

Data Mining:

Concepts and Techniques

Prerequisites

➤ Knowledge of:

- basic probability theory
- algorithms

Introduction

Motivation: Why data mining?

Why Data Mining?

➤ The Explosive Growth of Data: from terabytes(1000^4) to yottabytes(1000^8)

- Data collection and data availability

- ✓ Automated data collection tools, database systems, web

- Major sources of abundant data

- Business: Web, e-commerce, transactions, stocks, ...
 - Science: bioinformatics, scientific simulation, medical research ...
 - Society and everyone: news, digital cameras, ...

➤ Data rich but information poor!

- What does those data mean?
- How to analyze data?

➤ Data mining — Automated analysis of massive data sets

Why Mine Data? Commercial Viewpoint

➤ Lots of data is being collected and warehoused

- Web data, e-commerce
- purchases at department/ grocery stores
- Bank/Credit Card transactions



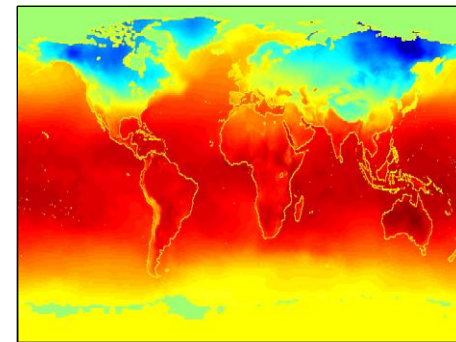
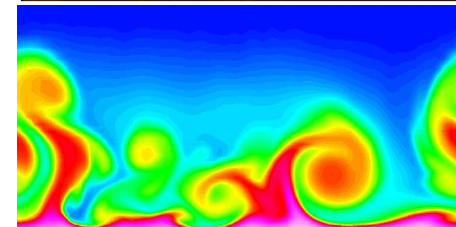
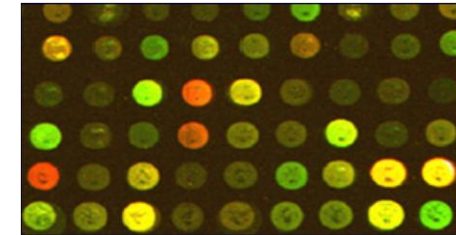
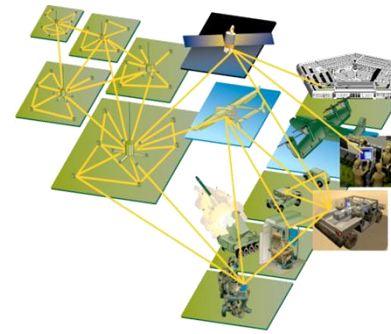
➤ Computers have become cheaper and more powerful

➤ Competitive Pressure is Strong

- Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
 - remote sensors on a satellite
 - telescopes scanning the skies
 - microarrays generating gene expression data
 - scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
 - in classifying and segmenting data
 - in Hypothesis Formation



Examples: What is (not) Data Mining?

❑ What is not Data Mining?

- Look up phone number in phone directory
- Query a Web search engine for information about “Amazon”

❑ What is Data Mining?

- Certain names are more prevalent in certain US locations (O’Brien, O’Rourke, O’Reilly... in Boston area)
- Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Why we need Data Mining? Cont'

- So, we need a **system** that will be capable of **extracting essence of information available** and that can **automatically generate report, views or summary of data for better decision-making.**

What Is Data Mining?



➤ Data mining (knowledge discovery from data)

- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data

➤ Alternative names and their “inside stories”:

- Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

➤ Patterns must be:

- **valid, novel, potentially useful, understandable**



Example of discovered patterns

- **Association rules:**

“80% of customers who buy *cheese* and *milk* also buy *bread*, and 5% of customers buy all of them together”

Cheese, Milk → Bread [sup = 5%, confid = 80%]

Main data mining tasks

- **Classification:**

mining patterns that can classify future data into known classes.

- **Association rule mining**

mining any rule of the form $X \rightarrow Y$, where X and Y are sets of data items.

- **Clustering**

identifying a set of similarity groups in the data

Main data mining tasks (cont ...)

➤ Sequential pattern mining:

A sequential rule: $A \rightarrow B$, says that event A will be immediately followed by event B with a certain confidence

➤ Deviation detection:

discovering the most significant changes in data

➤ Data visualization: using graphical methods to show patterns in data.

Why is data mining important?

- Rapid computerization of businesses produce huge amount of data
- How to make best use of data?
- A growing realization: knowledge discovered from data can be used for competitive advantage.

