



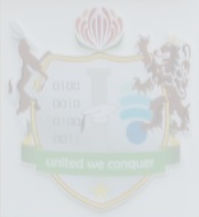
Faculty of
Information
Technology
**BELGIUM
CAMPUS**
ITVERSITY



Business Intelligence

G. Mudare

HAG
6L168




Data Mining

Introduction

- What is Data Mining
- What is Needed to Do Data Mining
- Business Data Mining
- Data Mining Tools

Abundance of data

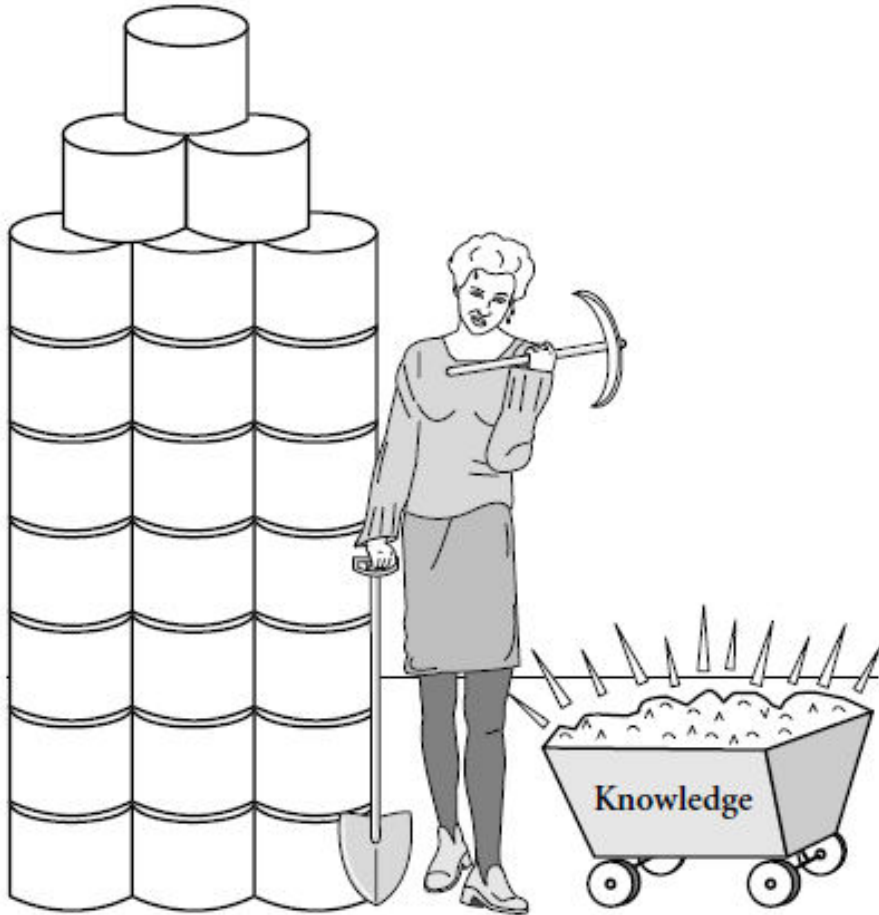


How can I analyze my data?

The abundance of data, coupled with the need for powerful data analysis tools, has been described as a *data rich but information poor* situation

What Is Data Mining?

Data mining or **KDD** refers to *extracting* or “*mining*” *knowledge* from *large amounts* of *data*.



- Data mining refers to the analysis of large quantities and observational data sets that are stored in computers to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.
- Data Mining, also popularly known as **Knowledge Discovery in Databases (KDD)**, refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases.
- Data mining refers to extracting or “mining” knowledge from large amounts of data

What Is Data Mining?

- With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making.

What Is Data Mining?

- Enormous amount of data stored in
 - files,
 - databases,
 - other repositories,
- Powerful means required for analysis and interpretation of such data and for the extraction of interesting knowledge that could help in decision-making.

- The definitions refer to "observational data," as opposed to "experimental data."
- Data mining typically deals with data that have already been collected for some purpose other than the data mining analysis
- Objectives of the data mining exercise play no role in the data collection strategy.
- In statistics, data are collected by using efficient strategies to answer specific questions.
- Data mining is actually part of the knowledge discovery process.
- The definition also mentions that the data sets examined in data mining are often large.

