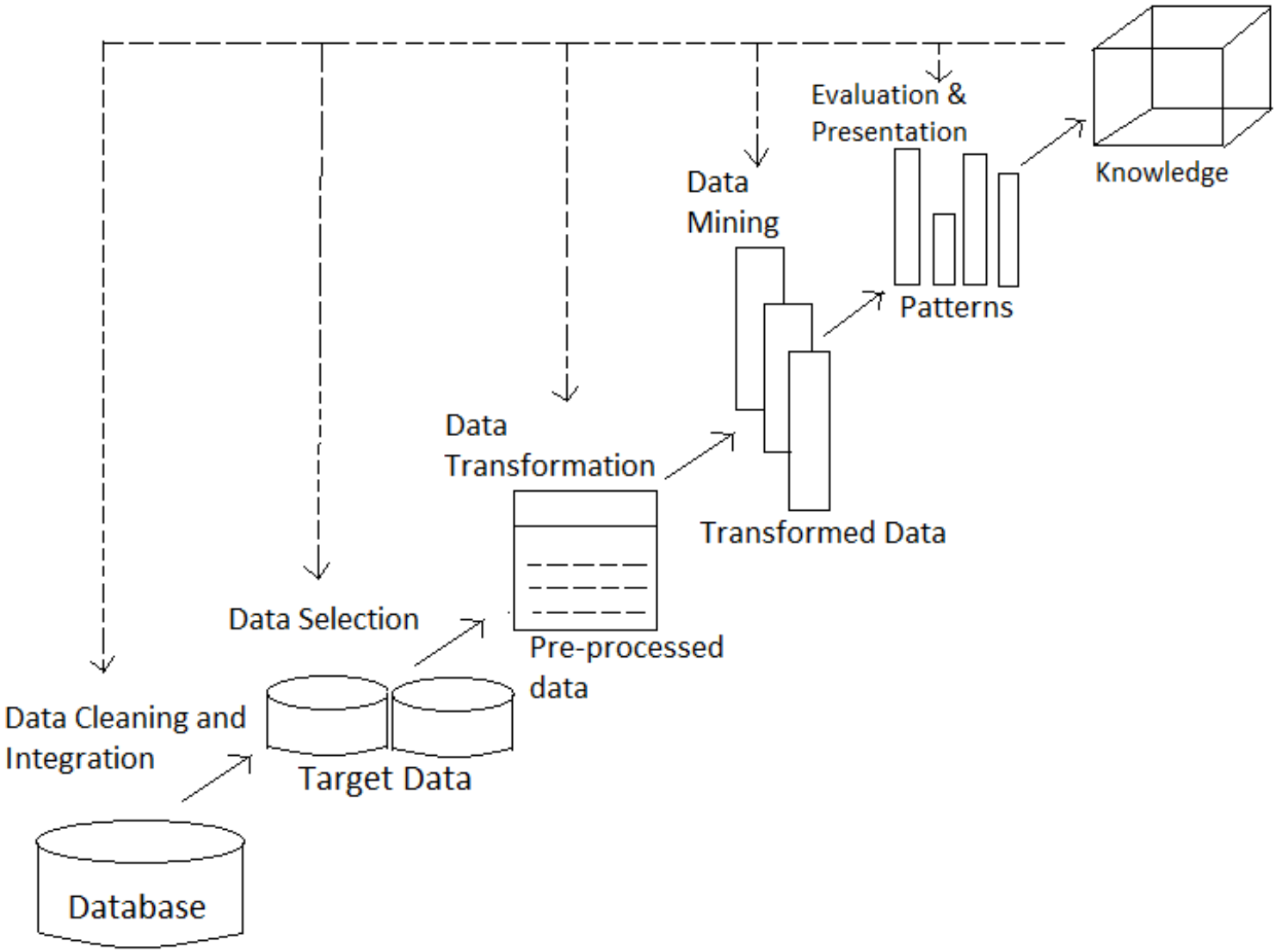
Introduction of Data Mining

Data Mining: Why?

- 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

Knowledge Discovery from Data (KDD) Process



1. **Data cleaning** (to remove noise and inconsistent data)

2. Data integration (where multiple data sources may be combined)

3. Data selection (where data relevant to the analysis task are retrieved from the database)

4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)

5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)

6. Pattern evaluation (to identify the truly interesting) patterns representing knowledge based on some interestingness measures

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user).

Types 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 Functionalities

- Concept/Class Description: Characterization and Discrimination
- Classification
- Clustering
- Mining Frequent Patterns, Associations, and Correlations
- Outlier Analysis
- Evolution Analysis

Major Issues in Data Mining

- **Mining methodology and user interaction issues:**
  - 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
  - Presentation and visualization of data mining results
  - Handling noisy or incomplete data
  - Pattern evaluation

- **Performance issues:**
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed, and incremental mining algorithms
- **Issues relating to the diversity of database types:**
  - Handling of relational and complex types of data
  - Mining information from heterogeneous databases and global information systems

Applications of Data Mining

- Financial Analysis
- Retail Industry
- Health care
- Telecommunication Industry
- Higher Education
- Criminal Investigation
- Intrusion Detection
- E-Commerce
- Research Analysis

- Watch out: Is everything “data mining”?
  - Simple search and query processing
  - (Deductive) expert systems

-

Reference

1. J. Han, M. Kamber - “Data Mining Concepts and Techniques”, Morgan Kaufmann, 3rd Edition.

Thank You