UNIT 2

Data Mining

UNIT 2

TOPICS TO BE COVERED

1. Basic Concepts of data Mining
2. Different types of data repositories
3. Data mining Functionalities
4. Concepts of interesting patterns
5. Data Mining tasks
6. Current Trends
7. Major issues and ethics in data mining

BASIC CONCEPTS OF DATA MINING

A Brief History of Data Mining Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
- More conferences on data mining
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.
- ACM Transactions on KDD starting in 2007

Conferences and Journals on Data Mining

- KDD Conferences
  - ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
  - SIAM Data Mining Conf. (SDM)
  - (IEEE) Int. Conf. on Data Mining (ICDM)
  - European Conf. on Machine Learning and Principles and practices of Knowledge Discovery and Data Mining (ECML-PKDD)
  - Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)
  - Int. Conf. on Web Search and Data Mining (WSDM)
- Other related conferences
  - DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, …
  - Web and IR conferences: WWW, SIGIR, WSDM
  - ML conferences: ICML, NIPS
  - PR conferences: CVPR,
- Journals
  - Data Mining and Knowledge Discovery (DAMI or DMKD)
  - IEEE Trans. On Knowledge and Data Eng. (TKDE)
  - KDD Explorations
  - ACM Trans. on KDD

Where to Find References? DBLP, CiteSeer, Google

- Data mining and KDD (SIGKDD: CDROM)
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
  - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
  - Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- AI & Machine Learning
  - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
  - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- Web and IR
  - Conferences: SIGIR, WWW, CIKM, etc.
  - Journals: WWW: Internet and Web Information Systems,
- Statistics
  - Conferences: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization
  - Conference proceedings: CHI, ACM-SIG Graph, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.

What kinds of data can be mined????

- As a general technology, data mining can be applied to any kind of data as long as the data are meaningful for a target application.
- The most basic forms of data for mining applications are: database data , data warehouse data, and transactional data.
- Data mining can also be applied to other forms of data repositories (e.g., data streams, ordered/sequence data, graph or networked data, spatial data, text data, multimedia data, and the WWW).
- However, algorithms and approaches may differ when applied to different types of data.

Flat files

- Flat files are actually the most common data source for data mining, especially at research level.
- Flat files are simple data files in text or binary format.
- The data in these files can be transactions, time series data, scientific measurements etc.

Relational databases

- Data mining algorithms using relational databases can be more versatile than data mining algorithms specifically written for flat files.
- Data mining can benefit from SQL for data selection.
- Relational databases are one of the most commonly available and richest information repositories

Data warehouse

- A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site.
- Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.
- To facilitate decision making, the data in a data warehouse are organized around major subjects (e.g., customer, item, supplier, and activity).
- The data are stored to provide information from a historical perspective, such as in the past 6 to 12 months, and are typically summarized.

Data warehouse Cond……

- A data warehouse is usually modeled by a multidimensional data structure, called a data cube.
- each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure such as count or sum (sales _amount).

Transactional Data

- A transaction database is a set of records representing transactions, each with a time stamp, an identifier and s et of items.
- Associated with transaction files could also be descriptive data for the items.
- Since relational databases do not allow nested tables, transactions are usually stored in flat files or stored in two normalized transaction tables, one for transaction and one for the transaction items.
- The typical data mining analysis on such data is called the **market basket analysis** or association rules in which associations between items occurring or in sequence.

Multimedia databases

- It includes video, images, audio and text media.
- They can be stored on object oriented databases or simply on a file system.
- multimedia is characterized by high dimensionality, which makes data mining even more challenging.
- It may require computer vision, computer graphics, image interpretation and natural language processing.

Spatial Databases

- In addition to usual data , it stores geographical data.
- Data like maps, and global regional positioning.
- It present new challenges to data mining algorithms.

Time series data

- It contains time related data.
- Like stock market data and logged activities.
- Data mining in such databases includes the study of trends and correlations between evaluations of different variables as well as a prediction of trends.

World Wide data

- It is the most heterogeneous and dynamic data repository.
- Data in the WWW is organized in interconnected documents.
- These documents can be audio, video, text etc.
- It is also called a web mining.

What is Data Mining??????

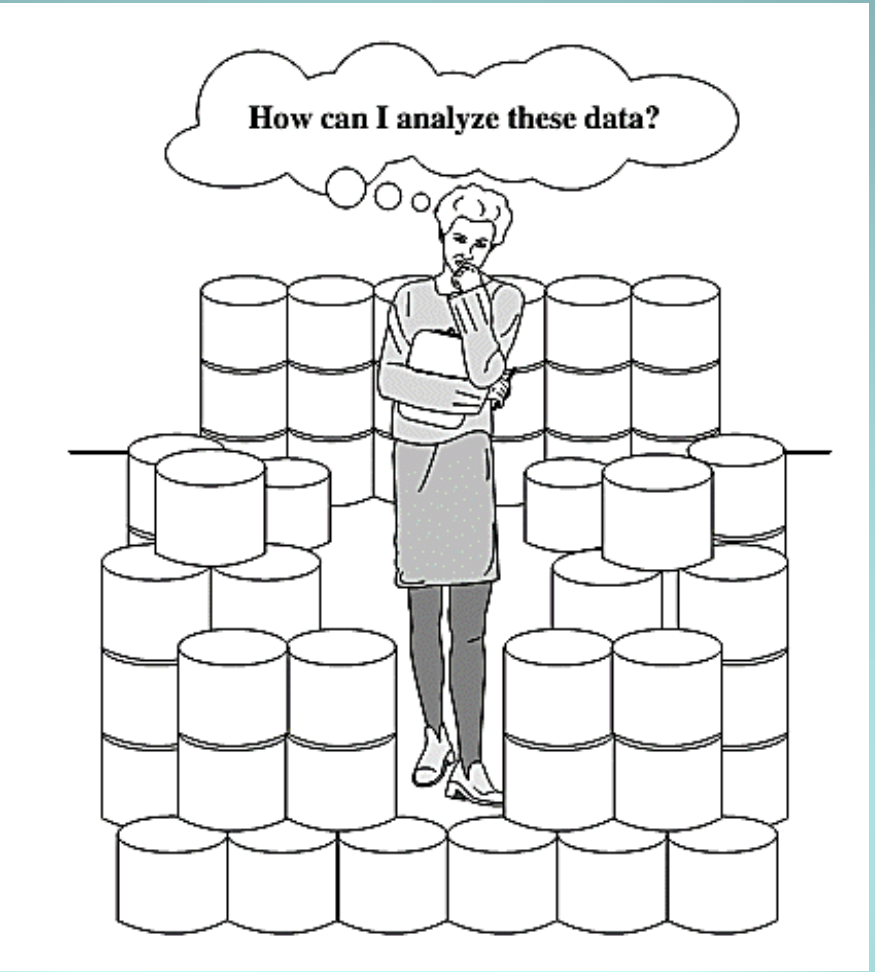
- It is also known as Knowledge Discovery from Data (KDD).

- We live in a world where vast amounts of data are collected daily. Analyzing such data is an important need.