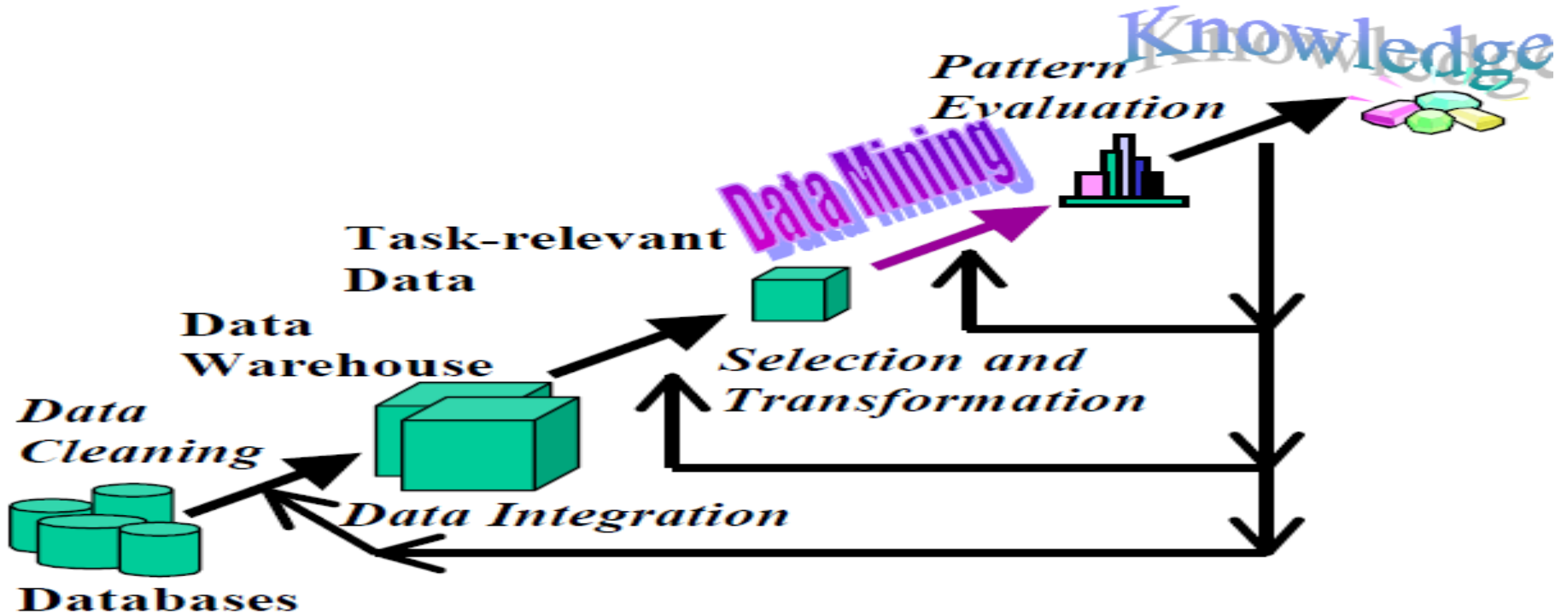
Problems That Arise With Large Amounts Of Data

- When we are faced with large bodies of data, new problems arise.
 - housekeeping issues of how to store or access the data,
 - how to determine the representativeness of the data
 - how to analyse the data in a reasonable period of time
 - how to decide whether an apparent relationship is merely a chance occurrence not reflecting any underlying reality.
- Often the available data comprise only a sample from the complete population (or, perhaps, from a hypothetical super population);
- the aim may be to generalize from the sample to the

Knowledge Discovery process

- Steps 1 to 4 are different forms of data pre-processing, where the data are prepared for mining.
- The data mining step may interact with the user or a knowledge base.
- The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base.
- Note that according to this view, data mining is only one step in the entire process, albeit an essential one because it uncovers hidden patterns for evaluation.

