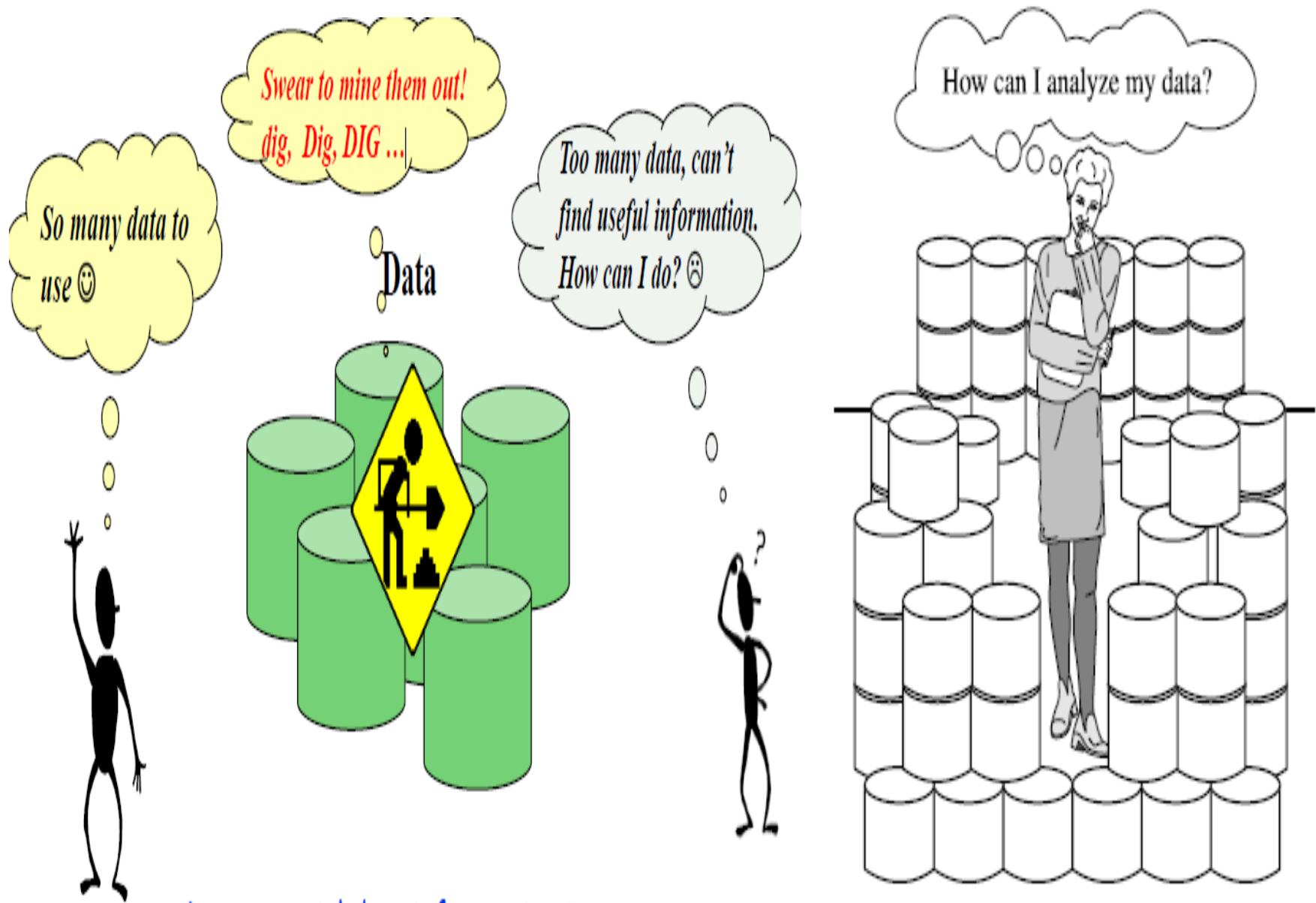
Moving toward the Information Age

- *“We are living in the information age” is a popular saying; however, we are actually living in the data age.*
- Terabytes or petabytes of data transfer into our computer networks, the World Wide Web, and various data storage devices every day from business, society, science and engineering, medicine, and almost every other aspect of daily life.
- Global backbone telecommunication networks carry tens of petabytes or more of data traffic every day.
- ✓ The explosive growing, widely available, and huge body of data makes our time truly the data age.
- ✓ Powerful and versatile tools are seriously needed to automatically uncover valuable information from the tremendous amounts of data and to transform such data into organized knowledge.
- This necessity has led to the birth of data mining.

Why Data Mining?

Data Mining as the Evolution of Information Technology

- Data mining can be viewed as a result of the natural evolution of information technology.
- The fast-growing, tremendous amount of data, collected and stored in large and numerous data repositories, has far exceeded our human ability for comprehension without powerful tools.
- As a result, data collected in large data repositories become “data tombs” data archives that are seldom visited.
- Consequently, important decisions are often made based not on the information-rich data stored in data repositories, because the decision maker does not have the tools to extract the valuable knowledge embedded in the vast amounts of data.
- The widening gap between data and information calls for the systematic development of data mining tools that can turn data tombs into “golden nuggets” of knowledge.



