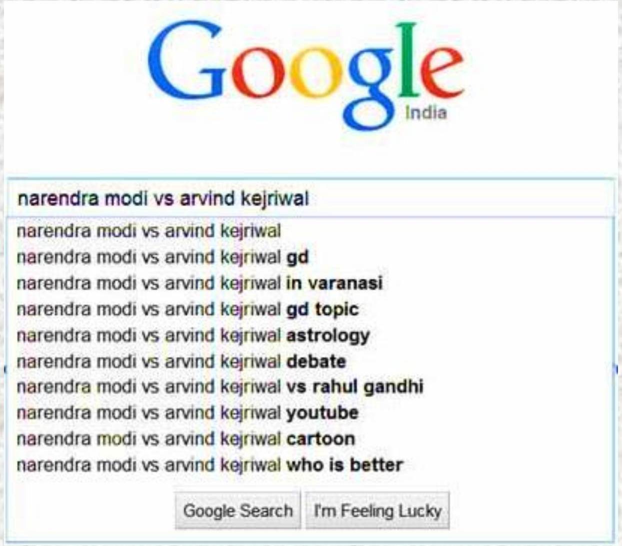
Chapter -2

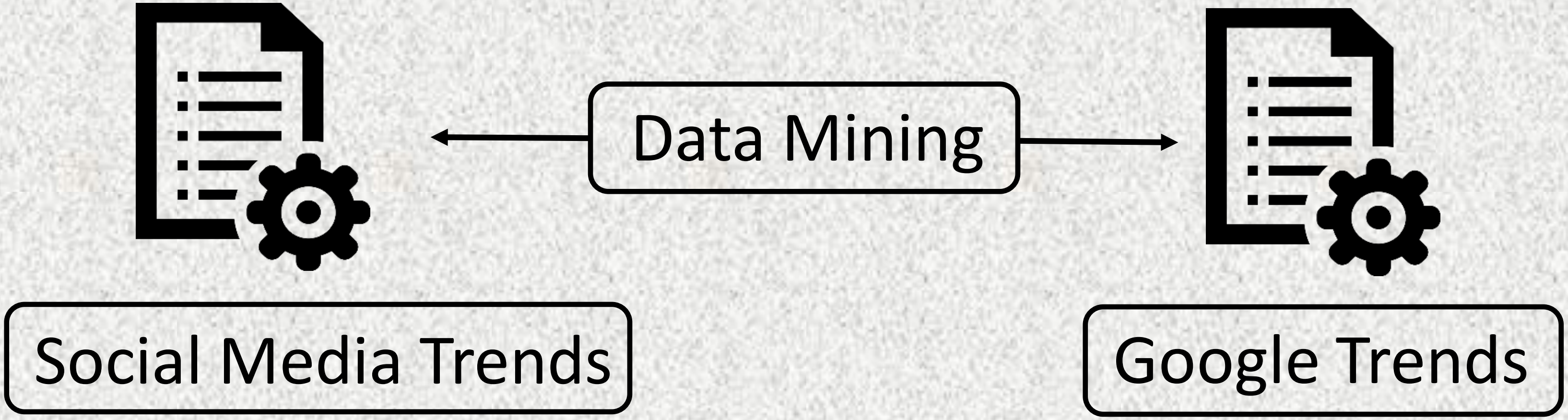
Introduction to Data Mining

Motivation : Why Data Mining?



narendra modi vs arvind kejriwal

narendra modi vs arvind kejriwal
narendra modi vs arvind kejriwal **gd**
narendra modi vs arvind kejriwal **in varanasi**
narendra modi vs arvind kejriwal **gd topic**
narendra modi vs arvind kejriwal **astrology**
narendra modi vs arvind kejriwal **debate**
narendra modi vs arvind kejriwal **vs rahul gandhi**
narendra modi vs arvind kejriwal **youtube**
narendra modi vs arvind kejriwal **cartoon**
narendra modi vs arvind kejriwal **who is better**



Data Mining

Social Media Trends

Google Trends

Motivation : Why Data Mining ?

