Unit – I

Introduction

- What motivated Data Mining?
- Why it is important?
- Data Mining – On what kind of data
- Data Mining Functionalities
- What kinds of patterns can be mined?
- Are all of the patterns interesting?
- Classification of Data Mining Systems
- Data Mining Task Primitives
- Integration of a Data Mining System with a Database or Data Warehouse System
- Major Issues in Data Mining

What Motivated Data Mining? Why it is important?

- Necessity is the mother of invention
- The major reason that Data Mining has attracted a great deal of attention in the information industry in recent years is *due to thee wide availability of huge amount of data and the imminent need for turning such data into useful information and knowledge.*
- The information and knowledge gained can be used for applications ranging from business management, production control and market analysis.

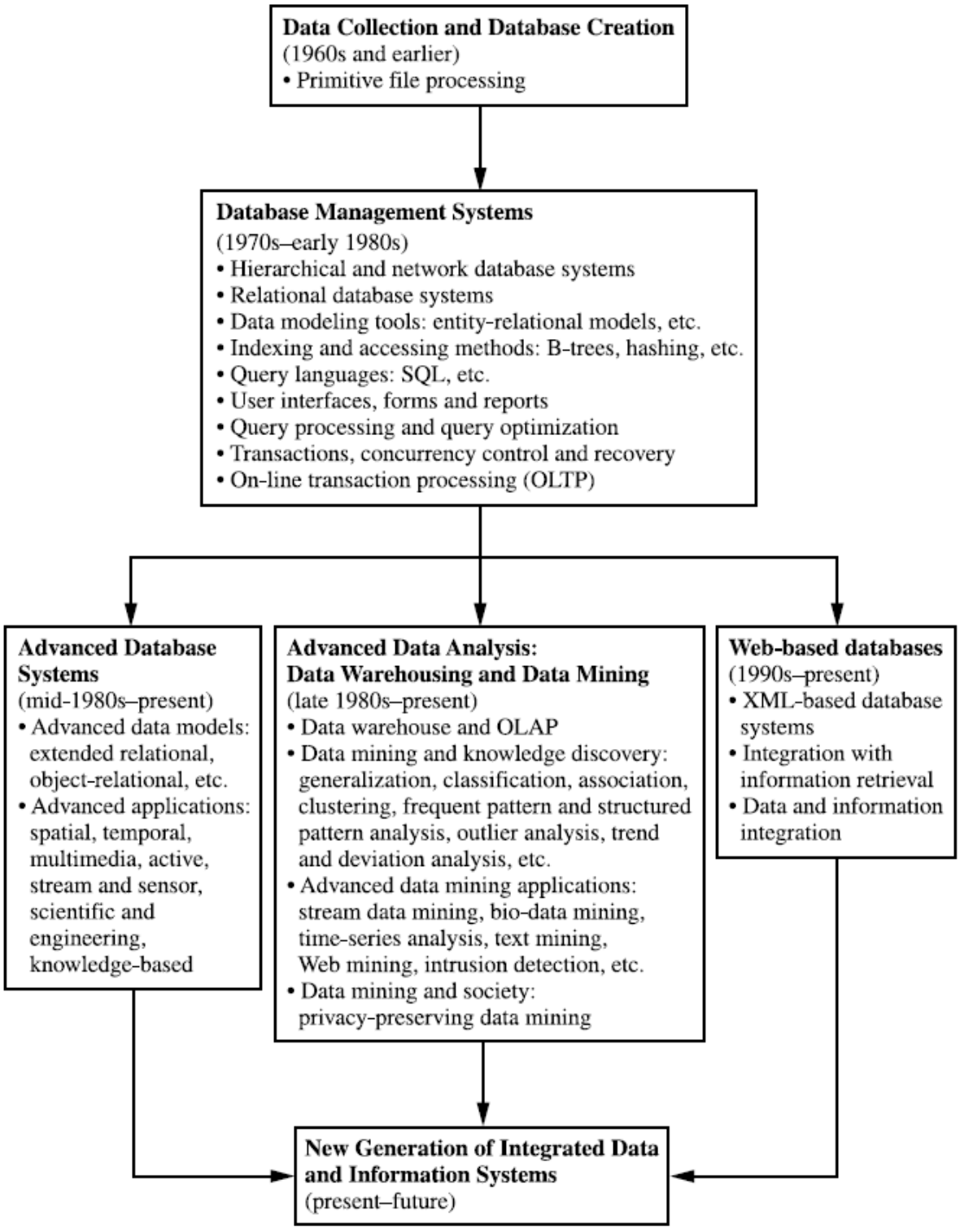
What is Data Mining?