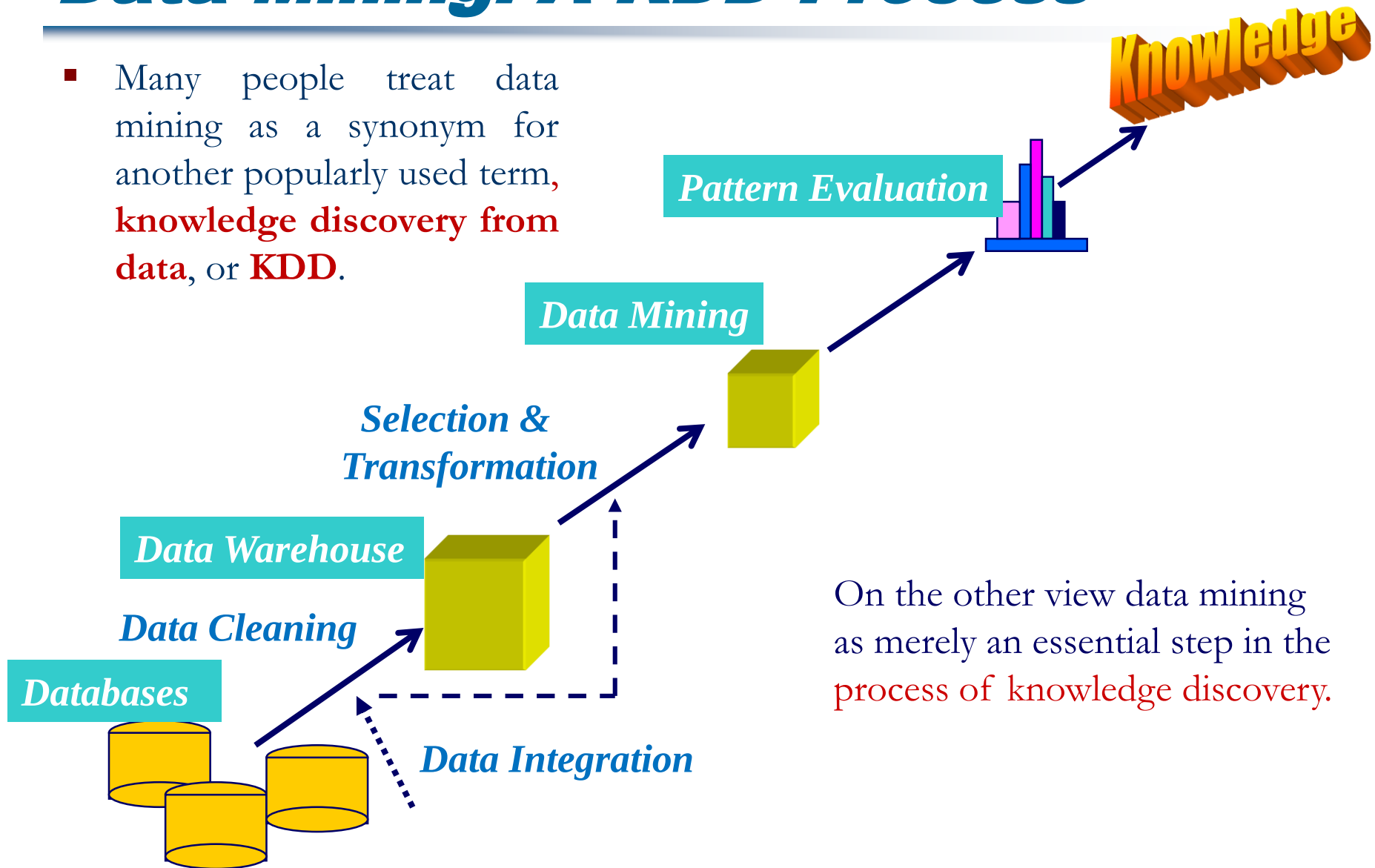
We are data rich, but information poor

What Is Data Mining?

- Data mining is extraction of **interesting** (non-trivial, implicit, previously unknown and potentially useful) **patterns or knowledge** from huge amount of data.
- Data mining: a misnomer?
- To refer to the **mining of gold** from **rocks or sand**, we say **gold mining** instead of **rock or sand mining**.
- Analogously, **data mining** should have been more appropriately named “**knowledge mining from data**,” which is unfortunately somewhat long.
- However, the shorter term, **knowledge mining** may not reflect the emphasis on mining from large amounts of data.
- In addition, many other terms have a similar meaning to data mining for example, **knowledge mining from data**, **knowledge extraction**, **data/pattern analysis**, **data archaeology**, and **data dredging**.

Data Mining: A KDD Process

- Many people treat data mining as a synonym for another popularly used term, **knowledge discovery from data**, or **KDD**.