- The major reason for using data mining techniques is requirement of useful information and knowledge from huge amounts of data.
- The information and knowledge gained can be used in many applications such as business management, production control etc. ... Data collection and database creation.
- Data are any facts, numbers, images or text that can be processed by a computer.
  - ❖ Huge amounts of data are widely available.
  - ❖ Usage of bar codes of commercial products
  - ❖ Store membership/Customer rewards program
  - ❖ Computerization of office work
  - ❖ Advanced data collection tools
  - ❖ World Wide Web

Motivation : Why Data Mining?

“Necessity is the Mother of Invention”

Tremendous Amount of Data

Solution

“Data Mining”

Extraction of interesting Knowledge from data in large databases

“It has been estimated that the amount of **information** in the world **doubles** every **10** months.”

- There is a tremendous increase in the amount of data recorded and stored on digital media as well as individual sources.

Why Data Mining? (Cont..)

"We are drowning in data, but starving for knowledge!"
"Data rich but Information poor"

- Since the 1960's, database and information technology has been changed systematically from primitive file processing systems to powerful database systems.
- The research and development in database systems since the 1970's has led to the development of relational database systems.

Why Data Mining? (Cont..)

Years	Evolution
Since 1960's	Data collection, database creation, IMS (hierarchical database system by IBM) and network DBMS
1970s	Relational data model, relational DBMS implementation
1980s	RDBMS, advanced data models, application-oriented DBMS (spatial, scientific, engineering, etc.)
1990s	Data mining, data warehousing, multimedia databases, and web databases
2000s	Stream data management and mining, Social Networks (Facebook, etc.), web technology (XML) and global information systems
At Present	Heterogeneous database systems, big data

Every day data **grows exponentially**,
but these **all data** are really **important to us**??

What is Data Mining?

1. Data mining refers to extracting or “mining” knowledge from large amounts of data.
2. “Knowledge mining from data” or “Knowledge mining”
3. “Extract knowledge from large data or databases”
4. “Knowledge discovery from database (KDD)”

What is Data Mining? (Cont..)

- It is the **computational process** of **discovering patterns** in **large data sets** involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.

The overall goal of the data mining process is to **extract information from a large data sets** and **transform it into an understandable structure** for further use.

What is Data Mining? (Cont..)

Data → Knowledge → Action → Goal

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**)

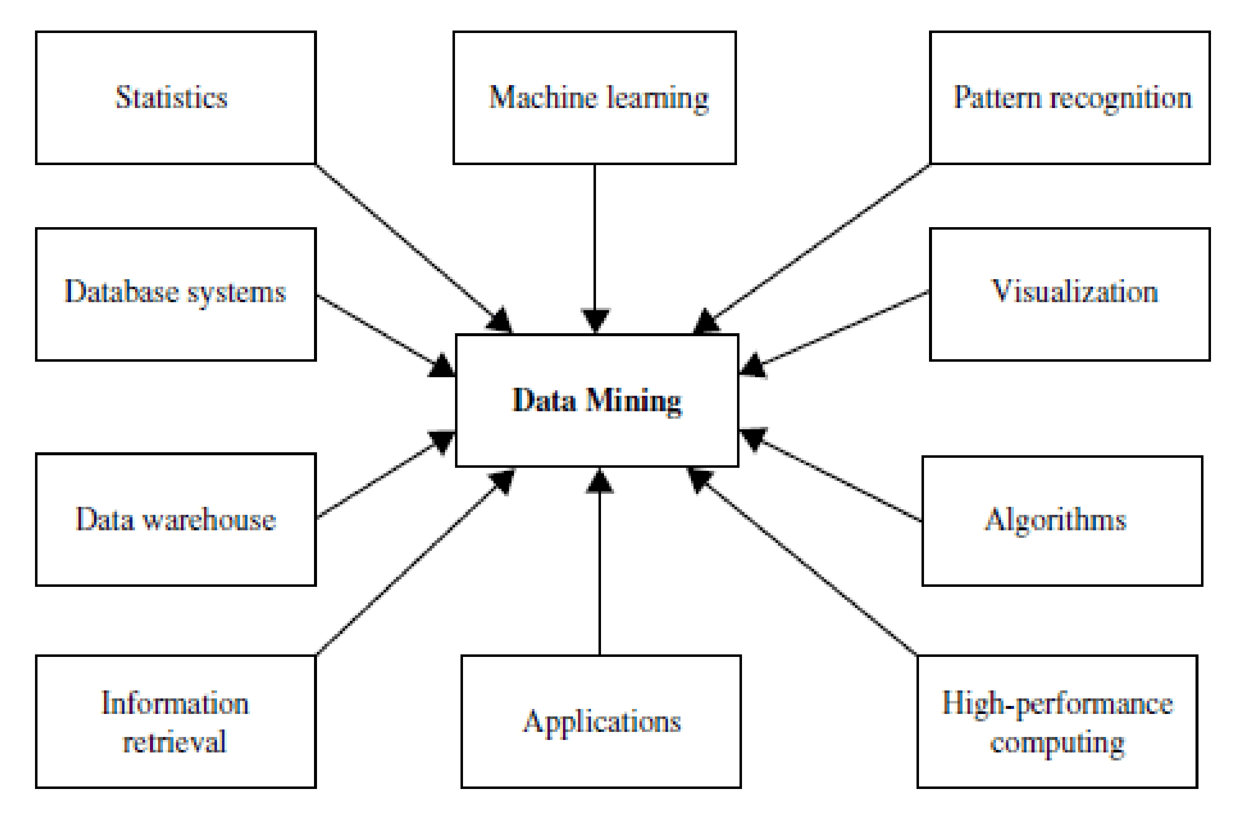
Road traffic (**data**) => Which road is likely to be congested? (**knowledge**) => Suggest better routes to drivers (**action**) => Save time and energy (**goal**)