"***We are living in the information age***" is a popular saying; however, we are actually "***living in the data age***".

- This explosive growth of available data volume is a result of the computerization of our society and the fast development of powerful data collection and storage tools.

The world is data rich but information poor



Data collected in large data repositories become “data tombs”. The data mining tools that can turn data tombs into “golden nuggets” of knowledge. Golden nuggets means “**small but valuable facts**”.

Data mining is also called as *knowledge mining from data, knowledge extracti haeology, and data dredging.*

The knowledge discovery process

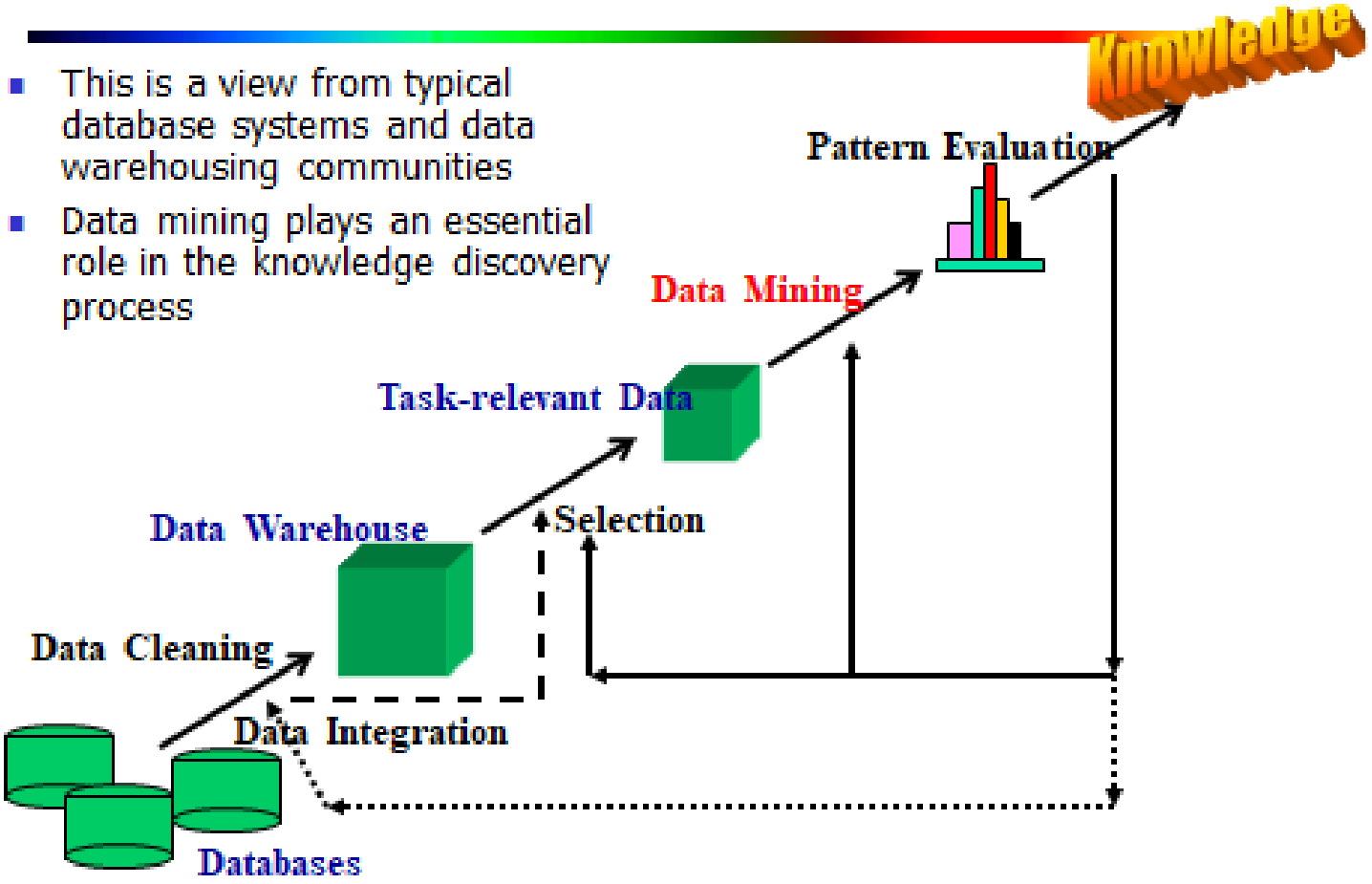
It is an iterative sequence of the following steps:

- **Data cleaning** (to remove noise and inconsistent data)
- **Data integration** (where multiple data sources may be combined)
- **Data selection** (where data relevant to the analysis task are retrieved from the database)
- **Data transformation** (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)

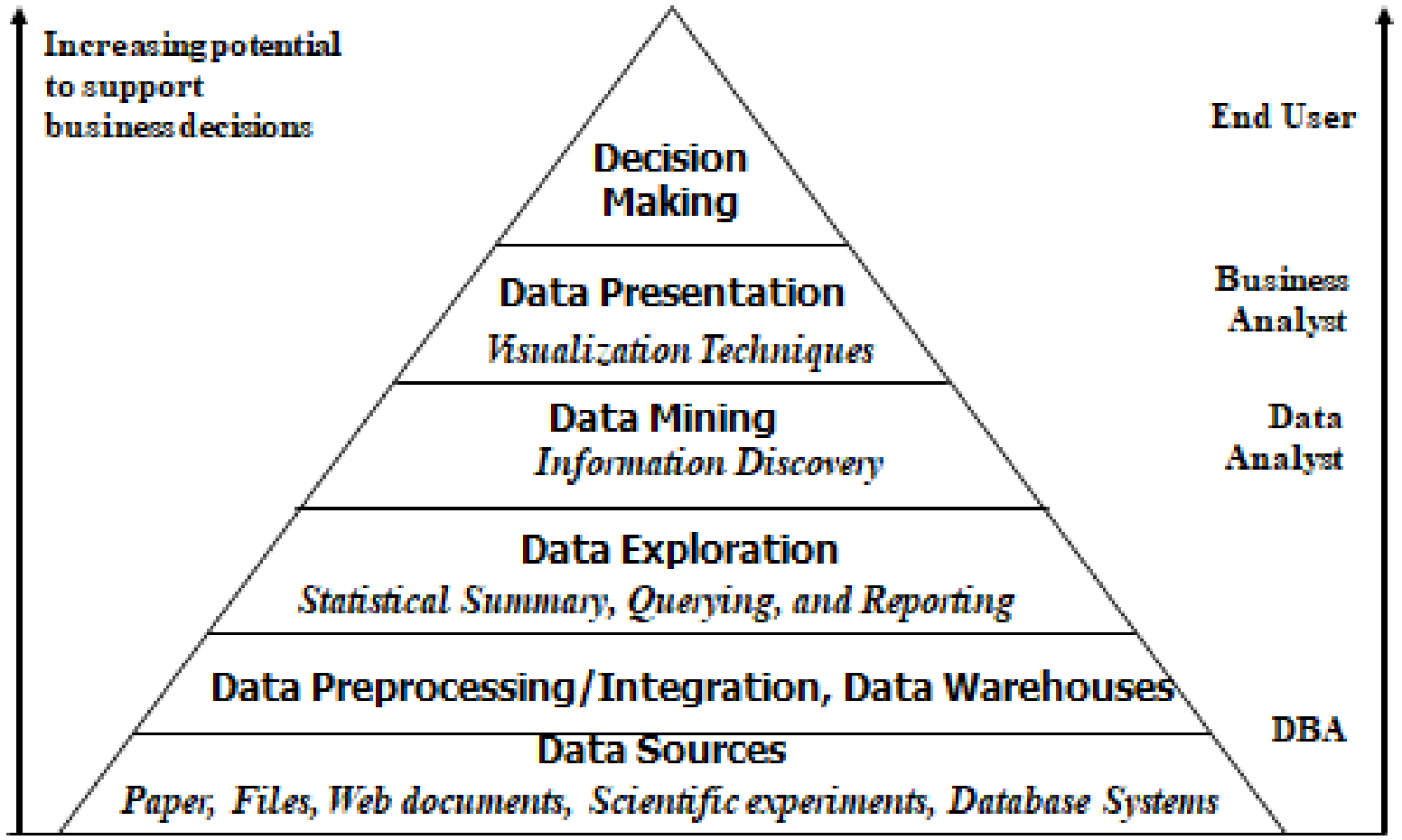
- **Data mining** (an essential process where intelligent methods are applied to extract data patterns)
- **Pattern evaluation** (to identify the truly interesting patterns representing knowledge based on interestingness measures)
- **Knowledge presentation** (where visualization and knowledge representation techniques are used to present mined knowledge to users)