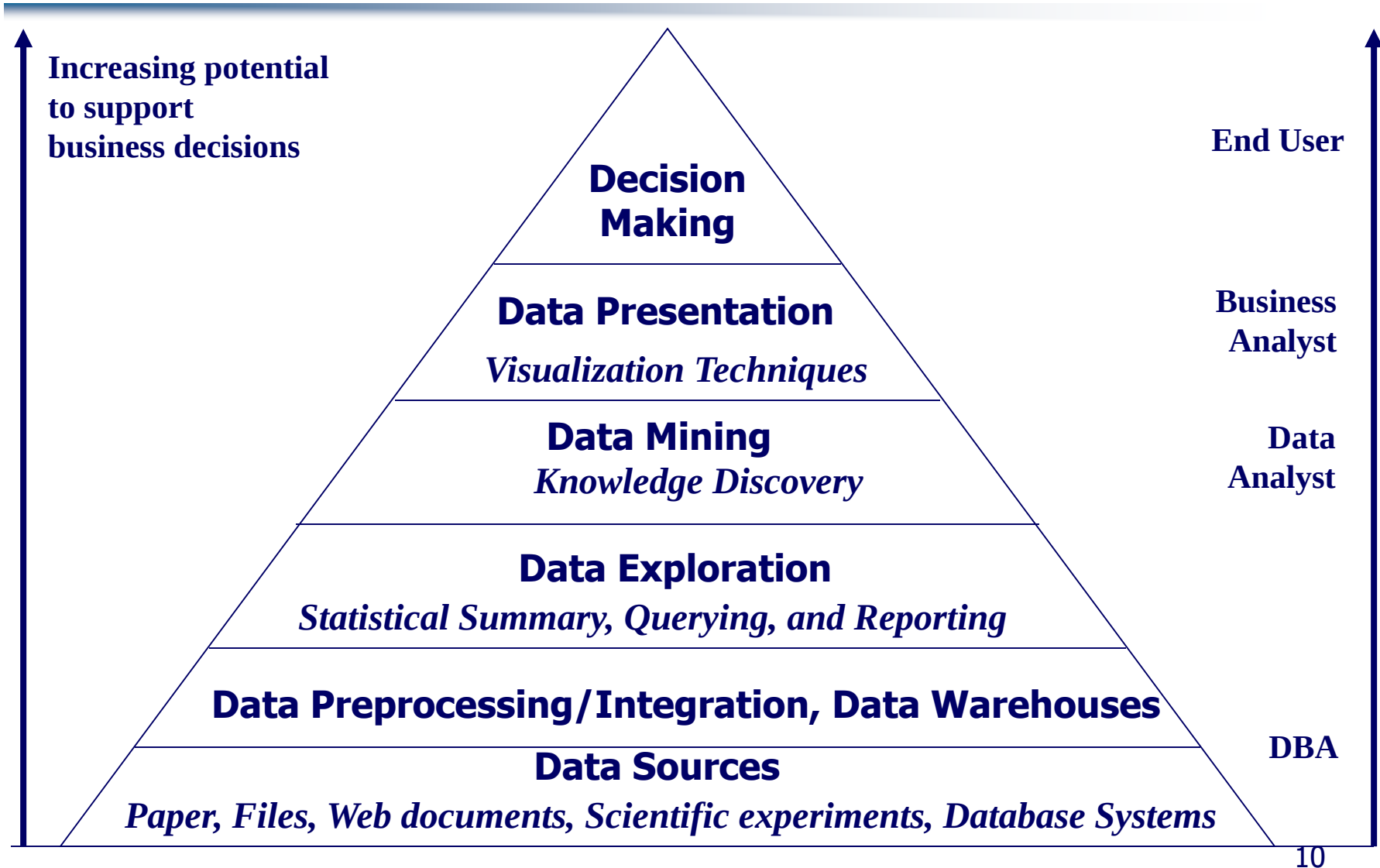
KDD Process

- The knowledge discovery process is an iterative sequence of the following steps:
 - ✓ **Data cleaning** (remove noise and inconsistent data)
 - ✓ **Data integration** (multiple data sources maybe combined)
 - ✓ **Data selection** (data relevant to the analysis task are retrieved from database)
 - ✓ **Data transformation** (consolidated into forms appropriate for mining)
 - ✓ **Data mining** (an essential process where intelligent methods are applied to extract data patterns)
 - ✓ **Pattern evaluation** (identify the truly interesting patterns)
 - ✓ **Knowledge presentation** (mined knowledge is presented to the user with representation techniques)
- **Data mining** is the *process* of discovering interesting patterns and knowledge from *large* amounts of data.
- The data sources can include databases, data warehouses, the Web, other information repositories.