Data Mining Functionalities

- Data mining tasks can be classified into two categories:
  1. **Descriptive**
  2. **Predictive**
- **Descriptive**
  - These tasks present the **general properties** of data stored in database.
  - The descriptive tasks are used to find out patterns in data.
  - **E.g.** : Cluster, correlation, trends etc.
- **Predictive**
  - These tasks **predict the value of one attribute on the bases of values of other attributes**.
  - **E.g.** : Customer/Product prediction at sales store

Domains of Data Mining Systems

- Data mining is an **interdisciplinary field**, joining of a set of disciplines, including database systems, statistics, machine learning, visualization and information science.



Classification of Data Mining Systems

- **Classification of data mining & Multi-Dimensional View of Data Mining are similar terms.**
- Classification of data mining based on..
  1. **Databases** to be mined
  2. **Knowledge** to be mined
  3. **Techniques/Methods** utilized
  4. **Application** adapted

Classification of Data Mining Systems

1. Classification according to the kinds of **databases** mined:
   - Classified **according to different criteria** (such as data models, or the types of data or applications involved), each of which **may require its own data mining technique**.
   - For instance, if classifying according to data models, we may have a relational, transactional, object-oriented, object-relational, or data warehouse mining system.
   - If classifying according to the **special data types**, we may have a **spatial, time-series, text or multimedia data** mining system or a **world-wide web** mining system.
   - Other system types include **heterogeneous data mining systems** and legacy data mining systems.

Classification of Data Mining Systems (Cont..)

2. Classification according to the kinds of **knowledge** mined:
    - Based on data mining functionalities,
        - Characterization
        - Discrimination
        - Association
        - Correlation analysis
        - Classification & prediction
        - Clustering
        - Outlier analysis

Classification of Data Mining Systems (Cont..)

3. Classification according to the kinds of **techniques** utilized:
   - These techniques can be described according to the **degree of user interaction** involved (e.g., autonomous systems, query-driven systems).
   - The methods of data analysis employed (e.g., database-oriented or data warehouse–oriented techniques, machine learning, statistics, visualization, pattern recognition, neural networks etc.)
   - A sophisticated data mining system will often adopt multiple data mining techniques for work out an effective, integrated technique which combines the merits of a few individual approaches.
   - **E.g.** Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.

Classification of Data Mining Systems (Cont..)

4. Classification according to the **Applications** adapted:
    - Retail
    - Telecommunication
    - Banking
    - Fraud analysis
    - Stock market analysis
    - Text mining
    - Web mining etc.