Knowledge Discovery (KDD) Process

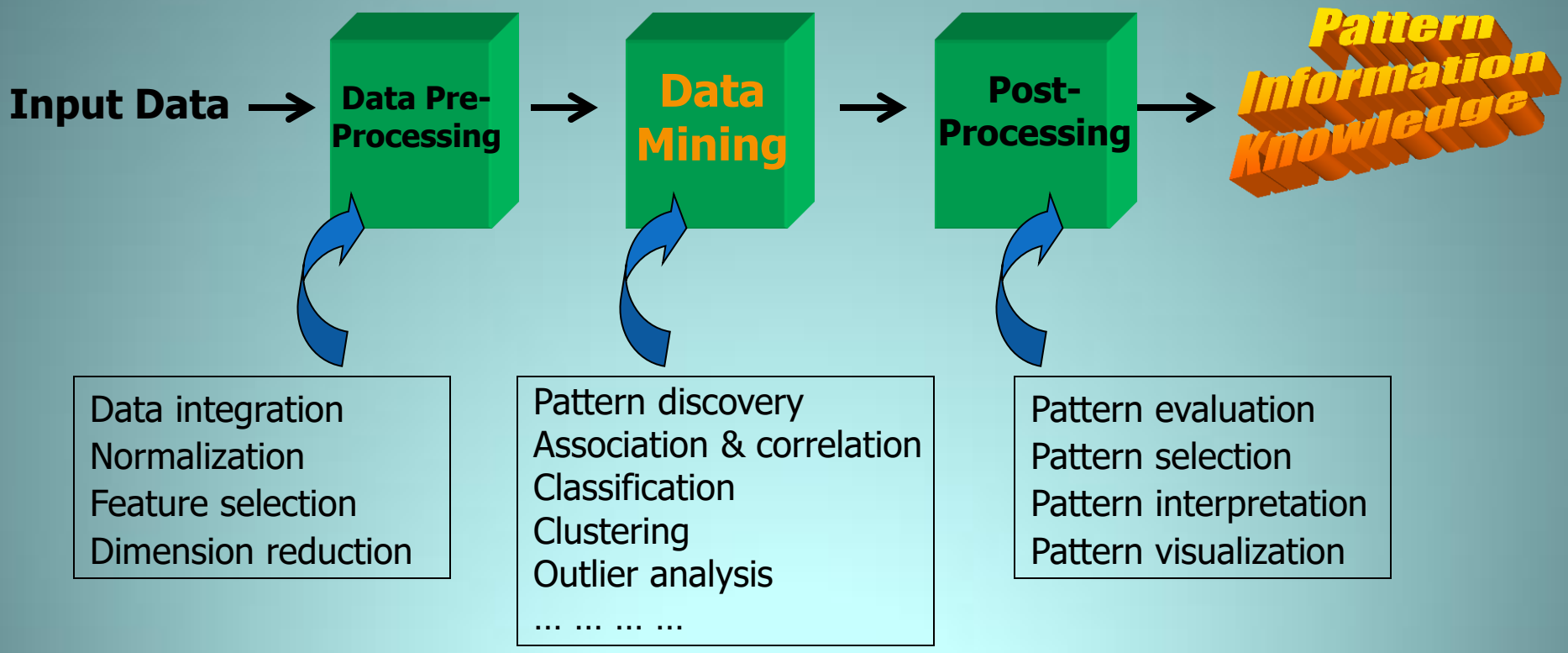
- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



Data Mining in Business Intelligence



KDD Process: A Typical View from ML and Statistics



- This is a view from typical machine learning and statistics communities

DATA MINING FUNCTIONALITIES

Data Mining Functionalities

Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks.

- **Class/Concept Description: Characterization and Discrimination:**

Data entries can be associated with classes or concepts.

For example, in the All Electronics store, classes of items for sale include computers and printers, and concepts of customers include big Spenders and budget Spenders. Such descriptions of a class or a concept are called class/concept descriptions.

Data characterization

- It is a summarization of the general characteristics or features of a target class of data.
- The data corresponding to the user-specified class are typically collected by a query.
- For example, to study the characteristics of software products with sales that increased by 10% in the previous year, the data related to such products can be collected by executing an SQL query on the sales database.
- There are several methods for effective data summarization or characterization. The data cube-based OLAP roll-up operation can be used to perform user-controlled data summarization along a specified dimension.
- The output of data characterization can be presented in various forms. Examples include **pie charts, bar charts, curves, multidimensional data cubes, and multidimensional tables, including crosstabs.**

Data discrimination

- Data discrimination is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.
- The target and contrasting classes can be specified by a user, and the corresponding data objects can be retrieved through database queries.
- For example, a user may want to compare the general features of software products with sales that increased by 10% last year against those with sales that decreased by at least 30% during the same period.