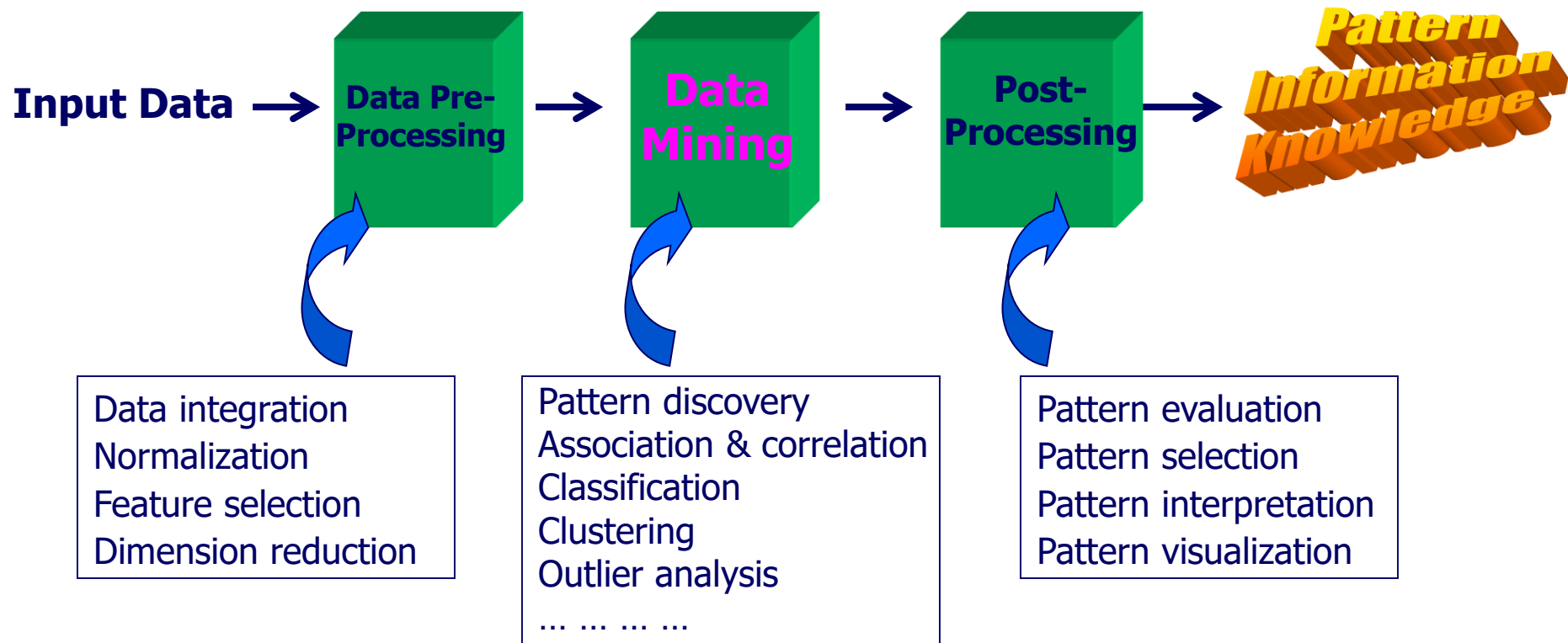
What is Business Intelligence (BI)?

- Business Intelligence is the process, technologies and tools needed to turn
 - ✓ data into information,
 - ✓ information into knowledge and
 - ✓ knowledge into plans that drive profitable business action
- BI comprises data warehousing, business analytic tools, and content/knowledge management

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

Database Processing vs. Data Mining Processing

	Database	Data mining	Comments
Query	Well defined Structured Query Language	• Poorly defined • No precise query language	The data miner might not know what he exactly wants to see
Data	Operational data	Non-Operational data	The data have been cleansed and modified to better support the mining process
Output	Precise and Subset of database	Not a subset of database	The output is some hidden useful patterns & knowledge in the database

Query Examples

❖ Database

- Find all credit applicants with first name 'Alex'.
- Identify customers who have purchased more than Birr 10,000 in the last month.
- Find all customers who have purchased **Bread**

❖ Data Mining

- Find all credit applicants who have no credit risks.
(classification)
- Identify customers with similar buying habits. (Clustering)
- Find all items which are frequently purchased with **Bread**.
(association rules)

Why the focus shifts to “Knowledge”

- We are living in dynamic/complex environment; an environment which is characterized by:
 - **Competitors**
 - Very strong competition
 - **Market**
 - Volatility of the market
 - The business landscape is changing rapidly and non-linearly
 - **Customers/Consumers**
 - Customers reaches to the level of prosumers
 - Prosumer are more educated consumer, who provide feedback regarding products/services they need
 - **Professionals**
 - The high turnover rate of professionals
 - Diminishing individual experience

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.
 - Descriptive vs. predictive data mining
 - Multiple/integrated functions and mining at multiple levels
- **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

What kind of patterns can be mined?