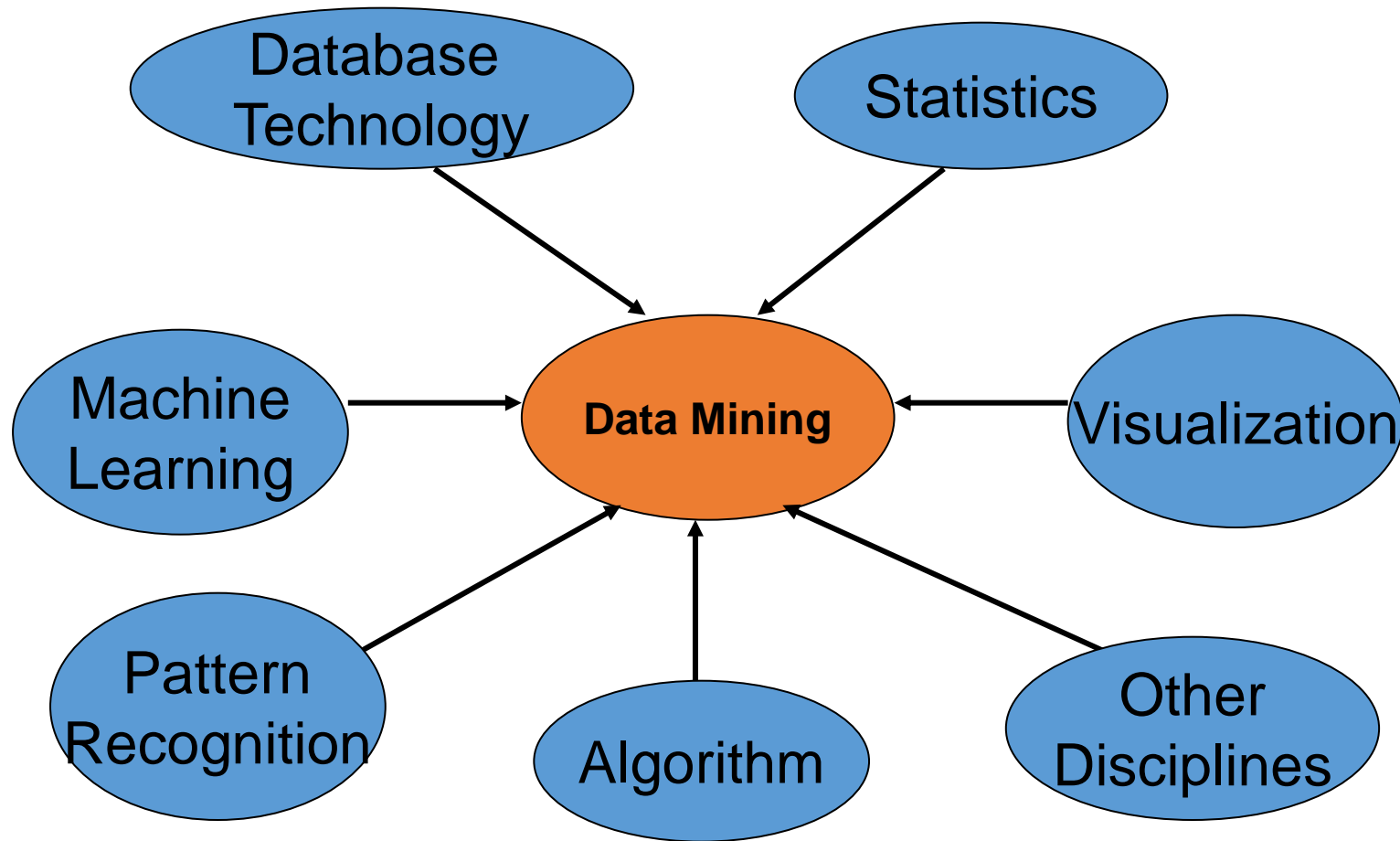
Why is data mining necessary? Cont'

- Make use of your data assets
- There is a big gap from stored data to knowledge; and the transition won't occur automatically.
- Many interesting things you want to find cannot be found using database queries
 - “find me people likely to buy my products”
 - “Who are likely to respond to my promotion”

Why data mining now?

- The data is abundant.
- The data is being warehoused.
- The computing power is affordable.
- The competitive pressure is strong.
- Data mining tools have become available