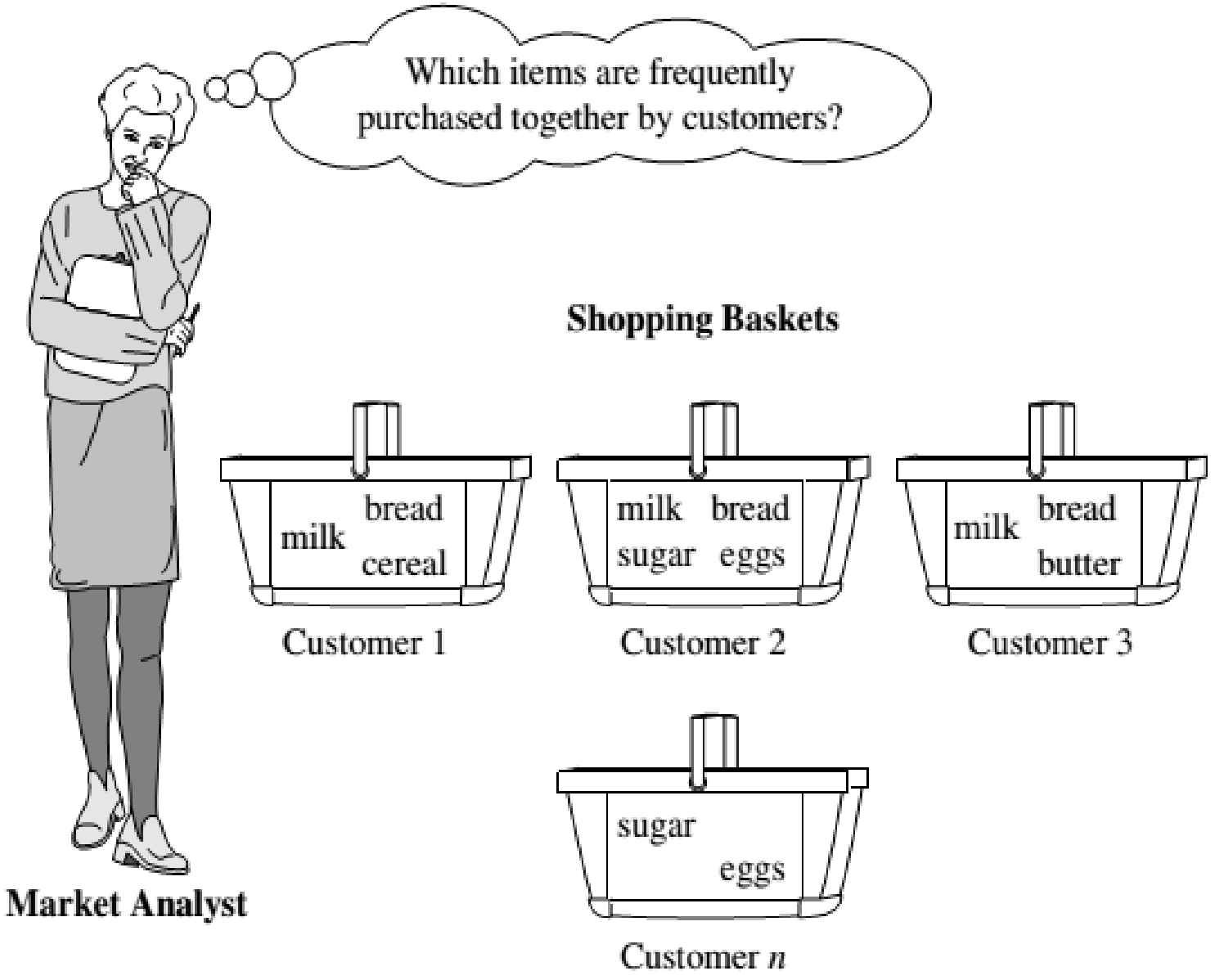
CONCEPT OF INTERESTING PATTERNS

Imagine that you are a sales manager, and you are talking to a customer who recently bought a PC and a digital camera from the store. What should you recommend to her next?

Frequent patterns and association rules are the knowledge that you want to mine in such a scenario.

Basic concepts

- Frequent patterns are patterns (e.g., itemsets, or subsequences) that appear frequently in a data set.
- For example, a set of items, such as **milk and bread**, that appear frequently together in a transaction data set is a frequent itemset.
- A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a (frequent) sequential pattern.
- Frequent pattern mining searches for recurring relationships in a given data set.


Which items are frequently purchased together by customers?
Shopping Baskets
milk
bread
cereal
Customer 1
milk
bread
sugar
eggs
Customer 2
milk
bread
butter
Customer 3
sugar
eggs
Customer n
Market Analyst

Market Basket Analysis: A Motivating Example

- If customers who purchase computers also tend to buy antivirus software at the same time, then placing the hardware display close to the software display may help increase the sales of both items.
- Market basket analysis can also help retailers plan which items to put on sale at reduced prices. If customers tend to purchase computers and printers together, then having a sale on printers may encourage the sale of printers as well as computers.

Association rules

- If we think of the universe as the set of items available at the store, then each item has a Boolean variable representing the presence or absence of that item.
- Each basket can then be represented by a Boolean vector of values assigned to these variables.
- The Boolean vectors can be analyzed for buying patterns that reflect items that are frequently associated or purchased together.
- These patterns can be represented in the form of **ASSOCIATION RULES**.

- For example, the information that customers who purchase computers also tend to buy antivirus software at the same time is represented in the following association rule:

$$computer \Rightarrow antivirus\_software\ [support = 2\%, confidence = 60\%].$$

- A support of 2% for Rule means that 2% of all the transactions under analysis show that computer and antivirus software are purchased together.
- A confidence of 60% means that 60% of the customers who purchased a computer also bought the software.

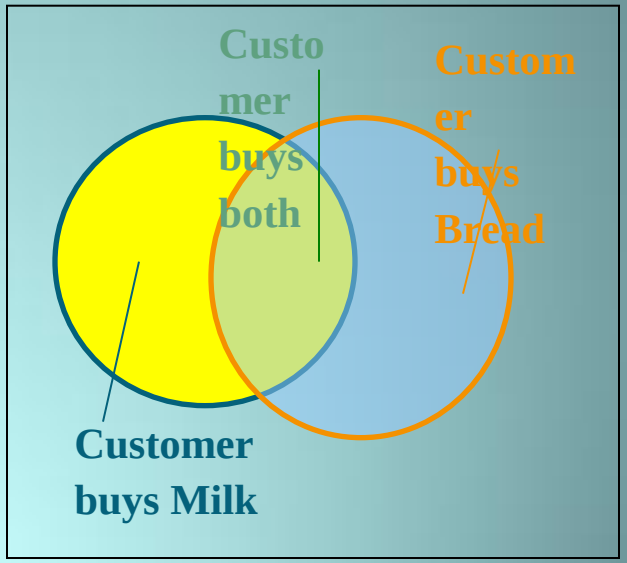
Association rules

- Typically, association rules are considered interesting if they satisfy both a **minimum support threshold** and a **minimum confidence threshold**. These thresholds can be a set by users or domain experts.

Frequent Patterns and Association Rules

Transaction-id	Items bought
10	A, B, D
20	A, C, D
30	A, D, E
40	B, E, F
50	B, C, D, E, F

- Itemset $X = \{x_1, \ldots, x_k\}$
- Find all the rules $X \rightarrow Y$ with minimum support and confidence
  - **support**, *s*, probability that a transaction contains $X \cup Y$.
  - **confidence**, *c*, conditional probability that a transaction having X also contains *Y*.