- We have observed various types of data and information repositories on which data mining can be performed.
- Let us now examine the kinds of patterns that can be mined.
- There are a number of ***data mining functionalities***. These include
 - ✓ Characterization and discrimination; the mining of frequent patterns, associations, and correlations;
 - ✓ Classification and regression;
 - ✓ Clustering analysis; and outlier analysis.
- Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks. In general, such tasks can be classified into two categories: **descriptive** and **predictive**.
- **Descriptive mining tasks** characterize properties of the data in a target data set.
- **Predictive mining tasks** perform induction on the current data in order to make predictions.

Data Mining Functionalities

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 Functionalities

Classification and Prediction

- 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 Functionalities

Cluster Analysis and Outlier Analysis

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 and minimizing interclass similarity

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

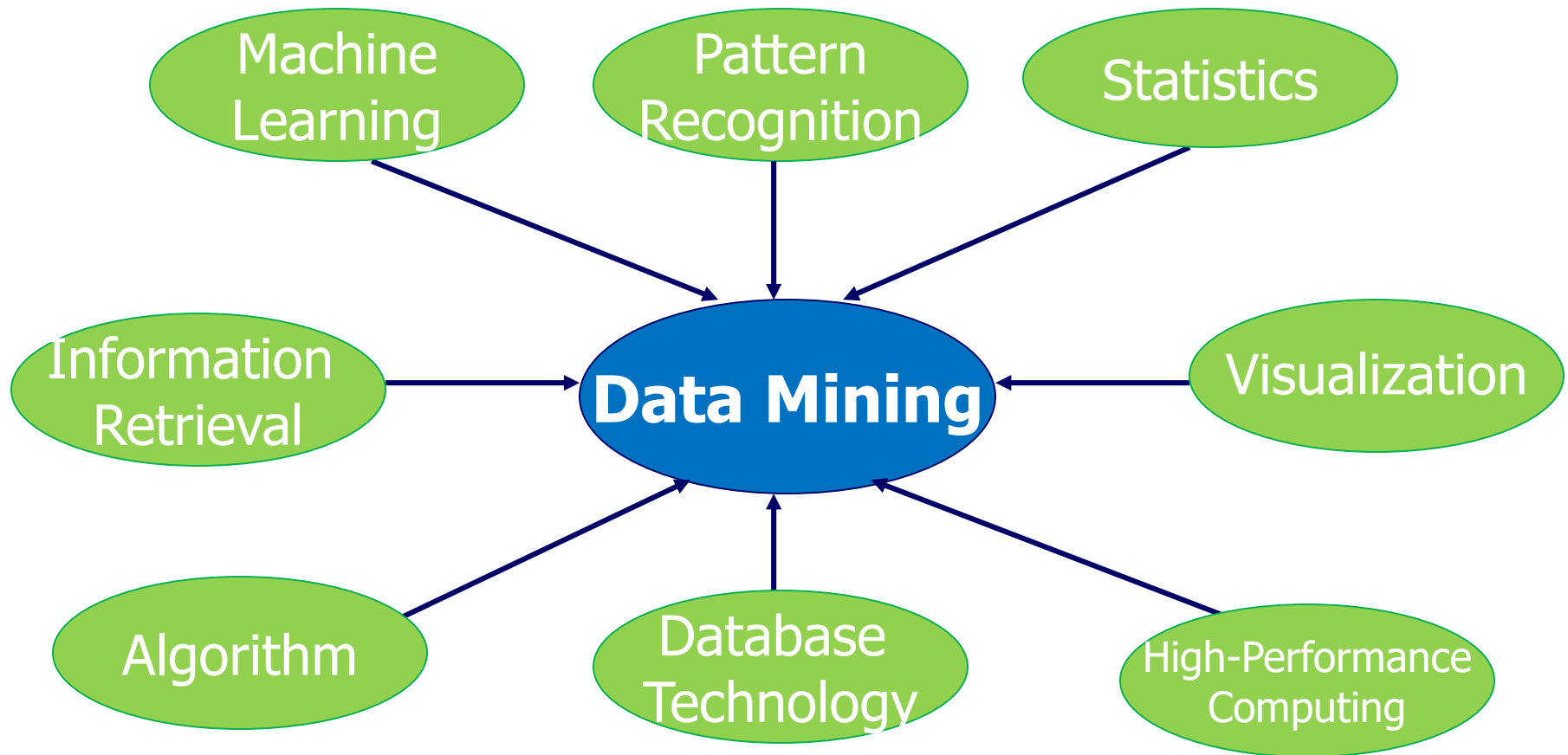
Are All of the Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting. *“What makes a pattern interesting? Can a data mining system generate all of the interesting patterns? Or, Can the system generate only the interesting ones?”*
- 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
 - validates some hypothesis that a user seeks to confirm
- An interesting measure represents *knowledge* !