Data Mining refers to extracting or mining knowledge from large amounts of data.

Synonyms to Data Mining:

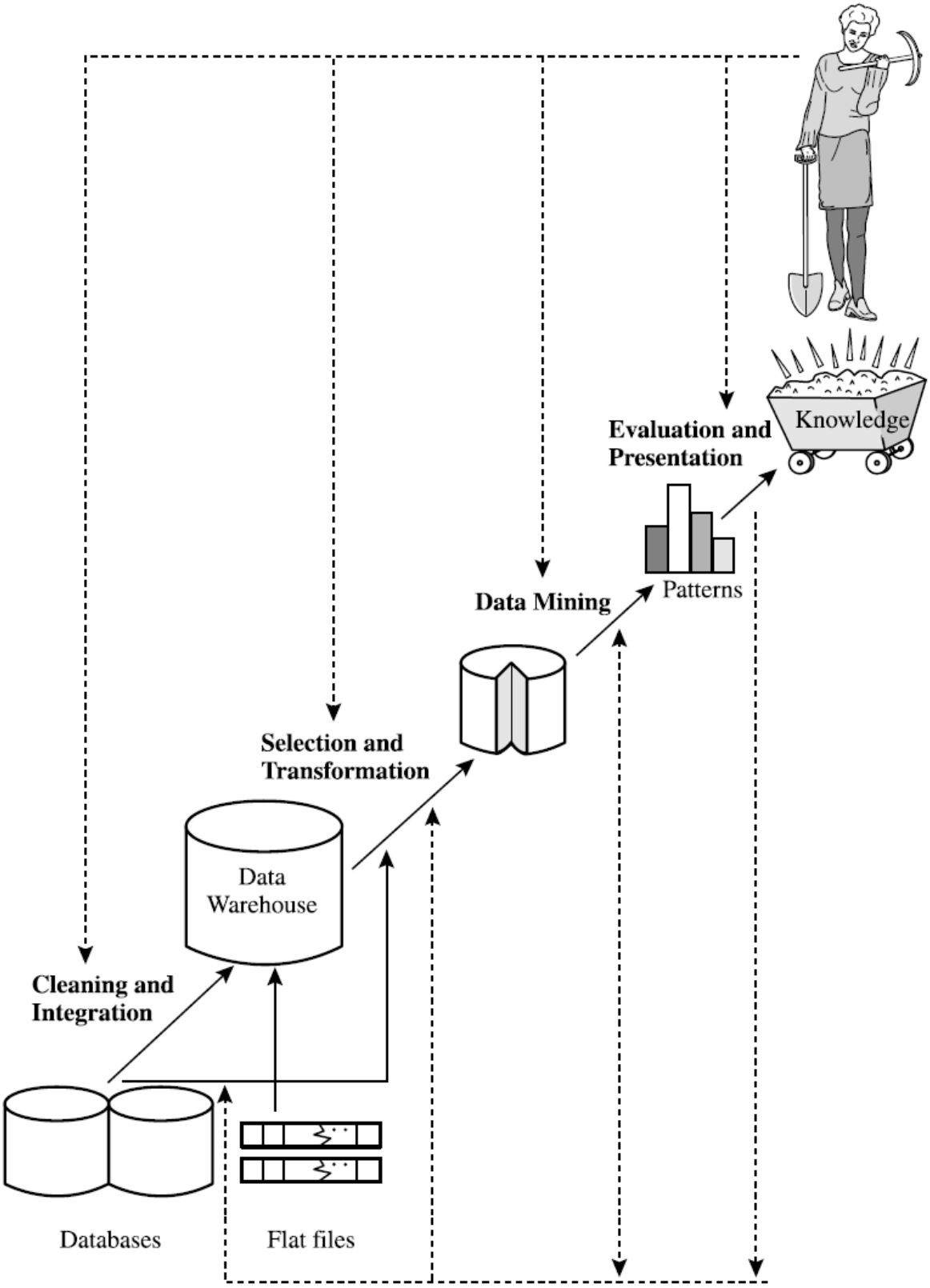
Knowledge mining from databases, knowledge extraction, data/ pattern analysis, data archeology and data dredging