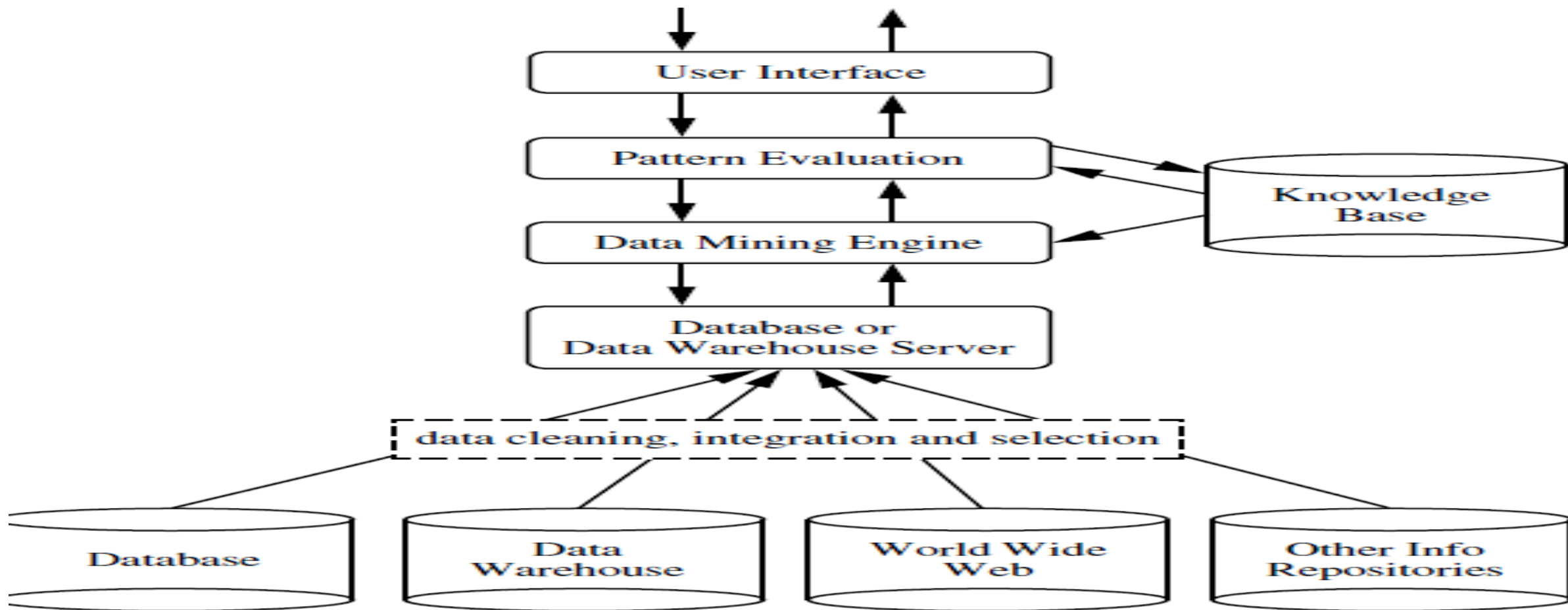
Data Mining is the core of Knowledge Discovery process



Major Components

- **Database, data warehouse, WorldWideWeb, or other information repository:**
- Data cleaning and data integration techniques may be performed on the data.
- **Database or data warehouse server:** The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.
- **Knowledge base:** This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns
- **Data mining engine:** This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.
- **Pattern evaluation module:** This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search toward interesting patterns.
- **User interface:** this component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

Architecture of a typical data mining system.



Data Mining—On What Kind of Data

- In principle, data mining should be applicable to any kind of data repository, as well as to transient data, such as data streams
- **Data repositories:**
 - relational databases,
 - data warehouses,
 - transactional databases,
 - advanced database systems (object Relational),
 - flat files,
 - data streams,
 - World Wide Web
 - **specific application-oriented databases**, such as spatial databases, time-series databases, text databases, and multimedia databases.

Data Mining Functionalities

What Kinds of Patterns Can Be Mined?