Are All of the Patterns Interesting?

- **Objective measures:** Based on statistics and structures of patterns, e.g., support, confidence, etc. (Rules that do not satisfy a threshold are considered uninteresting.)
- **Subjective measures:** Reflect the needs and interests of a particular user. Based on user's belief in the data.
- The second question “*Can a data mining system generate all of the interesting patterns?*” refers to the **completeness** of a data mining algorithm.
- Finally, the third question—“*Can a data mining system generate only interesting patterns?*” is an optimization problem in data mining.
- Measures of pattern interestingness are essential for the efficient discovery of patterns by target users.
- Such measures can be used after the data mining step to rank the discovered patterns according to their interestingness, filtering out the uninteresting ones.
- Objective and subjective measures need to be combined.

Data Mining: Confluence of Multiple Disciplines



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
- Fraud detection/Network intrusion detection
 - ✓ Which types of transactions are likely to be fraudulent, given the demographics and transactional history of a particular customer?

Major Issues in Data Mining

The major issues in data mining classified into five groups:

- **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

Data Mining System Classification

- A data mining system can be classified according to the following criteria:
 - ✓ Database Technology
 - ✓ Statistics
 - ✓ Machine Learning
 - ✓ Information Science
 - ✓ Visualization
- Apart from these, a data mining system can also be classified based on the kind of (a) databases mined, (b) knowledge mined, (c) techniques utilized, and (d) applications adapted.

Data Mining System Classification

Classification Based on the Databases Mined

- We can classify a data mining system according to the kind of databases mined. Database system can be classified according to different criteria such as data models, types of data, etc.
- And the data mining system can be classified accordingly.
- For example, if we classify a database according to the data model, then we may have a relational, transactional, object-relational, or data warehouse mining system.

Classification Based on the Techniques Utilized

- We can classify a data mining system according to the kind of techniques used. We can describe these techniques according to the degree of user interaction involved or the methods of analysis employed.

Data Mining System Classification

Classification Based on the Kind of Knowledge Mined

- We can classify a data mining system according to the kind of knowledge mined. It means the data mining system is classified on the basis of functionalities such as:
- Characterization
- Discrimination
- Association and Correlation Analysis
- Classification
- Prediction
- Clustering
- Outlier Analysis
- Evolution Analysis