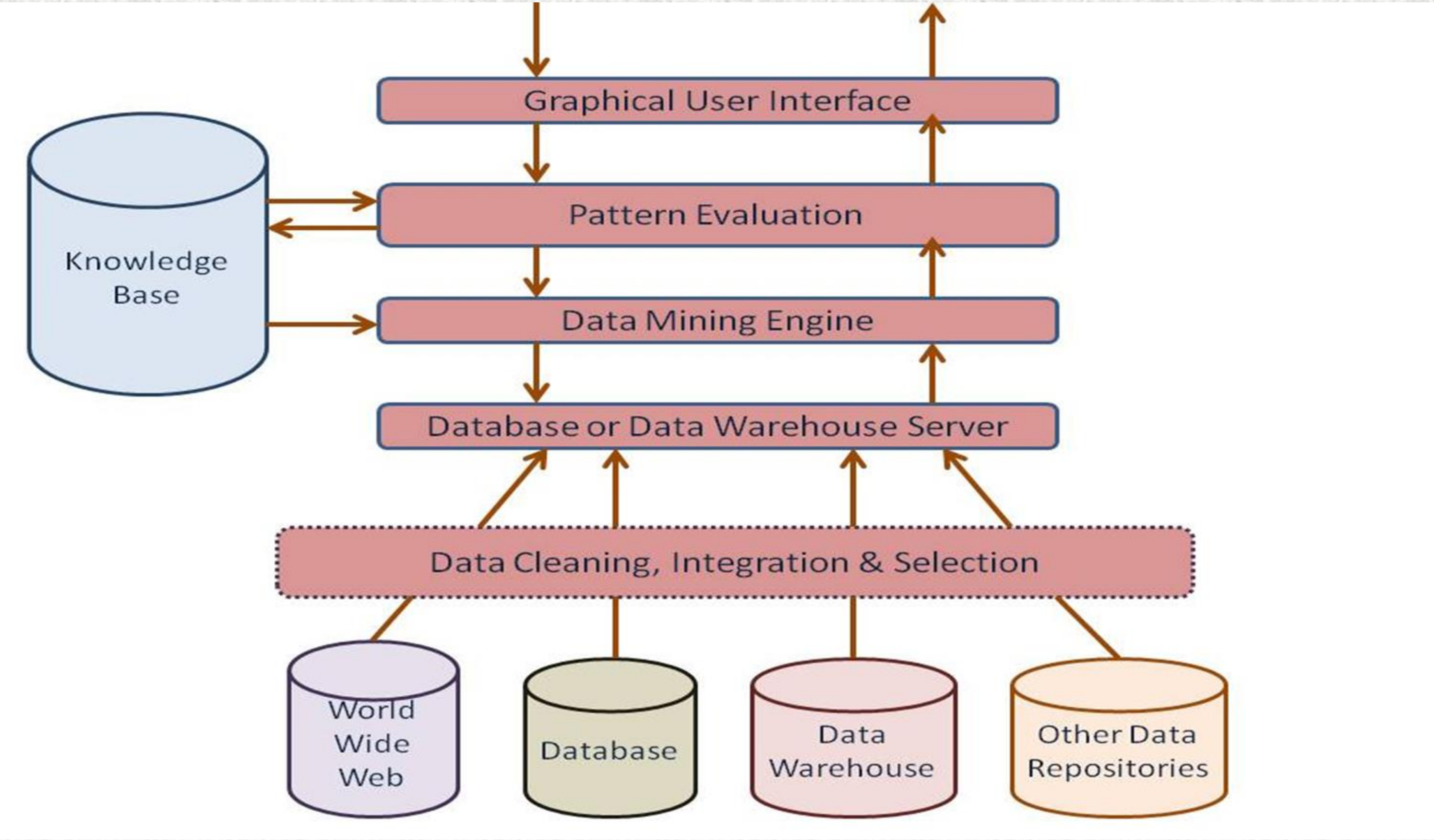
Data Mining—On what kind of data?

- **Relational Databases:**
  - A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data.
  - **E.g.** : SQL Server, Oracle etc.
- **Data Warehouses:**
  - A data warehouse is a repository of information collected from multiple sources.
  - Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.
  - **E.g.** : Stock Market, D-Mart, Big Bazar etc.