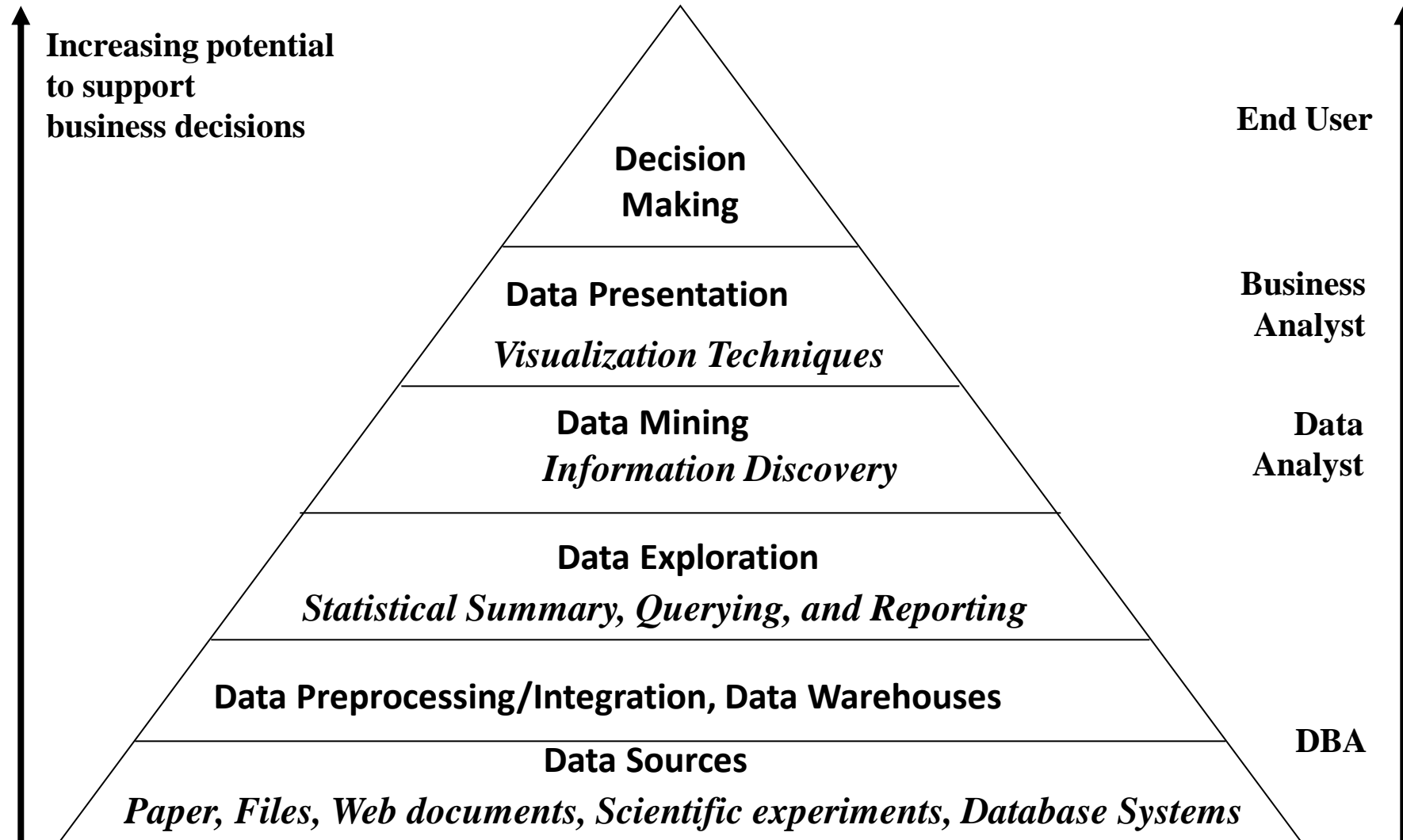
Related fields



- Data mining is an emerging multi-disciplinary field:

- ✓ **Statistics**
- ✓ **Machine learning**
- ✓ **Databases**
- ✓ **Information retrieval**
- ✓ **Visualization**
- etc.

Data Mining and Business Intelligence



Summary on DM Definition

- Data mining refers to extraction of information from large amount of data.
- In today's world data mining is very important because huge amount of data is present in companies and different type of organization.
- It becomes impossible for humans to extract information from this large data, so machine learning technology are used in order to process data fast enough to extract information from it.
- Data mining is used by companies in order to get customer preferences, determine price of their product and services and to analyse market.
- Data mining is also known as **knowledge discovery in Database (KDD)**

Why Data Mining?—Potential Applications

- Data analysis and decision support
 - Market analysis and management
 - Corporate analysis and Risk management
 - Fraud detection and detection of unusual patterns (outliers)

Why Data Mining?—Potential Applications Cont'

Market Analysis and Management

- Target marketing
 - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
 - Determine customer purchasing patterns over time
- **Cross-market analysis—Find** associations/co-relations between product sales, & predict based on such association
- **Customer profiling**
 - What types of customers buy what products
- **Customer requirement analysis**
 - Identifying the best products for different customers
 - Predict what factors will attract new customers

Corporate Analysis & Risk Management

- **Finance planning and asset evaluation**

- cash flow analysis and prediction
- contingent claim analysis to evaluate assets
- cross-sectional and time series analysis (financial-ratio, trend analysis, etc.)

- **Resource planning**

- summarize and compare the resources and spending

- **Competition**

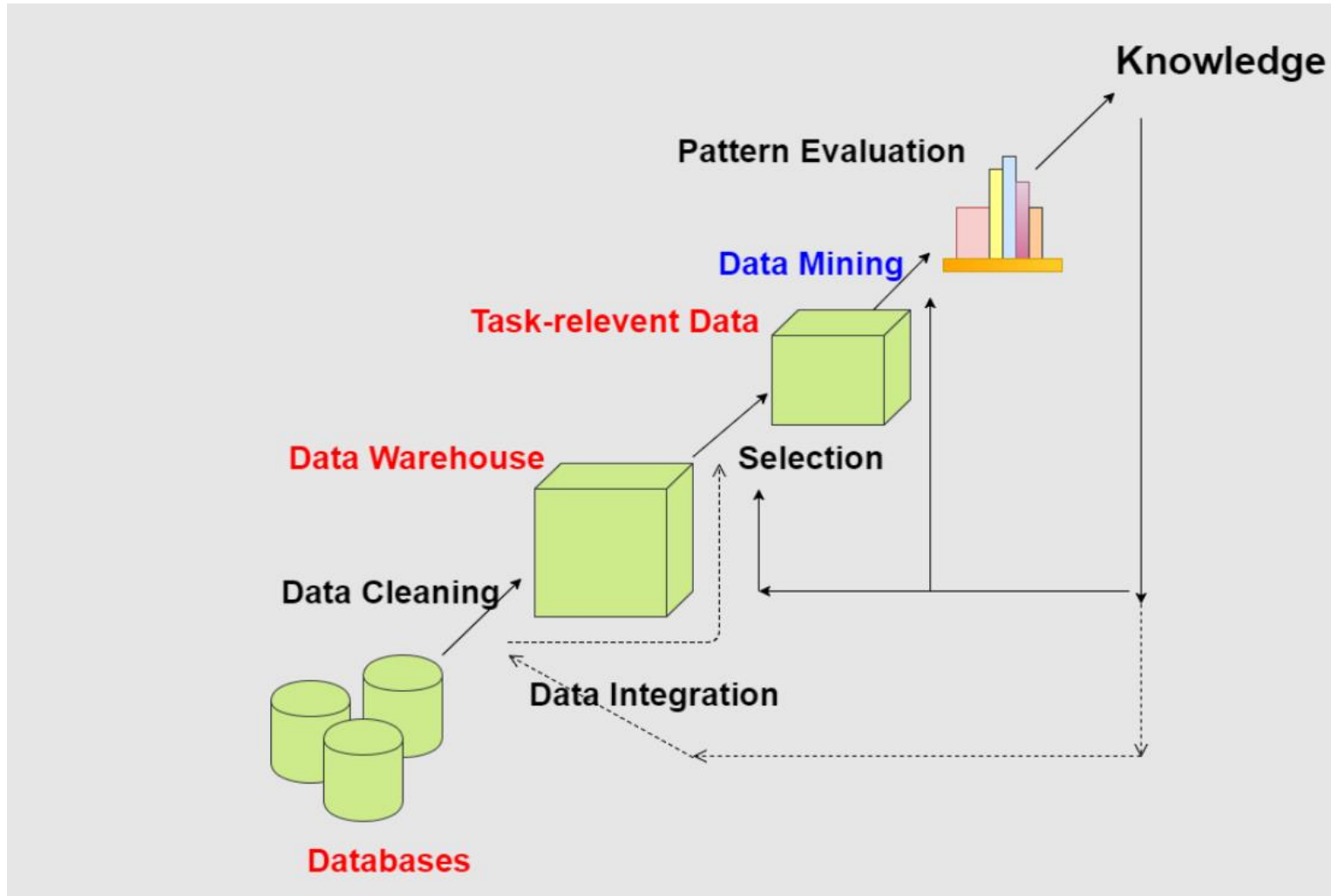
- monitor competitors and market directions
- group customers into classes and a class-based pricing procedure
- set pricing strategy in a highly competitive market