Let $sup_{min} = 50\%$, $conf_{min} = 50\%$

Freq. Pat.: {A:3, B:3, D:4, E:3, AD:3}

Association rules:

$A \rightarrow D$ (60%, 100%)

$D \rightarrow A$ (60%, 75%)

Frequent Patterns and Association Rules

$$support(A \Rightarrow B) = P(A \cup B)$$
$$confidence(A \Rightarrow B) = P(B|A).$$

- Rules that satisfy both a minimum support threshold (min_sup) and a minimum confidence threshold (min_conf) are called strong.
- A set of items is referred to as an itemset. An itemset that contains k items is a k-itemset. The set **{computer, antivirus software}** is a 2-itemset.
- The occurrence frequency of an itemset is the number of transactions that contain the itemset.
- This is also known, simply, as the **frequency, support count, or count** of the itemset.

Frequent Patterns and Association Rules

$$confidence(A \Rightarrow B) = P(B|A) = \frac{support(A \cup B)}{support(A)} = \frac{support\_count(A \cup B)}{support\_count(A)}.$$

- In general, association rule mining can be viewed as a two-step process:
  - **Find all frequent itemsets:** By definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count, min sup.
  - **Generate strong association rules from the frequent itemsets:** By definition, these rules must satisfy minimum support and minimum confidence.

Association analysis

$$buys(X, \text{“}computer\text{”}) \Rightarrow buys(X, \text{“}software\text{”}) \; [support = 1\%, confidence = 50\%],$$

- Where X is a variable representing a customer.
- A confidence, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that he/she will buy software as well.
- A 1% support means that 1% of all the transactions under analysis show that computer and software are purchased together.

$$age(X, \text{“}20..29\text{”}) \wedge income(X, \text{“}40K..49K\text{”}) \Rightarrow buys(X, \text{“}laptop\text{”})$$
$$[support = 2\%, confidence = 60\%].$$

Classification and Regression for Predictive Analysis

- Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts.

Eg : male or female

- The derived model is based on the analysis of a set of training data (i.e., data objects for which the class labels are known).

- The model is used to predict the class label of objects for which the class label is unknown.

- Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well.

“How is the derived model presented?”

The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks

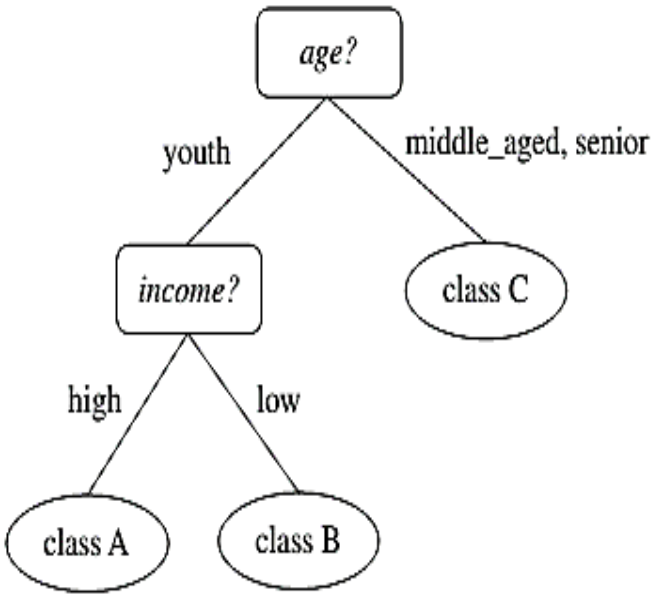
age(X, "youth") AND income(X, "high") ⟶ *class(X, "A")*