Data Mining—On what kind of data? (Cont..)

- **Transactional Databases:**
  - Transactional database consists of a file where each record represents a transaction.
  - A transaction typically includes a unique transaction identity number (TID) and a list of the items making up the transaction (such as items purchased in a store).
  - **E.g.** : Online shopping like Flipkart, Amazon etc.
- **Other Data**
  - Spatial data (Maps or Location)
  - Engineering design data (Design of Buildings, Offices Structures)
  - Hypertext and multimedia data (Including text, image, video, and audio data), the World Wide Web (a huge, widely distributed information repository made available on the Internet).

Data Mining Architecture



Data Mining Architecture (Cont..)

- **Data Mining Engine:**
  - It is essential to the data mining system and ideally consists of a set of **functional modules (knowledge) & methods** for different tasks such as…
    - Characterization
    - Association
    - Correlation analysis
    - Classification & prediction
    - Cluster analysis
    - Outlier analysis

Data Mining Architecture (Cont..)

- **Pattern Evaluation Module:**
  - This component typically employs **interestingness measures** and **interacts with the data mining modules** so it is focus in the search towards **interesting patterns**.
  - The pattern evaluation module integrated with the mining module, depending on the implementation of the data mining method used.

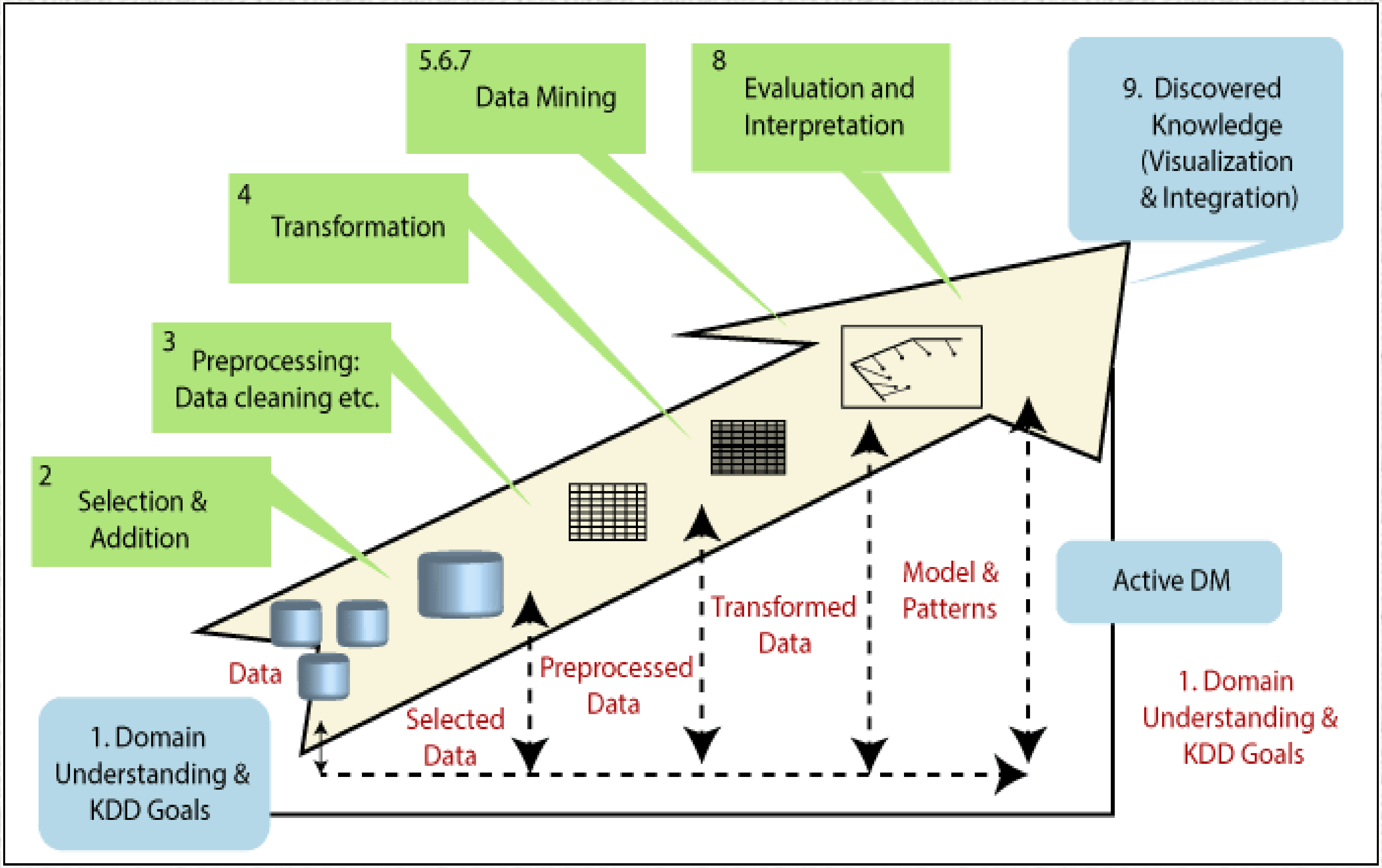
Data Mining Architecture (Cont..)