Other data mining applications

- **Fraud detection:**
identifying credit card fraud, intrusion detection
- Scientific data analysis
- Text (news group, email, documents) and web mining
- Bioinformatics and bio-data analysis
- Any application that involves a large amount of data ...

Data Mining: Knowledge Discovery (KDD) Process

- **Data mining** is the core part of the knowledge discovery process.



Knowledge Discovery Process (KDP)

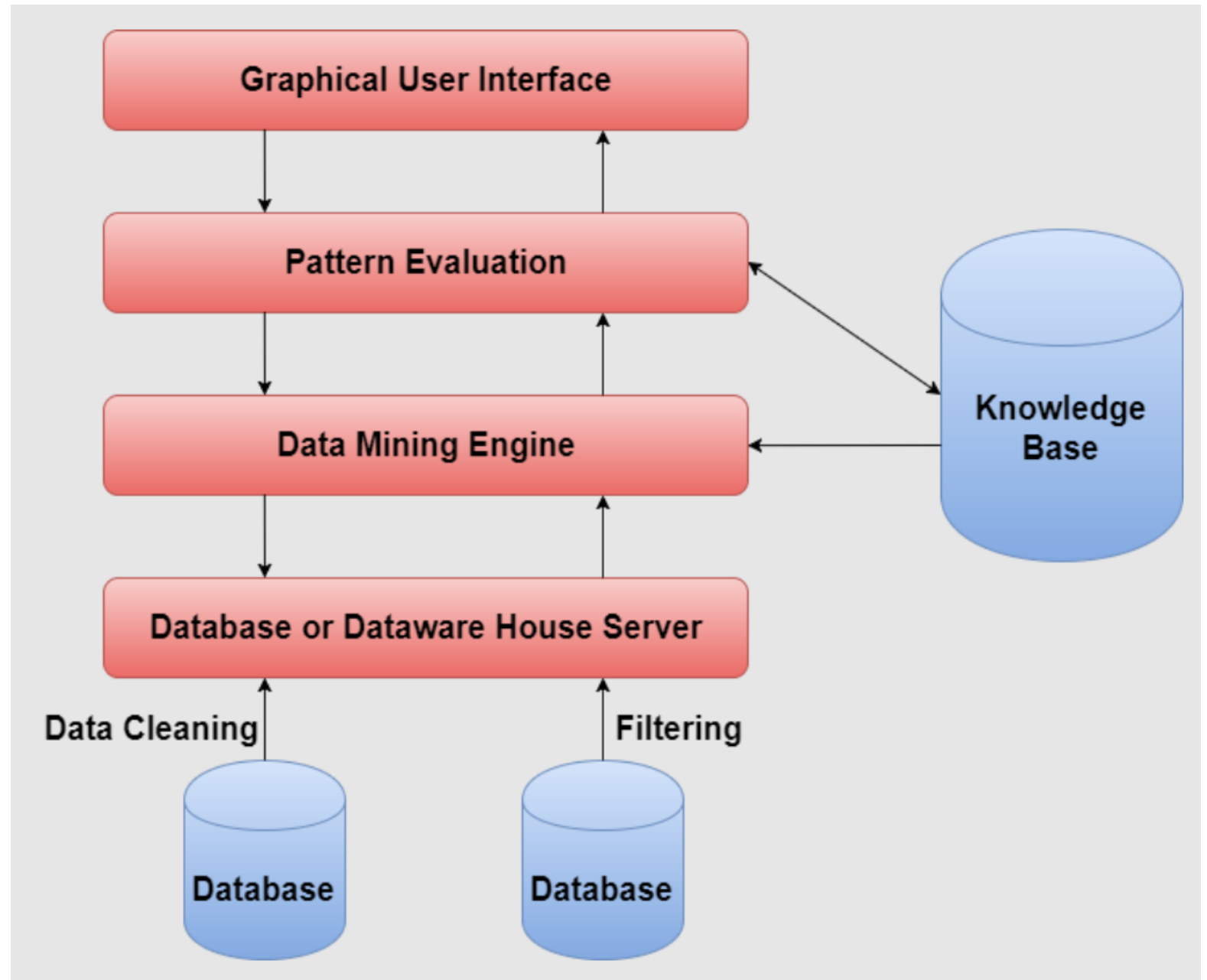
- **KDD process** is a process of finding knowledge in data, it does this by using **data mining methods (algorithms)** in order to extract demanding knowledge from large amount of data.