age(X, "youth") AND income(X, "low") ⟶ *class(X, "B")*

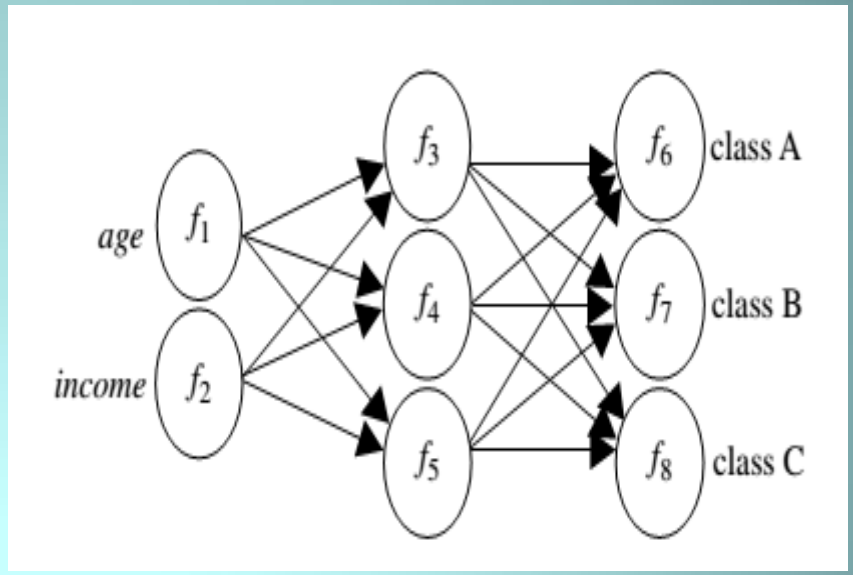
age(X, "middle_aged") ⟶ *class(X, "C")*

age(X, "senior") ⟶ *class(X, "C")*

IF - THEN Rules



Decision Tree



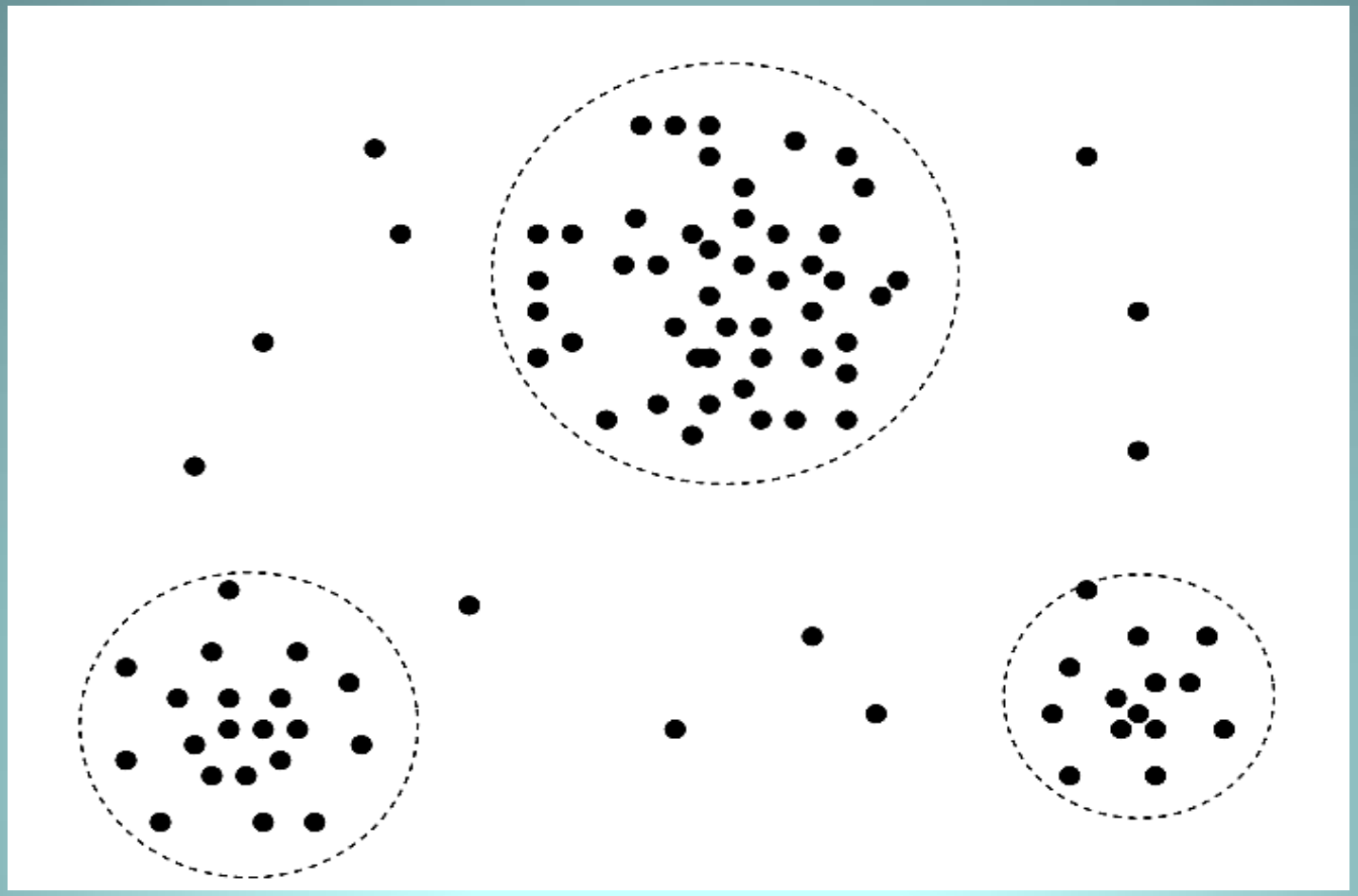
Neural Net

Cluster analysis

- Unlike classification and prediction, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label.
- In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels.
- The objects are clustered or grouped based on the principle of

"maximizing the intra-class similarity and minimizing the interclass similarity"

- That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are rather dissimilar to objects in other clusters.



Outlier Analysis

- Data objects that do not match with the general behavior or model of the data. Most analysis discard outliers as noise or exceptions.

- Outliers may be detected using statistical tests, or using distance measures where objects that are a substantial distance from any other cluster are considered outliers.

- **Example:** outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of extremely large amounts for a given account number in comparison to regular charges incurred by the same account

Are All Patterns Interesting?

- Only a small fraction of the patterns potentially generated would actually be of interest to any given user.
- a pattern is interesting if it is:
  - easily understood by humans
  - valid on new or test data with some degree of certainty
  - potentially useful
  - Novel (New)

DATA MINING TASKS

Summarization

Classification

Association

Clustering

Trend Analysis

Data Mining Function: (1) Generalization (Summarization)

- Information integration and data warehouse construction
  - Data cleaning, transformation, integration, and multidimensional data model
- Data cube technology
  - Scalable methods for computing (i.e., materializing) multidimensional aggregates
  - OLAP (online analytical processing)
- Multidimensional concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet region

Data Mining Function: (2) Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association, correlation vs. causality
  - A typical association rule
    - Diaper → Beer [0.5%, 75%] (support, confidence)
  - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

Data Mining Function: (3) Classification

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - 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 class labels
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …

Data Mining Function: (4) Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications

Data Mining Function: (5) Outlier Analysis

- Outlier analysis
  - Outlier: A data object that does not comply with the general behavior of the data
  - Noise or exception? — One person's garbage could be another person's treasure
  - Methods: by product of clustering or regression analysis, …
  - Useful in fraud detection, rare events analysis

Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

- Sequence, trend and evolution analysis
  - Trend, time-series, and deviation analysis: e.g., regression and value prediction
  - Sequential pattern mining
    - e.g., first buy digital camera, then buy large SD memory cards
  - Periodicity analysis
  - Motifs and biological sequence analysis
    - Approximate and consecutive motifs
  - Similarity-based analysis
- Mining data streams
  - Ordered, time-varying, potentially infinite, data streams

Structure and Network Analysis

- Graph mining
  - Finding frequent subgraphs (e.g., chemical compounds), trees (XML), substructures (web fragments)
- Information network analysis
  - Social networks: actors (objects, nodes) and relationships (edges)
    - e.g., author networks in CS, terrorist networks
  - Multiple heterogeneous networks
    - A person could be multiple information networks: friends, family, classmates, …
  - Links carry a lot of semantic information: Link mining
- Web mining
  - Web is a big information network: from PageRank to Google
  - Analysis of Web information networks
    - Web community discovery, opinion mining, usage mining, …

