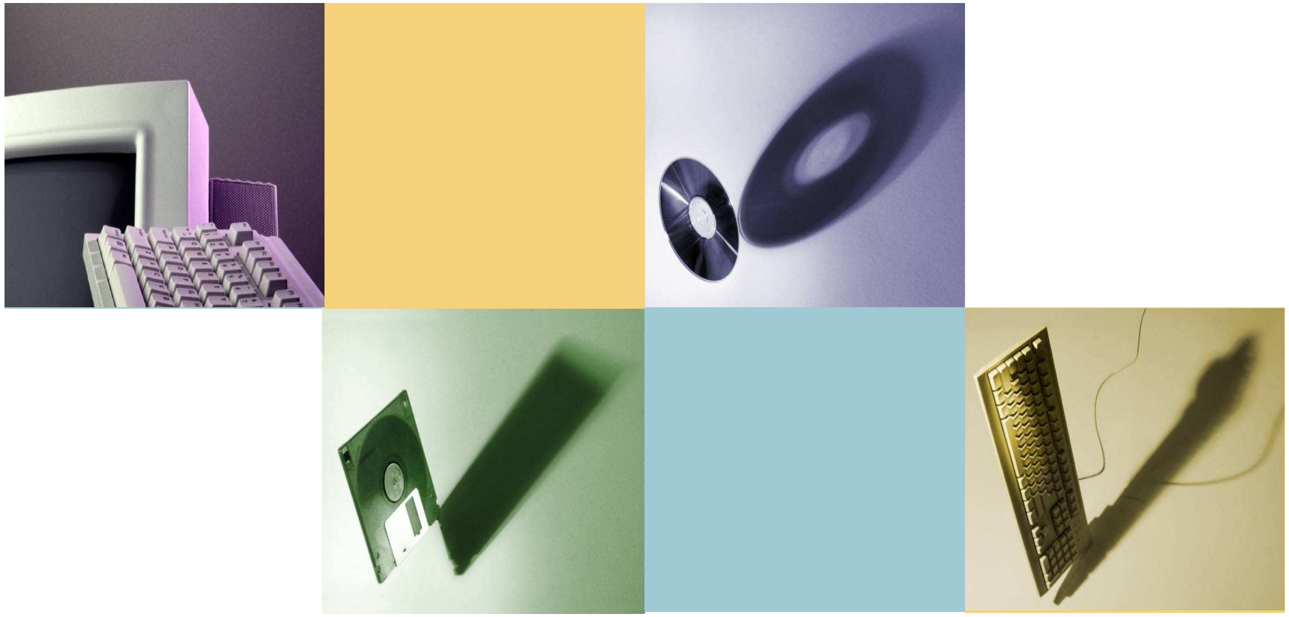
Introduction to Data Mining



Outline

- What motivated data mining?
- What is data mining?
- Data mining – on what kind of data?
- Data mining functionalities
- Are all of the patterns interesting?
- Classification of data mining systems
- Major issues in data mining
- Summary

What Motivated Data Mining?

- 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 (OLAP).
  - Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases.

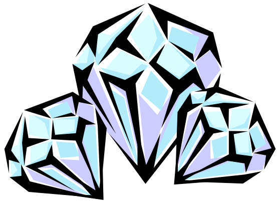
Evolution of Database Technology

- **1960s and earlier:**
  - Data collection, database creation
- **1970s – early 1980s:**
  - Hierarchical and network database systems
  - Relational database systems, SQL language.
- **Mid 1980s – present:**
  - Advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- **Late 1980s – present:**
  - Data warehousing and data mining
- **1990s – present:**
  - Web database, XML-based Database Systems
  - Web mining
- **2000 – …:**
  - New Generation of Integrated Information Systems

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:
  - Data mining: a misnomer?
  - Knowledge discovery(mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- What is not data mining?



  - (Deductive) query processing.
  - Expert systems or small statistical programs

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.