- **Knowledge base:**
  - Knowledge base 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 is such as **user beliefs**, which can be used to assess a pattern’s interestingness based on its unexpectedness, may also be included.

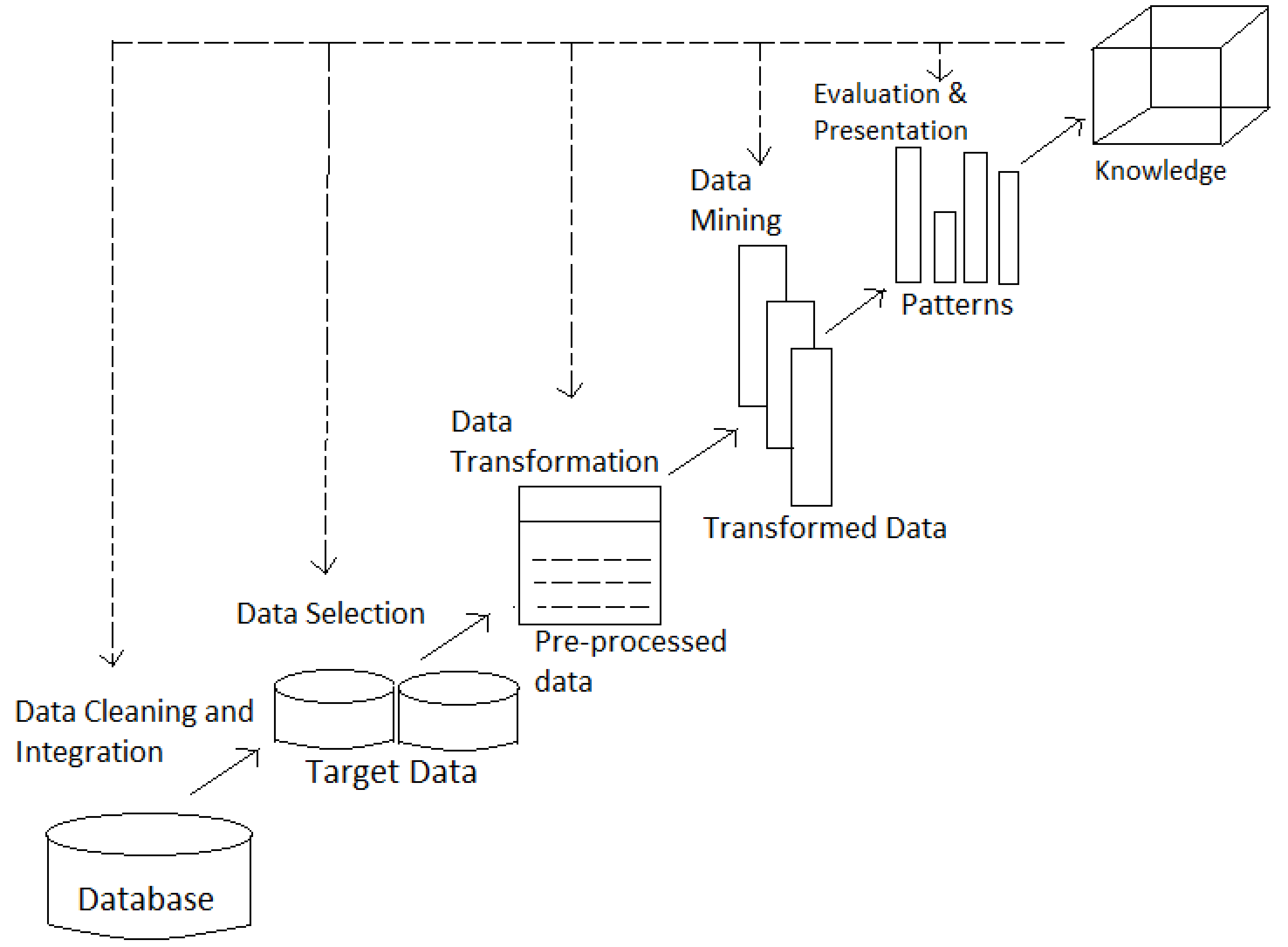
KDD (Knowledge Discovery in Databases) Process