Popular synonym is *"Knowledge Discovery in Databases (KDD)".*



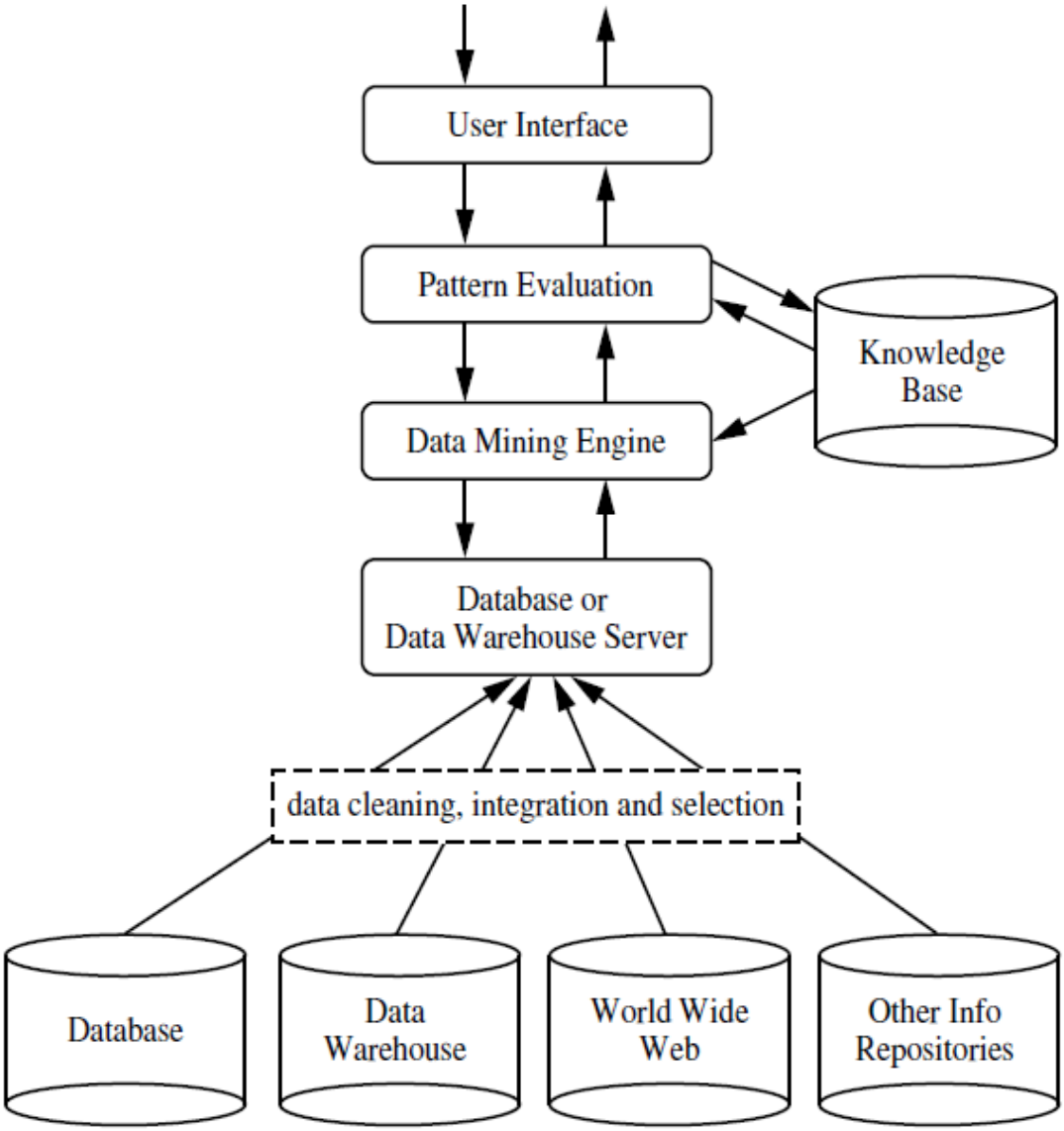
Evolution of Database Technology

Knowledge Discovery as a Process



1. **Data Cleaning:** remove noise and inconsistent data
2. **Data Integration:** multiple data sources to be combined
3. **Data Selection:** data relevant to analysis task are retrieved from the database
4. **Data Transformation:** data are transformed or consolidated into apt forms for mining. This is done by doing summary or aggregation operations
5. **Data Mining:** mining methods are applied for extracting patterns
6. **Pattern Evaluation:** identifies truly interesting patterns based on some interesting measures.
7. **Knowledge Representation:** visualization and knowledge representation techniques are used to present the discovered knowledge

Architecture of Data Mining System



Components of Data Mining System:

- Database, data warehouse or other information repository
- Database or data warehouse server
- Knowledge base
- Data mining engine
- Pattern evaluation module
- Graphical user interface

Database, data warehouse, World Wide Web, or other information repository: This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories. Data cleaning and data integration techniques are performed.
Database or data warehouse server: it is responsible for fetching relevant data from repository based on data mining task.
Knowledge Base: this is the domain knowledge which is used to guide the mining process. Includes concept hierarchies, thresholds or interestingness and meta data.
Data Mining Engine: contains functional modules like:

- Classification
- Association
- Cluster analysis
- Evolution analysis
- Outlier analysis

Pattern Evaluation Module: employs interestingness measures. Interacts with data mining engine in search of interesting patterns.
Graphical User Interface: allows user to interact with the system by providing data mining query or task.

Data Mining – On What Kind of Data?

1. Relational Databases
2. Data Warehouses – Data Cube
3. Transactional Databases – Transactional Data Set
4. Advanced Database Systems and Advanced Database Applications
    a. Object Oriented Databases
    b. Object Relational Databases
    c. Spatial Databases
    d. Temporal and Time Series Databases
    e. Text Databases & Multimedia Databases
    f. Heterogeneous Databases and Legacy Databases
    g. World Wide Web

Data Mining Functionalities – What Kinds of Patterns Can Be Mined?

1. **Concept/ Class Description: Characterization and Discrimination:**

Data can be associated with classes or concepts. It is useful to describe individual classes or concepts. Such descriptions are called class/ concept descriptions. These are:

(a) *Data Characterization:* by summarizing the data of the class (target class).

(b) *Data Discrimination:* by comparison of target class with one or set of comparative classes.

The output of data characterization can be presented in various forms like pie-charts, bar charts, curves, multidimensional data cubes and tables.

Discrimination descriptions are expressed in rule forms are referred as discriminant rules.

2. **Association Analysis:**

It is the discovery of association rules showing attribute value conditions that occur frequently together in a given set. Association analysis is widely used for market basket or transaction analysis.

Rules are of the form, X=>Y

Uses *support* and *confidence* values

Ex: contains(T, “Computer”) => contains(T, “Software”)

[support=1%, confidence=50%]

3. **Classification and Prediction:**

Classification is the process of finding a set of models that describe and distinguish classes or concepts. Classification predicts a class of objects whose class label is unknown which is based on the analysis of training dataset (objects whose class label is known). This is represented by simple “IF-THEN” rules.

A decision tree is a flow chart like tree structure where each node denotes a test on the attribute value, each branch represents an outcome of the test and tree leaves represents classes. A neural network when used for classification is typically a collection of neuron like processing units with weighted connections between the units.

Classification can be used for predicting the class label of data objects. In some applications users wish to predict some missing or unavailable data values rather than class labels. This is usually the case when the predicted values are numerical data and is often specifically referred to as *Prediction.*

4. **Cluster Analysis:**

Clustering analyzes data objects without consulting a known class label. These objects are clustered based on the principle:

“maximizing the intra-class similarity and minimizing the inter-class similarity”

Each cluster can be viewed as a set of objects from which rules for that cluster can be derived.

This also facilitates grouping formation i.e. hierarchy of classes.

5. **Outlier Analysis:**

A database may contain some objects which do not fit into model of data. Such objects are called *outliers.* Most of the mining methods exclude outliers as noise or exceptions during mining. Outliers in some applications like fraud detection proved to be interesting. Analysis of outlier data is referred as outlier mining.

Outliers are identified by using statistical test like distribution or probability model or distance measures. Rather than these deviations based methods identify outliers by examining differences in the main characteristics of objects in a group.

6. Evolution Analysis:

Evolution analysis describes and models regularities or trends for objects whose behavior changes over time. This includes characterization, discrimination, association, classification or clustering of time related data.

Are all of the Patterns Interesting?

No, only small fractions of the patterns are interesting for analysis.

A pattern is interesting if,

1. It is easily understood
2. Valid on new or test data with some degree of certainty
3. Useful potentially
4. Novel

An interesting pattern represents knowledge.

Classification of Data Mining Systems

Data Mining is an interdisciplinary field and is extending its research to generate wide variety of Data Mining Systems.