Market Analysis and Management (1)

- Data sources for analysis:
  - Credit card, loyalty cards, discount coupons, customer complaint calls, public lifestyle studies
- Target marketing:
  - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
- Determine customer purchasing patterns over time
  - Conversion of single to a joint bank account: marriage, etc.
- Cross-market analysis
  - Associations/co-relations between product sales.
  - Prediction based on the association information.

Market Analysis and Management (2)

- Customer profiling
  - What types of customers buy what products (clustering or classification).
- Identifying customer requirements
  - Identifying the best products for different customers.
  - Use prediction to find what factors will attract new customers.
- Provides summary information
  - Various multidimensional summary reports.
  - Statistical summary

Corporate Analysis and Risk Management

- Finance planning and asset evaluation
  - Cash flow analysis and prediction.
  - Cross-sectional and time series analysis (financial-ratio, trend analysis, etc).
- Resource planning
  - Summarize and compare the resources and spending.
- Competition
  - Monitor competitors and market directions.
  - Group customers into classes and a class-based pricing procedure.
  - Set pricing strategy in a highly competitive market.

Fraud Detection and Management (1)

- Applications
  - Widely used in health care, retail, credit card services, telecommunications (phone card fraud), etc.
- Approach
  - Use historical data to build models of fraudulent behavior and use data mining to help identify similar instances.
- Examples:
  - auto insurance: detect a group of people who stage accidents to collect on insurance.
  - money laundering: detect suspicious money transactions (US Treasury's Financial Crimes Enforcement Network).
  - medical insurance: detect professional patients and ring of doctors and ring of references.

Fraud Detection and Management (2)

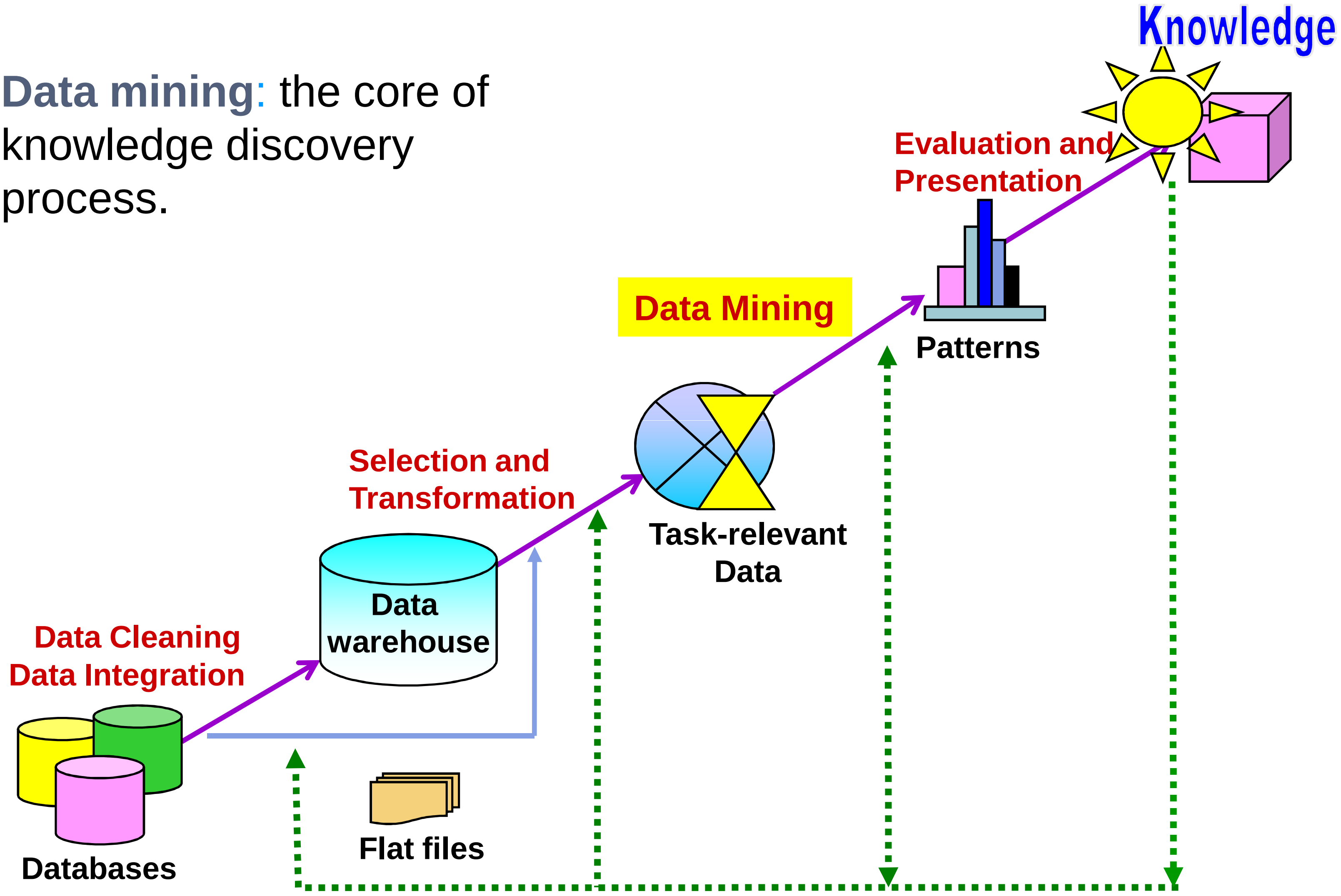
- Detecting inappropriate medical treatment
  - Australian Health Insurance Commission identifies that in many cases blanket screening tests were requested (save Australian $1m/yr).
- Detecting telephone fraud
  - Telephone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm.
  - British Telecom identified discrete groups of callers with frequent intra-group calls, especially mobile phones, and broke a multimillion dollar fraud.
- Retail
  - Analysts estimate that 38% of retail shrink is due to dishonest employees.