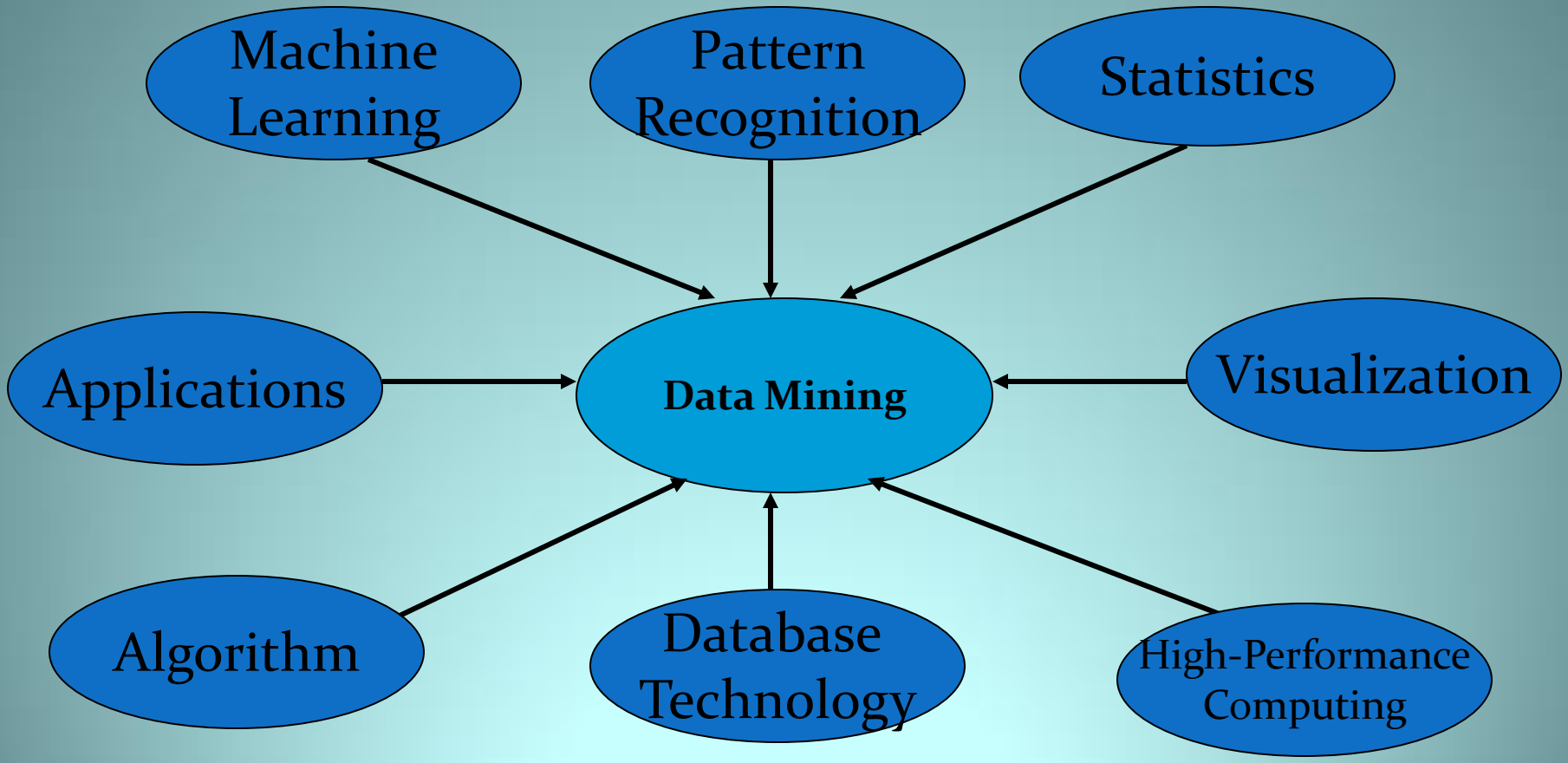
Evaluation of Knowledge

- Are all mined knowledge interesting?
  - One can mine tremendous amount of "patterns" and knowledge
  - Some may fit only certain dimension space (time, location, …)
  - Some may not be representative, may be transient, …
- Evaluation of mined knowledge → directly mine only interesting knowledge?
  - Descriptive vs. predictive
  - Coverage
  - Typicality vs. novelty
  - Accuracy
  - Timeliness
  - …

DATA MINING TECHNIQUES

- Statistical Approaches
- Machine Learning Approaches
- Database Oriented Approaches
- Other Approaches (Neural Network Approach)

Data Mining: Confluence of Multiple Disciplines



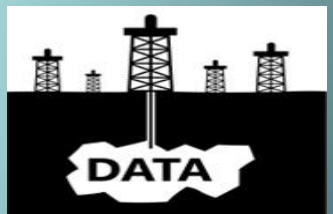
Why Confluence of Multiple Disciplines?

- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

CURRENT TRENDS

IMPORTANT FUTURE TRENDS IN DATA MINING

- Data mining is one of the most widely used methods to extract data from different sources and organize them for better usage.
- In spite of having different commercial systems for data mining, a lot of challenges come up when they are actually implemented.
- With rapid evolution in the field of data mining, companies are expected to stay abreast with all the new developments.



- Complex algorithms form the basis for data mining as they allow for data segmentation to identify various trends and patterns, detect variations, and predict the probabilities of various events happening.
- The raw data may come in both analog and digital format, and is inherently based on the source of the data.
- Companies need to keep track of the latest data mining trends and stay updated to do well in the industry and overcome challenging competition.

- Businesses which have been slow in adopting the process of data mining are now catching up with the others. Extracting important information through the process of data mining is widely used to make critical business decisions. In the coming decade, we can expect data mining to become as ubiquitous as some of the more prevalent technologies used today. Some of the key data mining trends for the future include

- **Multimedia Data Mining**
- This is one of the latest methods which is catching up because of the growing ability to capture useful data accurately. It involves the extraction of data from different kinds of multimedia sources such as audio, text, hypertext, video, images, etc. and the data is converted into a numerical representation in different formats. This method can be used in clustering and classifications, performing similarity checks, and also to identify associations.

- **Ubiquitous Data Mining**
- This method involves the mining of data from mobile devices to get information about individuals. In spite of having several challenges in this type such as complexity, privacy, cost, etc. this method has a lot of opportunities to be enormous in various industries especially in studying human-computer interactions.

- **Distributed Data Mining**
- This type of data mining is gaining popularity as it involves the mining of huge amount of information stored in different company locations or at different organizations. Highly sophisticated algorithms are used to extract data from different locations and provide proper insights and reports based upon them.

- **Spatial and Geographic Data Mining**
- This is new trending type of data mining which includes extracting information from environmental, astronomical, and geographical data which also includes images taken from outer space. This type of data mining can reveal various aspects such as distance and topology which is mainly used in geographic information systems and other navigation applications.

- **Time Series and Sequence Data Mining**
- The primary application of this type of data mining is study of cyclical and seasonal trends. This practice is also helpful in analyzing even random events which occur outside the normal series of events. This method is mainly being use by retail companies to access customer's buying patterns and their behaviors.

- **. Data Mining Dominance In The Pharmaceutical And Health Care Industries**
- Both the pharmaceutical and health care industries have long been innovators in the category of data mining. In fact, the recent rapid development of coronavirus vaccines is directly attributed to advances in data mining techniques for pharmaceutical testing, more specifically — in signal detection during the clinical trial process for new drugs. In health care, specialized data mining techniques are being used to analyze DNA sequences for creating custom therapies, make better informed diagnoses, and more.

- Increasing Automation In Data Mining
- Earlier incarnations of data mining involved manual coding by specialists with a deep background in statistics and programming. Modern techniques are highly automated, with AI/ML replacing most of these previously manual processes for developing pattern-discovering algorithms. Today's data mining solutions typically integrate ML and big data stores to provide both advanced data management functionality alongside sophisticated data analysis techniques.

- . Embedded Data Mining
- Data mining features are increasingly finding their way into a myriad of enterprise software use cases, from sales forecasting in CRM SaaS platforms to cyber threat detection in intrusion detection/prevention systems. The embedding of data mining into vertical market software applications enables prediction capabilities for any number of industries and opens up new realms of possibilities for unique value creation.
- .

- Rise Of Spatial And Geographic Data Mining
- With the new space race currently underway, more focus than ever has been placed on data mining for a myriad of commercial space-related use cases: zero-gravity cancer research, spacecraft design/testing, and — appropriately enough — asteroid mining, among others.
- Back on Earth, spatial and geographic data mining have already become fixtures of life through geographic information system (GIS) offerings, such as GPS-powered navigation and Google Maps.