- Knowledge discovery in databases is a process of an iterative sequence of the following steps:
  1. **Selection**
  2. **Preprocessing**
  3. **Transformation**
  4. **Data Mining**
  5. **Pattern Evaluation**
  6. **User Interface (Visualization of Pattern or Knowledge)**

KDD Process


5.6.7
Data Mining
8
Evaluation and Interpretation
9. Discovered Knowledge (Visualization & Integration)
4
Transformation
3
Preprocessing: Data cleaning etc.
2
Selection & Addition
Data
Selected Data
Preprocessed Data
Transformed Data
Model & Patterns
Active DM
1. Domain Understanding & KDD Goals
1. Domain Understanding & KDD Goals

KDD Process


Evaluation &
Presentation
Knowledge
Data
Mining
Patterns
Data
Transformation
Transformed Data
Data Selection
Pre-processed
data
Data Cleaning and
Integration
Target Data
Database

KDD Process (Cont..)

- **Data Selection:** Where data relevant to the analysis task are retrieved from the database.
- **Data Cleaning:** To remove noise and inconsistent data.
- **Data Integration:** Where multiple data sources may be combined.
- **Data Transformation**: Where data are transformed or consolidated into appropriate forms for mining by performing summary or aggregation operations.
- **Data Mining**: An essential process where intelligent methods are applied in order to extract data patterns.
- **Pattern Evaluation**: To identify the truly interesting patterns representing knowledge based on some interestingness measures.
- **Knowledge Presentation**: Where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

Data Mining Issues

- Data mining issues can be classified into five categories:

1. **Mining Methodology**
2. **User Interaction**
3. **Efficiency and Scalability**
4. **Diversity of Database Types**
5. **Data Mining and Society**

1) Mining Methodology

- **Mining various and new kinds of knowledge**
  - Data mining covers a wide spectrum of data analysis and knowledge discovery tasks, so these tasks may use the same database in **different ways and require the development of numerous data mining techniques**.
- **Mining knowledge in multidimensional space**
  - When searching for knowledge in large data sets, we can explore the data in multidimensional space.
  - That is, we can search for interesting patterns among **combinations of dimensions (attributes)** at varying levels of abstraction. Such mining is known as (exploratory) **multidimensional data mining**.

1) Mining Methodology (Cont..)

- **Data mining—an interdisciplinary effort**
  - The power of data mining can be substantially enhanced by integrating new methods from multiple disciplines.
  - For example, to mine data with natural language **text**, it makes sense to fuse data mining methods of **information retrieval** and **natural language processing**.
- **Handling uncertainty, noise, or incompleteness of data**
  - Data often contain **noise, errors, exceptions, uncertainty** or **incomplete**.
  - Errors and noise may confuse the data mining process, leading to the derivation of erroneous patterns.