Other Applications

- Sports
  - IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage for New York Knicks and Miami Heat.
- Astronomy
  - JPL and the Palomar Observatory discovered 22 quasars with the help of data mining.
- Internet Web Surf-Aid
  - IBM Surf-Aid applies data mining algorithms to Web access logs for market-related pages to discover customer preference and behavior pages, analyzing effectiveness of Web marketing, improving Web site organization, etc.

Data Mining: A KDD Process

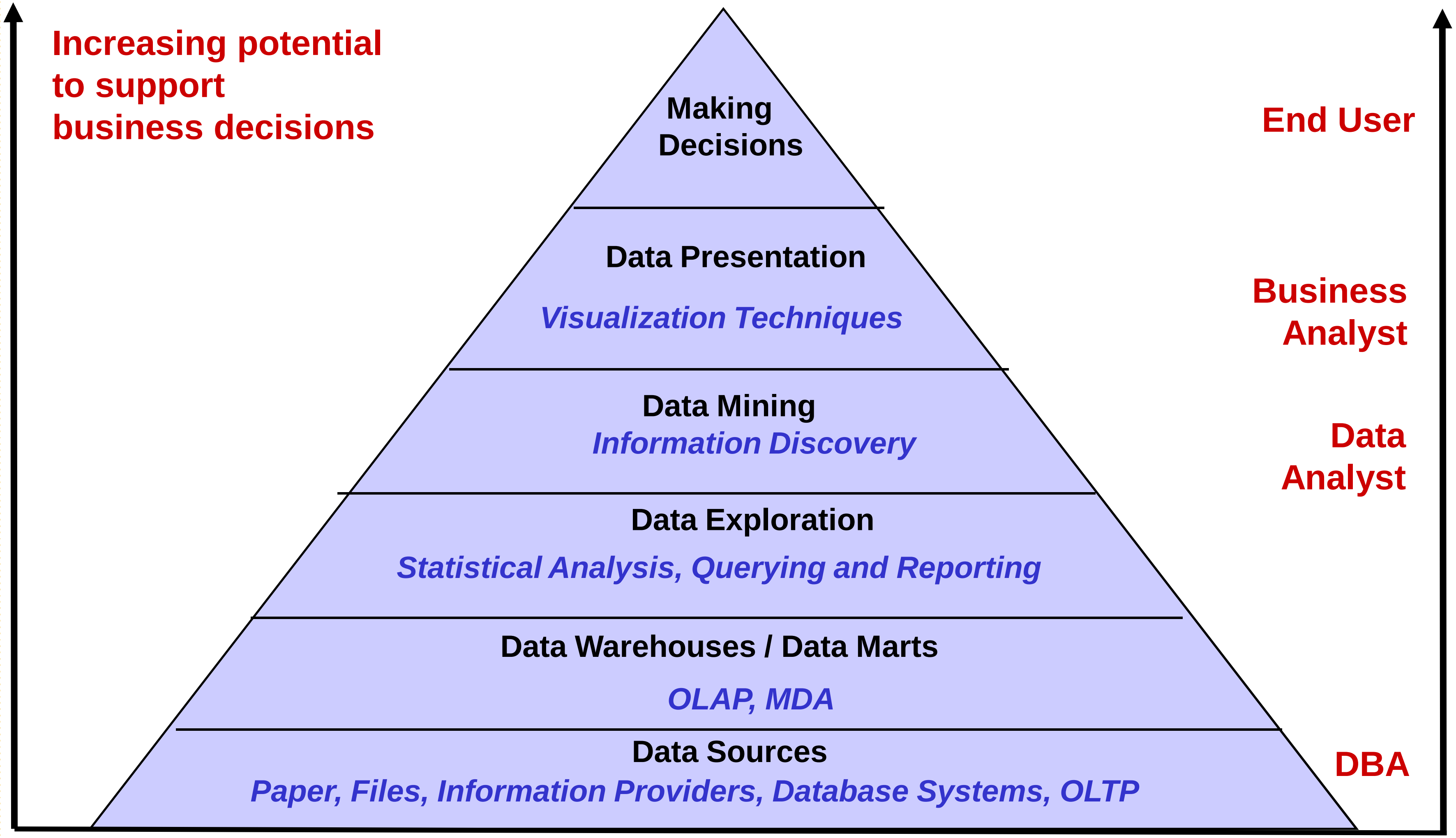
- **Data mining**: the core of knowledge discovery process.