Knowledge Discovery Process may consist of the following steps:

1. **Data cleaning:** noise and inconsistent data is removed.
 2. **Data integration:** multiple data sources may be combined.
 3. **Data selection:** data relevant to the analysis task are retrieved from the database.
 4. **Data transformation:** data are transformed into forms appropriate for mining by performing summary or aggregation operations.
 5. **Data mining:** data mining methods (algorithms) are applied in order to extract data patterns.
 6. **Pattern evaluation:** data patterns are identified based on some interesting measures.
 7. **Knowledge presentation:** visualization and knowledge representation techniques are used to present the mined knowledge to the user
- A **data pattern** defines the way in which the **data** collected (semi-structured **data**) can be structured, indexed, and made available for searching

Data Mining Architecture

- Data mining architecture has many elements like Data Warehouse, Data Mining Engine, Pattern evaluation, User Interface and Knowledge Base.



- **Data Warehouse:** a **data warehouse** is a place which store information collected from multiple sources under unified schema.
- Information stored in a data warehouse is critical to organizations for the process of decision-making.
- **Data Mining Engine:** **Data Mining Engine** is the core component of data mining process which consists of various modules that are used to perform various tasks like **clustering, classification, prediction and correlation analysis.**

- **Pattern Evaluation:** Pattern Evaluation is responsible for finding various patterns with the help of Data Mining Engine.
- **User Interface:** User Interface provides communication between user and data mining system.
- It allows user to use the system easily even if user doesn't have proper knowledge of the system.
- **Knowledge Base:** Knowledge Base consists of data that is very important in the process of data mining.
- Knowledge Base provides input to the data mining engine which guides data mining engine in the process of pattern search.

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

Database Processing vs. Data Mining Processing

- **Query**

- Well defined
- SQL

- **Data**

- Operational data

- **Output**

- Precise
- Subset of database

- **Query**

- Poorly defined
- No precise query language

- **Data**

- Not operational data

- **Output**

- Fuzzy (not precise)
- Not a subset of database

Query Examples

- **Database**

- Find all credit applicants with last name of Smith.
- Identify customers who have purchased more than \$10,000 in the last month.
- Find all customers who have purchased milk

- **Data Mining**

- Find all credit applicants who are poor credit risks. **(classification)**
- Identify customers with similar buying habits. **(Clustering)**
- Find all items which are frequently purchased with milk.
(association rules)

Data Mining: Classification Schemes

➤ Decisions in data mining

- Kinds of databases to be mined
- Kinds of knowledge to be discovered
- Kinds of techniques utilized
- Kinds of applications adapted

➤ Data mining tasks

- Descriptive data mining
- Predictive data mining

Decisions in Data Mining

- **Databases to be mined**

- Relational, transactional, object-oriented, object-relational, active, spatial, time-series, text, multi-media, heterogeneous, WWW, etc.

- **Knowledge to be mined**

- Characterization, discrimination, association, classification, clustering, trend, deviation and outlier analysis, etc.
- Multiple/integrated functions and mining at multiple levels

- **Techniques utilized**

- Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, neural network, etc.

- **Applications adapted**

- Retail, telecommunication, banking, fraud analysis, DNA mining, stock market analysis, Web mining, Weblog analysis, etc.

Data Mining Tasks

➤ Prediction Tasks

- Use some variables to predict unknown or future values of other variables

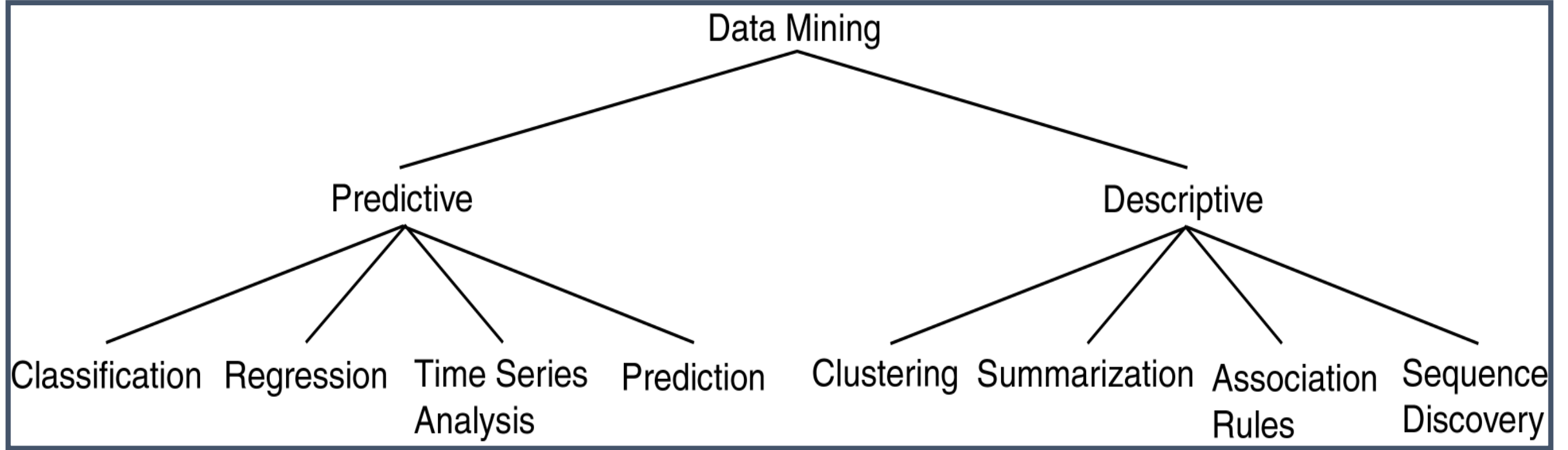
➤ Description Tasks

- Find human-interpretable patterns that describe the data.