Data Mining System Classification

Classification Based on the Applications Adapted

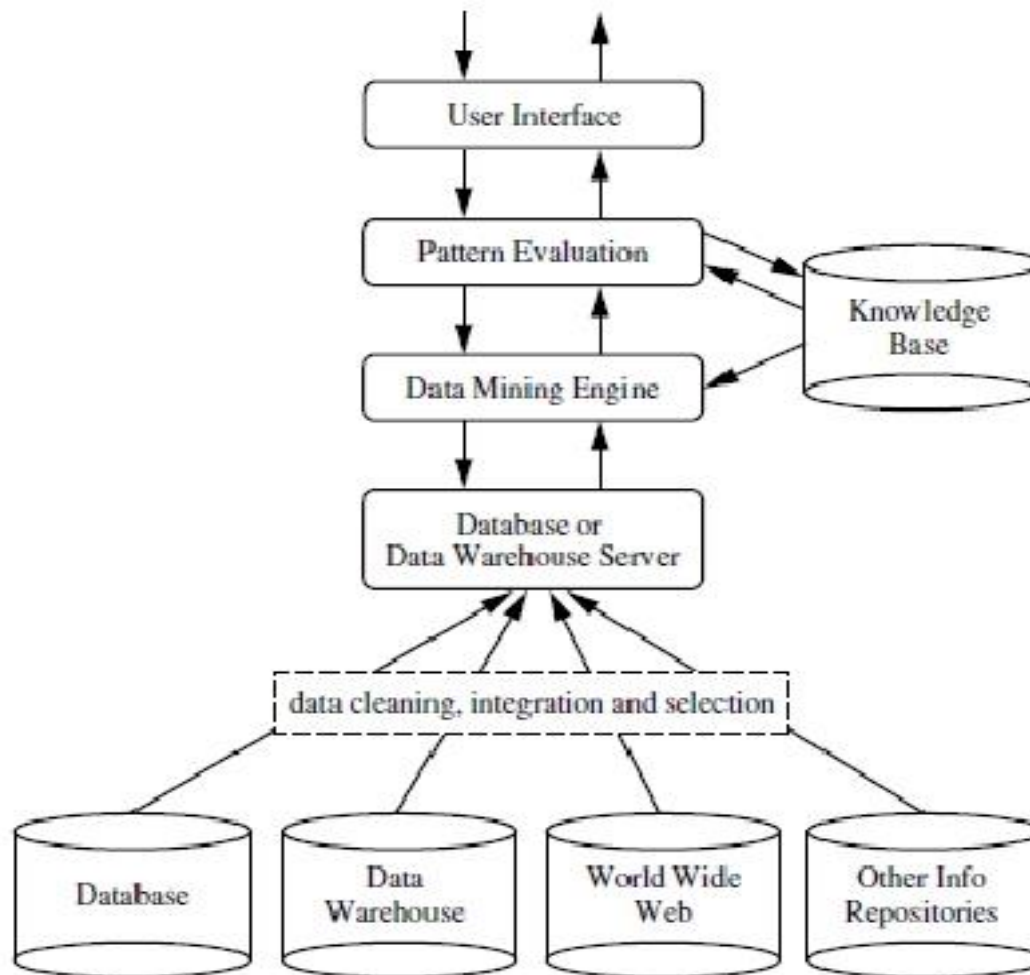
- We can classify a data mining system according to the applications adapted. These applications are as follows:
 - ✓ Finance
 - ✓ Telecommunications
 - ✓ DNA
 - ✓ Stock Markets
 - ✓ E-mail

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

Architecture of Data Mining

- A typical data mining system may have the following major components



Architecture of Data Mining

Knowledge Base:

- This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns.
- Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction.
- Knowledge such as user beliefs, which can be used to assess a pattern's interestingness based on its unexpectedness, may also be included.
- Other examples of domain knowledge are additional interestingness constraints or thresholds, and metadata (e.g., describing data from multiple heterogeneous sources).

Architecture of Data Mining

Data Mining Engine:

- This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.

Pattern Evaluation Module:

- This component typically employs interestingness measures interacts with the data mining modules so as to focus the search toward interesting patterns.
- It may use interestingness thresholds to filter out discovered patterns. Alternatively, the pattern evaluation module may be integrated with the mining module, depending on the implementation of the datamining method used.

Architecture of Data Mining

User interface:

- This module communicates between **users** and the **data mining system**, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results.
- In addition, this component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.