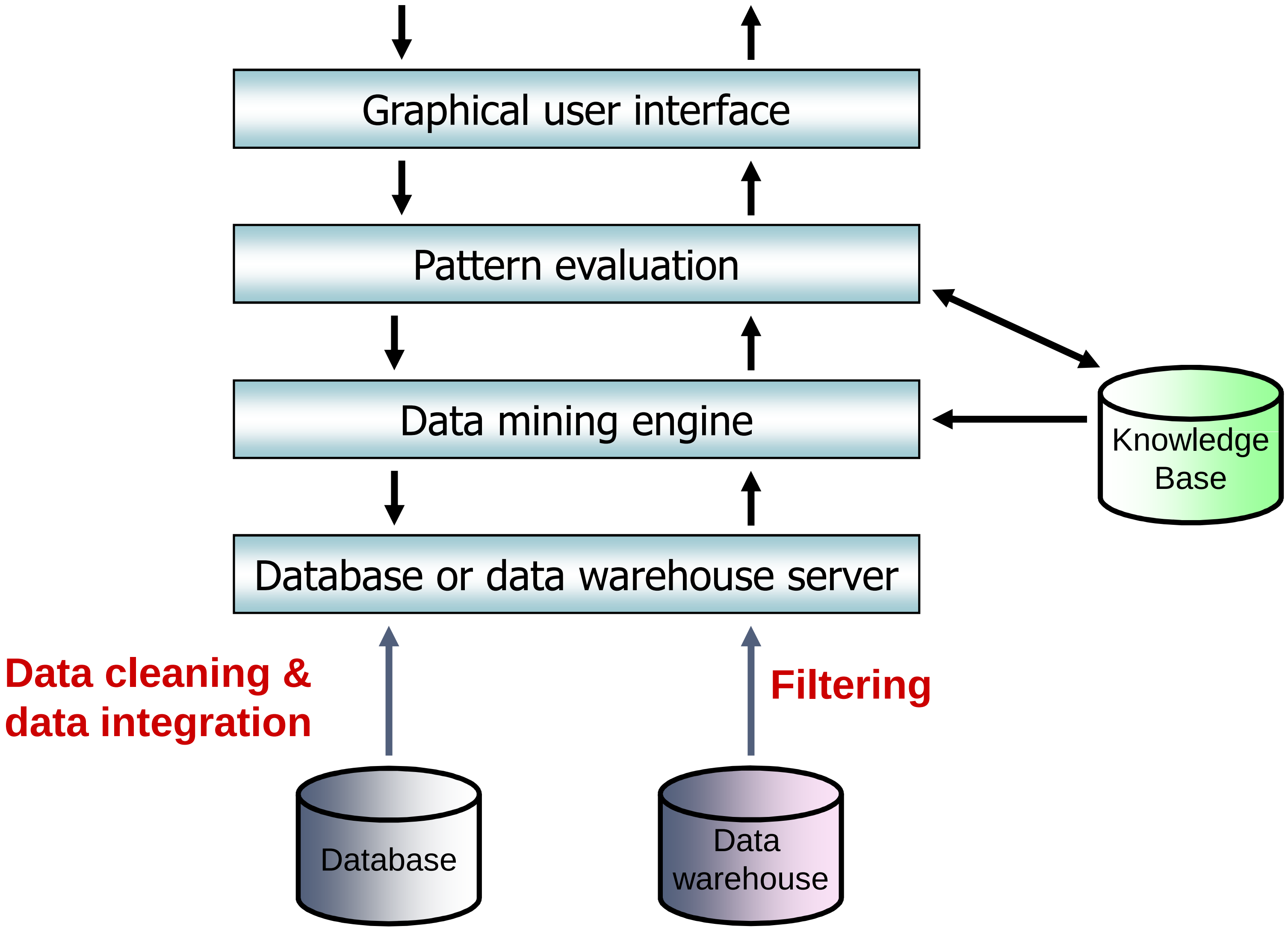
Steps of a KDD Process

- 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



Architecture: Typical Data Mining System



Graphical user interface

Pattern evaluation

Data mining engine

Database or data warehouse server

Knowledge Base

Data cleaning & data integration

Filtering

Database

Data warehouse

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
    - The WWW

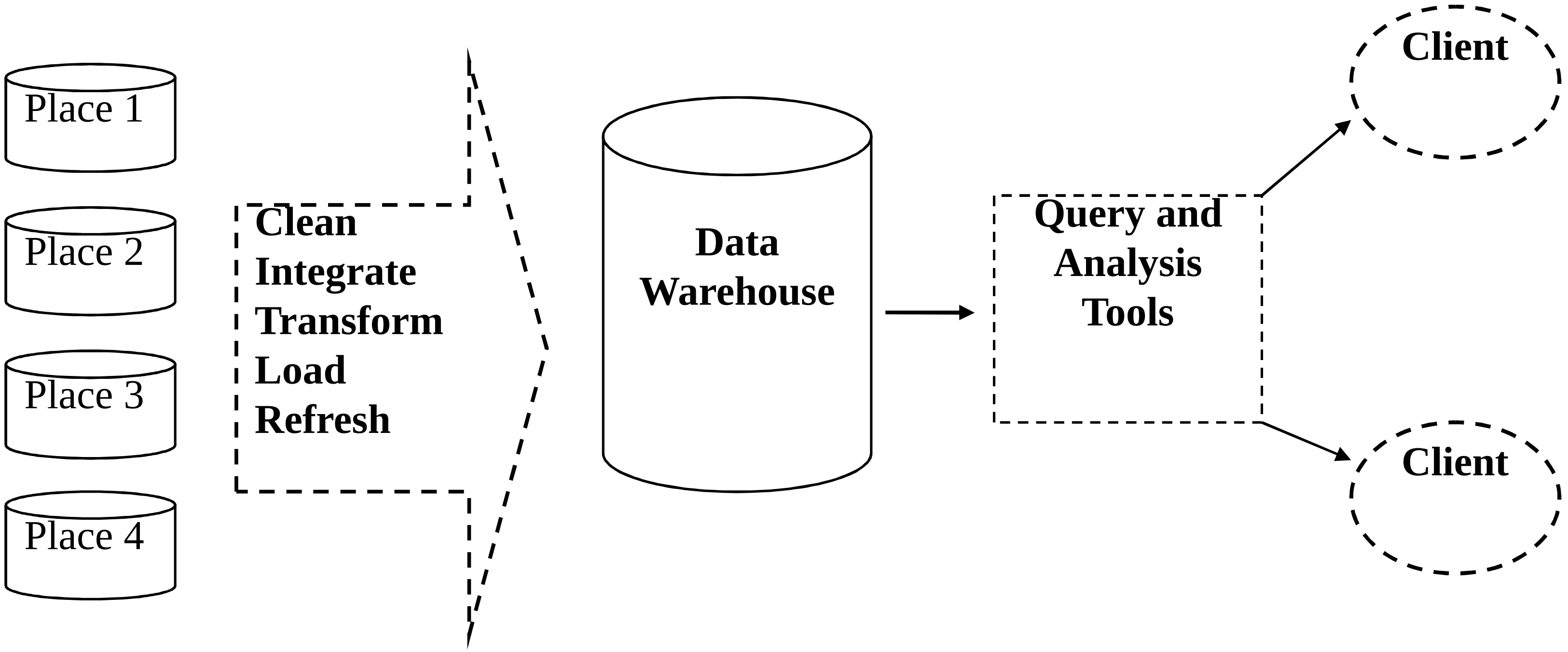
Relational Databases

Cust_ID	Name	Address	Age	Income	Category
123 ----	M.Kannan -------	123, south st, --------	34 --	34000 ------	2 -----

Data Warehouses

A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually resides at a single site.

Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.



Transactional Databases

A transactional database consists of a file where each record represents a transaction. A transaction typically includes a unique transaction identity number (trans_ID) and a list of the items making up the transaction.

Advanced Data and Information Systems and Advanced Applications

- ***Object-Relational Databases***

- A set of **variables** that describe the object (also called attributes)
- A set of **messages** that the object can use to communicate with other objects
- A set of **methods**, where each method holds the code to implement a message.

- ***Temporal Databases, Sequence Databases, and Time-Series Databases***

- Temporal database typically stores relational data that including time-related attributes.
- Data mining techniques can be used to find the characteristics of object, evolution or the trend of changes for objects in the database.

- ***Spatial Databases and Spatiotemporal Databases***

- Spatial database contain spatial-related information
- Geographic database, very large-scale integration or computed-aided design databases, and medical and satellite image databases.
- Geographic databases are commonly used in vehicle navigation and dispatching systems.

Text Databases and Multimedia Databases

- Text databases are databases that contain word descriptions for objects
- These word descriptions are usually not simple keywords
- By mining text data, one may uncover general and concise descriptions of the text documents, keyword or content associations
- Multimedia databases store image, audio, and video data
- Content-based retrieval, voice-mail systems, video-on-demand systems, the World Wide Web, and speech-based user interfaces recognize spoken commands

Heterogeneous Databases and Legacy Databases

- A heterogeneous database consists of a set of interconnected, autonomous component database

Data Streams

- Data flow in and out of an observation platform (or window) dynamically
- Power supply, network traffic, stock exchange, telecommunication, web click streams video surveillance, and weather or environment monitoring

- ***The World Wide Web***
- Capturing user access patterns in a distributed information environment is called Web usage mining (or Weblog mining).
- Automated Web page clustering and classification help group and arrange web pages in a multidimensional manner based on their contents.
- Web community analysis helps to identify hidden Web social networks and communities and observe their evolution.