■ Common data mining tasks

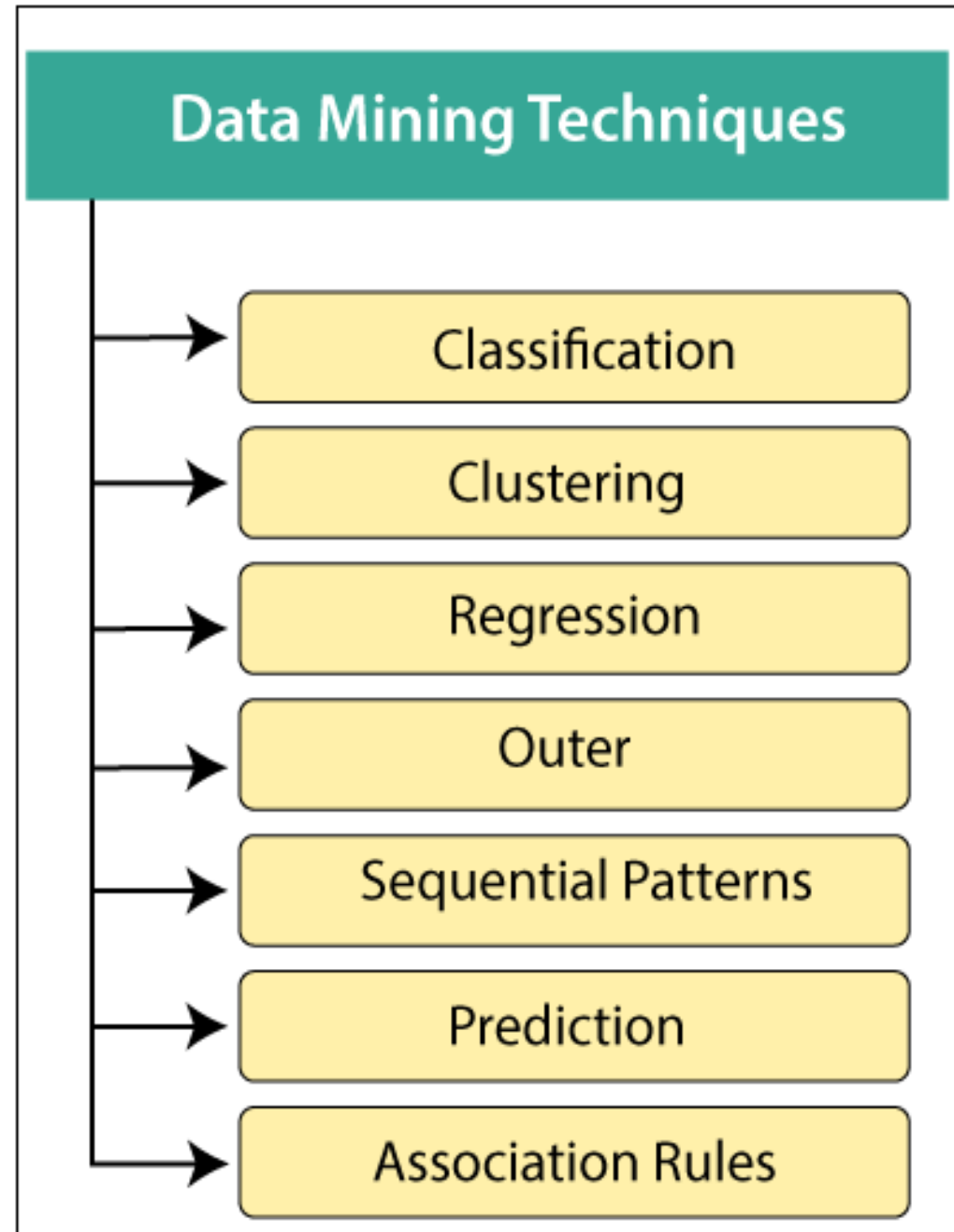
- **Classification** [Predictive]
- **Clustering** [Descriptive]
- **Association Rule Discovery** [Descriptive]
- **Sequential Pattern Discovery** [Descriptive]
- **Regression** [Predictive]
- **Deviation Detection** [Predictive]

Data Mining Models and Tasks



Data Mining Techniques

- Data mining includes the utilization of refined data analysis tools to find previously unknown, valid patterns and relationships in huge data sets.
- These tools can incorporate statistical models, machine learning techniques, and mathematical algorithms, such as neural networks or decision trees.
- Thus, data mining incorporates analysis and prediction.



Data Mining Functionalities (1)

- **Multidimensional concept description: Characterization and discrimination**
 - Data can be associated with **classes or concepts**.
 - It can be useful to **describe individual classes and concepts** in summarized, concise, and yet **precise terms**.
 - Such **descriptions** of a class or a concept are called **class/concept descriptions**.
 - These descriptions can be derived via (1) **data characterization, by summarizing the data of the class under study (often called the target class)** in general terms, or (2) **data discrimination, by comparison of the target class with one or a set of comparative classes (often called the contrasting classes)**, or (3) both **data characterization and discrimination**.

Data characterization:

- Data characterization is a summarization of the general characteristics or features of a **target class of data**.

Data discrimination:

- Data discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes.

Classification Definition

- This data mining technique helps to classify data in different classes.