Data Mining Systems can be categorized according to various criteria, as follows:

a) Classification according to the kinds of databases mined
b) Classification according to the kinds of knowledge mined (classification & prediction, association analysis, clustering etc)
c) Classification according to the kinds of techniques used
    - Categorized based on the underlying data mining techniques employed
    - These techniques can be described according to the degree of user interaction (Ex: autonomous systems, query driven systems) or the methods of data analysis employed (database or data warehouse oriented techniques, pattern recognition, visualization, neural networks)
    - A sophisticated data mining system adopts multiple techniques
d) Classification according to the applications adapted

    Data mining systems can be categorized based on applications. For ex: data mining systems can be tailored to areas like finance, telecommunications, stock market etc.

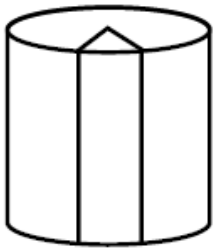
Data Mining Task Primitives

Each user will have a data mining task to perform some data analysis. This data mining task can be of a data mining query which is given as input to data mining system.

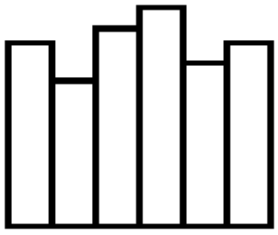
A data mining query is defined as set of data mining task primitives. These primitives allows user to interact with the data mining system effectively during the process of knowledge discovery. They are:

a) *The set of task relevant data to be mined*

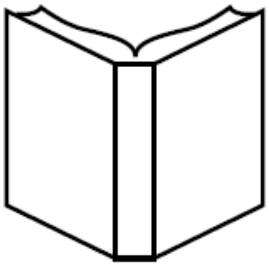
Specifies the portion of database or data warehouse in which the user is interested. This includes relevant attributes of database or dimensions of interest of data warehouse.



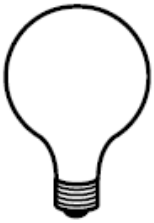
Task-relevant data
Database or data warehouse name
Database tables or data warehouse cubes
Conditions for data selection
Relevant attributes or dimensions
Data grouping criteria



Knowledge type to be mined
Characterization
Discrimination
Association/correlation
Classification/prediction
Clustering



Background knowledge
Concept hierarchies
User beliefs about relationships in the data



Pattern interestingness measures
Simplicity
Certainty (e.g., confidence)
Utility (e.g., support)
Novelty

Visualization of discovered patterns
Rules, tables, reports, charts, graphs, decision trees, and cubes
Drill-down and roll-up

b) *The kind of knowledge to be mined*
Specifies data mining functions to be performed
- Characterization
- Classification and prediction
- Association
- Clustering

c) *The background knowledge to be used in the discovery process*
Knowledge about the domain to be mined is helpful in guiding the knowledge discovery process.

d) *The interestingness measures and the thresholds for pattern evaluation.*
They may be used to guide the mining process or after discovery to evaluate the discovered patterns. Different kinds of knowledge have different interesting measures.
Ex: support and confidence are used to measure the interestingness of association rules.

e) *The expected representation for visualizing the discovered patterns.*
This refers to representation of discovered knowledge. This is done by using charts, graphs, decision trees etc.

Integration of Data Mining System with a Database or Data Warehouse System

Four types. They are:

1) No Coupling
Data Mining System will not utilize any function of database or data warehouse system. It fetches data and process data (from file system) using some data mining algorithms and stores the result in another file.

2) Loose Coupling
Data Mining System will use some facilities of database or data warehouse system. Fetching data from a repository is managed by these systems, performing data mining and then stores results either in a file or in database or in data warehouse at some designated place.

3) Semi-tight Coupling
Data Mining System is linked to database/ data warehouse system. Efficient implementations of few data mining primitives are provided in database/ data warehouse system.
These primitives include sorting, indexing, aggregation, histogram analysis, pre-computation of some statistical measures like sum, count, standard deviation etc.

4) Tight Coupling
Data Mining System is smoothly integrated into database/ data warehouse system. Data mining system is treated as functional component of information system. It facilitates efficient implementations of data mining functions, high system performance and an integrated information processing environment.

Issues in Data Mining

Data Mining issues are related to *mining methodologies, user interaction, performance* and *diverse data types.*

a) *Mining Methodology and User Interaction Issues*
   - Mining different kinds of knowledge in databases
   - Interactive mining of knowledge at multiple levels abstraction
   - Incorporation of background knowledge
   - Data Mining Query Languages and adhoc data mining
   - Presentation and visualization of data mining results
   - Handling noise or incomplete data
   - Pattern evaluation

b) *Performance Issues*
   - Efficiency and scalability of data mining algorithms
   - Parallel, distributed and incremental mining algorithms

c) *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