Data Mining Vendor Consolidation

- If history is any indication, significant product consolidation in the data mining space is imminent as larger database vendors acquire data mining tooling startups to augment their offerings with new features.
- The current, fragmented market and broad range of players in the data mining arena resembles the adjacent big data vendor landscape — one that continues to undergo consolidation

- Application areas of Data Mining
  - Business Application
  - Science Application

Business Application

- From Traditional Areas sucha s business and science , to new areas such as sports Data Mining is being used these days.
- Data Mining has been Successfully used database marketing, Retail Analysis, Stock Selection, Credit approval etc. and of course many mor.
- Mining historical consumer analysis, Pattern checking extracting customer profiles.
- Shopping transactions for sales campaign
- Credit and loan related information
-

In the area of Science

- Astronomy, Molecular biology, Medicine, Geology,
- As an example Jet propulsion Lab at California Institute of Technology has devloped a data mining system which can classify the sky objects such as stars in the satellite images.

Other Applications of Data Mining in trend

- Health Care Management
- Tax Fraud Detection
- Money Laundering Monitoring
- Sports
- E.g. Advanced Scout system developed by IBM has been used by coaches of more than a dozen teams in National Basketball Association to improve their game.

Example: Medical Data Mining

- Health care & medical data mining – often adopted such a view in statistics and machine learning
- Preprocessing of the data (including feature extraction and dimension reduction)
- Classification or/and clustering processes
- Post-processing for presentation

MAJOR ISSUES AND ETHICS IN DATA MINING

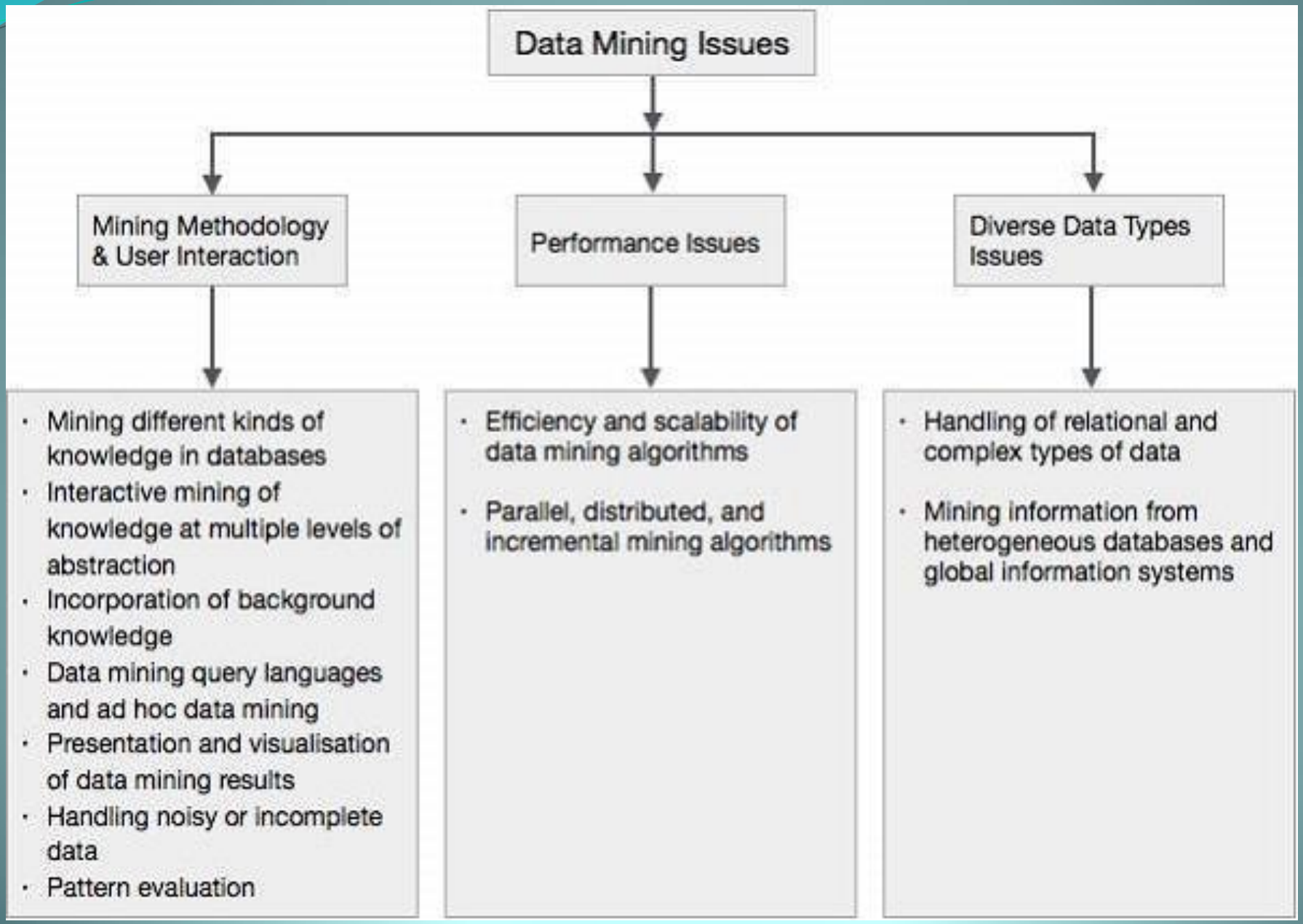
Major Issues in Data Mining

- Mining Methodology
  - Mining various and new kinds of knowledge
  - Mining knowledge in multi-dimensional space
  - Data mining: An interdisciplinary effort
  - Boosting the power of discovery in a networked environment
  - Handling noise, uncertainty, and incompleteness of data
  - Pattern evaluation and pattern- or constraint-guided mining
- User Interaction
  - Interactive mining
  - Incorporation of background knowledge
  - Presentation and visualization of data mining results

Major Issues in Data Mining

- Efficiency and Scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed, stream, and incremental mining methods
- Diversity of data types
  - Handling complex types of data
  - Mining dynamic, networked, and global data repositories
- Data mining and society
  - Social impacts of data mining
  - Privacy-preserving data mining
  - Invisible data mining

- Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources. These factors also create some issues. Here in this tutorial, we will discuss the major issues regarding –
- Mining Methodology and User Interaction
- Performance Issues
- Diverse Data Types Issues


Data Mining Issues
Mining Methodology & User Interaction
Performance Issues
Diverse Data Types 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 visualisation of data mining results
Handling noisy or incomplete data
Pattern evaluation
Efficiency and scalability of data mining algorithms
Parallel, distributed, and incremental mining algorithms
Handling of relational and complex types of data
Mining information from heterogeneous databases and global information systems

Mining Methodology and User Interaction Issues

It refers to the following kinds of issues –

Mining different kinds of knowledge in databases – Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.

Interactive mining of knowledge at multiple levels of abstraction – The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.

- **Incorporation of background knowledge** – To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.
- **Data mining query languages and ad hoc data mining** – Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.

- **Presentation and visualization of data mining results** – Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.
- **Handling noisy or incomplete data** – The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- **Pattern evaluation** – The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

- Performance Issues
- There can be performance-related issues such as follows –
- **Efficiency and scalability of data mining algorithms** – In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
- **Parallel, distributed, and incremental mining algorithms** – The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

- Diverse Data Types Issues
- **Handling of relational and complex types of data** – The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.
- **Mining information from heterogeneous databases and global information systems** – The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from them adds challenges to data mining.