Classification and prediction

➤ Construct models (functions) that 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 or missing numerical values

Classification: Definition cont'

- Given a collection of records (*training set*)
 - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
 - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

- Classification technique is most **common data mining technique**.
- In classification method we use mathematical techniques such as **decision trees, neural network and statistics in order to predict unknown records**.
- This technique helps in deriving important information about data.
- Let assume you have set of records, each record contains a set of attributes and depending upon this attributes you will be able to **predict unseen or unknown records**.
- For example, you have given all records of employees who left the company, with classification technique you can predict who will probably leave the company in a future period.

- **Data mining techniques can be classified by different criteria, as follows:**
 - i. **Classification of Data mining frameworks as per the type of data sources mined:**
 - This classification is as per the type of data handled.
 - For example, multimedia, spatial data, text data, time-series data, World Wide Web, and so on..
 - ii. **Classification of data mining frameworks as per the database involved:**
 - This classification based on the data model involved.
 - For example:
 - Object-oriented database, transactional database, relational database, and so on..

iii. Classification of data mining frameworks as per the kind of knowledge discovered:

- This classification depends on the types of knowledge discovered or data mining functionalities.
- For example, discrimination, classification, clustering, characterization, etc.
- some frameworks tend to be extensive frameworks offering a few data mining functionalities together..

iv. Classification of data mining frameworks according to data mining techniques used:

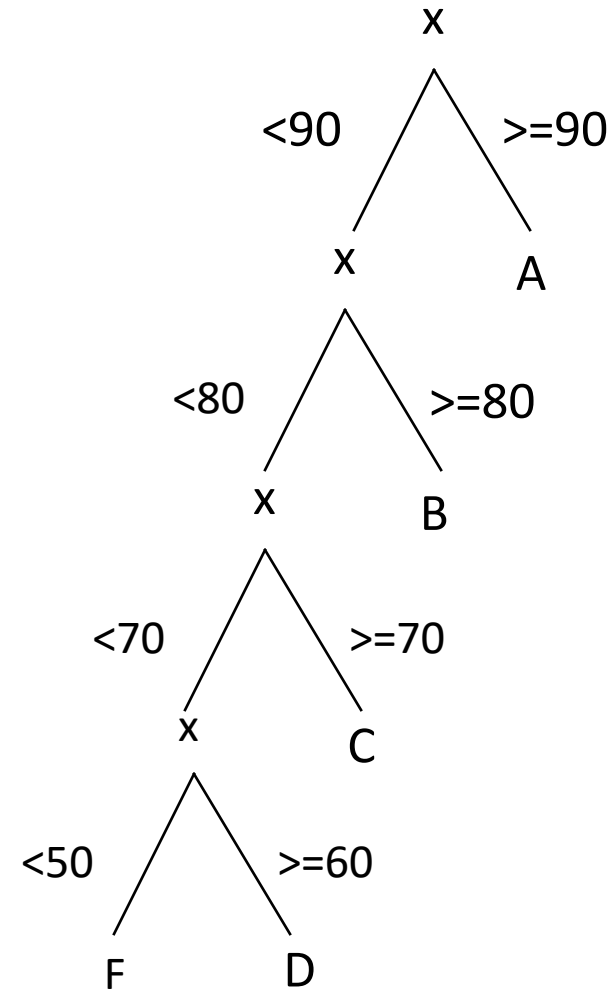
- This classification is as per the data analysis approach utilized, such as **neural networks, machine learning, genetic algorithms, visualization, statistics, data warehouse-oriented or database-oriented**, etc.
- The classification can also take into account, the level of user interaction involved in the data mining procedure, such as query-driven systems, autonomous systems, or interactive exploratory systems.

Classification Examples

- Teachers classify students' grades as A, B, C, D, or F.
- Predict when a river will flood.
- Identify individuals with credit risks.
- Speech recognition
- Pattern recognition

Classification Ex: Grading


- If $x \geq 90$ then grade =A.
- If $80 \leq x < 90$ then grade =B.
- If $70 \leq x < 80$ then grade =C.
- If $60 \leq x < 70$ then grade =D.
- If $x < 50$ then grade =F.





Classification Ex: Letter Recognition


View letters as constructed from 5 components:




 Letter A

 Letter B

 Letter C

 Letter D

 Letter E

 Letter F

Classification Techniques

- Approach:
 1. Create specific model by evaluating training data (or using domain experts' knowledge).
 2. Apply model developed to new data.
- Classes must be predefined
- Most common techniques use **DTs, NNs, or are based on distances or statistical methods.**

Classification Requirements

The two important steps of classification are:

1. Model construction

- A predefined class label is assigned to every sample tuple or object. These tuples or subset data are known as training data set.
- The constructed model, which is based on training set is represented as classification rules, decision trees or mathematical formulae.

2. Model usage

- The constructed model is used to perform classification of unknown objects.
- A class label of test sample is compared with the resultant class label.
- Accuracy of model is compared by calculating the percentage of test set samples, that are correctly classified by the constructed model.
- Test sample data and training data sample are always different.

Classification vs Prediction

Classification	Prediction
It uses the prediction to predict the class labels.	It is used to assess the values of an attribute of a given sample.
For example: If the patients are grouped on the basis of their known medical data and treatment outcome, then it is considered as classification.	For example: If a classification model is used to predict the treatment outcome for a new patient, then it is prediction.

- **Prediction:** It deals with some variables or fields, **which are available in the data** set to **predict unknown values regarding other variables of interest.**
- Numeric prediction is the type of predicting continuous or ordered values for given input.
- **For example:** The company may wish to predict the potential sales of a new product given with its price.
- The most widely used approach for numeric prediction is **regression.**

Issues related to Classification and Prediction

1. Data preparation

Data preparation consist of data cleaning, relevance analysis and data transformation.