2) User Interaction

- **Interactive mining**
  - The data mining process should be **highly interactive**. Thus, it is important to build **flexible user interfaces** and an exploratory mining environment, facilitating the user's interaction with the system.
- **Incorporation of background knowledge**
  - **Background knowledge, constraints, rules, and other information** regarding the domain under study should be incorporated into the knowledge discovery process.
- **Presentation and visualization of data mining results**
  - How any system can present data mining results, vividly(clear image in mind) and flexibly ?, so that the **discovered knowledge can be easily understood and directly usable by humans**.

3) Efficiency and Scalability

- **Efficiency and scalability of data mining algorithms**
  - Data mining **algorithms** must be **efficient and scalable** in order to effectively extract information from huge amounts of data lies in many data repositories or in dynamic data streams.
  - In other words, the **running time** of a data mining algorithm must be **predictable, short, and acceptable by applications**.
  - Efficiency, scalability, performance, optimization, and the ability to **execute in real time** are key criteria for **new mining algorithms**.
- **Parallel, distributed, and incremental mining algorithms**
  - The giant size of many data sets, the **wide distribution of data**, and the **computational complexity** of some data mining methods are factors that motivate the development of parallel and distributed data-intensive mining algorithms.

4) Diversity of Database Types

- **Handling complex types of data**
  - Data mining is how to uncover knowledge from **stream**, **time-series**, **sequence**, **graph**, **social network**, and **multirelational data**.
  - In mining various types of attributes are available and also different types of data in database or dataset.
- **Mining dynamic, networked, and global data repositories**
  - Data from multiple sources are connected by the Internet and various **kinds of networks** like **distributed** and **heterogeneous global information systems**.
  - The discovery of knowledge from **different sources** of **structured**, **semi-structured**, or **unstructured** challengeable.
  - **Web Mining**, **multisource data mining** and **information network mining** have become **challenging and fast-evolving data mining fields**.

5) Data Mining and Society

- **Social impacts of data mining**
  - With data mining penetrating our everyday lives, it is important to study the impact of data mining on society, How can we used at a mining technology to **benefit our society**? **How can we guard against its misuse**?
- **Privacy-preserving data mining**
  - Data mining will help in scientific discovery, business management, economy recovery, and security protection (e.g., the real-time discovery of intruders and cyber attacks).
  - However, it poses the risk of **disclosing** an **individual's personal information**.
- **Invisible data mining**
  - We cannot expect everyone in society to learn and master in data mining techniques.
  - For example, when purchasing items online, users may be unaware that the store is likely collecting data on the buying patterns of its customers, which may be **used to recommend other items** for **purchase in the future**.

Integration of data mining with database or data warehouse

- DB and DW systems, possible integration schemes include no coupling, loose coupling, semi -tight coupling, and tight-coupling.
- We examine each of these schemes, as follows:
- **1**.**No coupling**:
  - No coupling means that a DM system will not utilize any function of a DB or DW system.
  - It may fetch data from a particular source (such as a file system), process data using some data mining algorithms, and then store the mining results in another file.
- 2**.Loose coupling**:
  - Loose coupling means that a DM system will use some facilities of a DB or DW system, fetching data from a data repository managed by these systems, performing data mining, and then storing the mining results either in a file or in a designated place in a database or data Warehouse.
  - Loose coupling is better than no coupling because it can fetch any portion of data stored in databases or data warehouses by using query processing, indexing, and other system facilities.

Continue..

  - However, many loosely coupled mining systems are main memory-based. Because mining does not explore data structures and query optimization methods provided by DB or DW systems, it is difficult for loose coupling to achieve high scalability and good performance with large data sets.

- **3.Semitight coupling**:
  - Semitight coupling means that besides linking a DM system to a DB/DW system, efficient implementations of a few essential data mining primitives (identified by the analysis of frequently encountered data mining functions) can be provided in the DB/DW system.
  - These primitives can include sorting, indexing, aggregation, histogram analysis, multi way join, and precomputation of some essential statistical measures, such as sum, count, max, min ,standard deviation.
- **4.Tight coupling**:
  - Tight coupling means that a DM system is smoothly integrated into the DB/DW system.
  - The data mining subsystem is treated as one functional component of information system.
  - Data mining queries and functions are optimized based on mining query analysis, data structures, indexing schemes, and query processing methods of a DB or DW system.