2. Evaluation of classification methods

- i) **Predictive accuracy:** This is an ability of a model to predict the class label of a new or previously unseen data.
- ii) **Speed and scalability:** It refers to the time required to construct and use the model and increase efficiency in disk- resident databases.

3. Inter-predictability:

It is an understanding and insight provided by the model.

Category of Classification Algorithms

Generative

A generative classification algorithm models the distribution of individual classes. It tries to learn the model which creates the data through estimation of distributions and assumptions of the model. You can use generative algorithms to predict unseen data.

A prominent generative algorithm is the Naive Bayes Classifier.

Discriminative

It's a rudimentary classification algorithm that determines a class for a row of data. It models by using the observed data and depends on the data quality instead of its distributions.

Logistic regression is an excellent type of discriminative classifiers.

Classifiers in Machine Learning

- Classification is a highly popular aspect of data mining.
- As a result, machine learning has many classifiers:

1. Decision Trees

2. Bayesian Classifiers

3. K-Nearest Neighbour

5. Support Vector Machines

6. Linear Regression

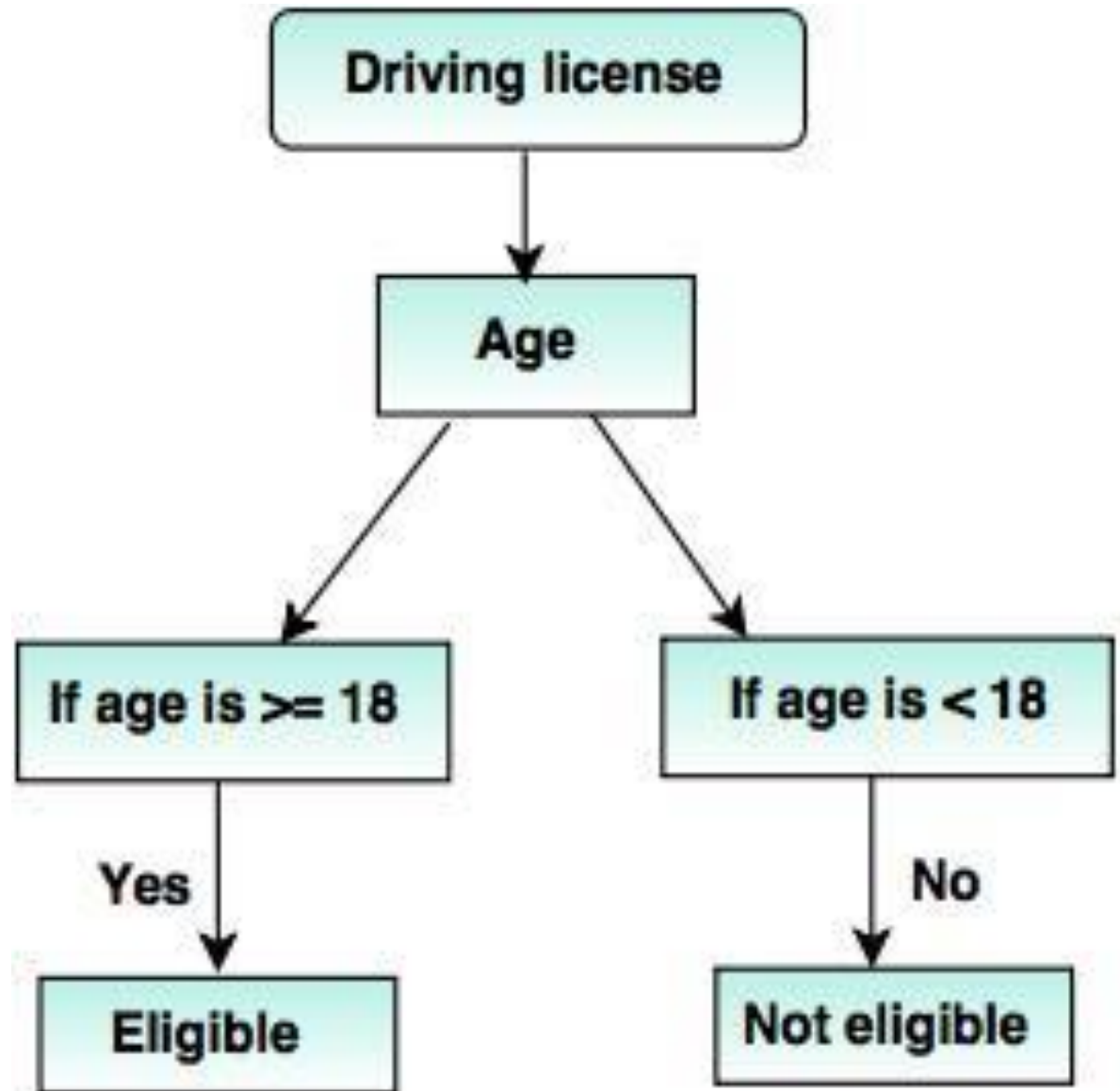
7. Logistic Regression

Decision tree

- A decision tree performs the classification in the form of **tree structure**.
- It breaks down the dataset into small subsets and a decision tree can be designed simultaneously.
- The final result is a **tree with decision node**.

For example:

- The following decision tree can be designed to declare a result, whether an applicant is eligible or not eligible to get the driving license.



Decision tree

- In the image above, you can see that **population** is classified into **four different groups** based on **multiple attributes** to identify 'if they will play or not'.