1. **Concept/Class Description: Characterization and Discrimination :**
summarizing the data of the class under study
2. **Mining Frequent Patterns, Associations, and Correlations:**
Patterns that occur frequently in data.
3. **Classification and Prediction**
use the model to predict the class of objects whose class label is unknown.
4. **Cluster Analysis:** organization of data in classes
5. **Outlier Analysis:** data elements that cannot be grouped in a given class or cluster. (*exceptions or surprises*)
6. **Evolution Analysis:** describes and models regularities or trends for objects whose behaviour changes over time



Faculty of
Information
Technology
**BELGIUM
CAMPUS**
ITVERSITY



Business Intelligence

G. Mudare

Data Mining Functionalities

Data Mining Functionalities

- Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks.
- In general, data mining tasks can be classified into two categories:
 1. **Descriptive** mining tasks characterize the general properties of the data in the database.
 2. **Predictive** mining tasks perform inference on the current data in order to make predictions.

Data mining functionalities

1. Concept/Class Description: Characterization and Discrimination
2. Mining Frequent Patterns, Associations, and Correlations
3. Classification and Prediction
- 4. Cluster Analysis**
5. Outlier Analysis
6. Evolution Analysis

Concept/Class Description

Characterization and Discrimination

- Data can be associated with classes or concepts.
- E.g., classes of items for sale include computers and printers, and concepts of customers include big Spenders and budget Spenders.
- 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)
 - (2) **Data discrimination**, by comparison of the target class with one or a set of comparative classes (often called the contrasting classes),
 - (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. collected by a database query.
- **Eg. to study the characteristics of software products whose sales increased by 10% in the last year,**
- Produce a description summarizing the characteristics of customers who spend more than \$1,000 a year at the electronics store. (Eg general profile of the customers, such as they are 40–50 years old, employed, and have excellent credit ratings.)
- System should allow users to drill down on any dimension, such as on occupation in order to view these customers according to their type of employment.

Data discrimination

- Comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes.
- Eg. compare the general features of software products whose sales increased by 10% in the last year with those whose sales decreased by at least 30% during the same period.
- A data mining system should be able to compare two groups of Electronics customers, such as those who shop for computer products regularly versus those who rarely shop for such products
 - Result could be a **comparative profile of the customers such as,**
 - **80% of the customers who frequently purchase computer products are between 20 and 40 years old and have a university education, whereas 60% of the customers who infrequently buy such products are either seniors or youths, and have no university degree.**
 - **Drilling down on a dimension, such as occupation, or adding new dimensions, such as income level, may help in finding even more discriminative features between the two classes**

Mining Frequent Patterns, Associations, and Correlations

- Frequent patterns, as the name suggests, are patterns that occur frequently in data.
- Kinds of frequent patterns:
 1. **Itemsets** → items that frequently appear together in a transactional data set, such as milk and bread
 2. **subsequences** → subsequence, such as the pattern that customers tend to purchase first a PC, followed by a digital camera, and then a memory card, is a (frequent) sequential pattern.
 3. **Substructures**. Refer to different structural forms, such as graphs, trees, or lattices, which may be combined with **itemsets** or **subsequences**.

Example

- Association analysis. Suppose, as a marketing manager of the Electronics shop, would like to determine which items are frequently purchased together within the same transactions.
- An example of such a rule, mined from the Electronics transactional database, is
- ***Buys(x; "computer") → buys(x; "software") [support = 1%; confidence = 50%]***
- ***"computer → software [1%, 50%]"***.
- 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 she will buy software as well.
- A **1% support** means that 1% of all of the transactions under analysis showed that computer and software were purchased together

Another Example

- Suppose, instead, that we are given the Electronics shop relational database relating to purchases.
- A data mining system may find association rules like
- *$age(X, "20:::29") \wedge income(X, 20K:::29K) \rightarrow buys(X, "CD player")$ [support = 2%, confidence = 60%]*
- The rule indicates that of the Electronics customers under study, 2% are 20 to 29 years of age with an income of 20,000 to 29,000 and have purchased a CD player. There is a 60% probability that a customer in this age and income group will purchase a CD player.
- *Association rules* are *discarded* as uninteresting *if they do not satisfy* both *a minimum support threshold and a minimum confidence threshold*.
- Additional analysis can be performed to uncover interesting statistical correlations between associated attribute-value pairs.

Classification and Prediction

- Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown.
- Derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known). “
- Derived model may be represented by
 - ❑ (IF-THEN) rules,
 - ❑ decision trees,
 - ❑ mathematical formulae,
 - ❑ neural networks
- Others methods include:
 - Naïve bayesian classification, support vector machines, and k-nearest neighbour classification.

IF-Then

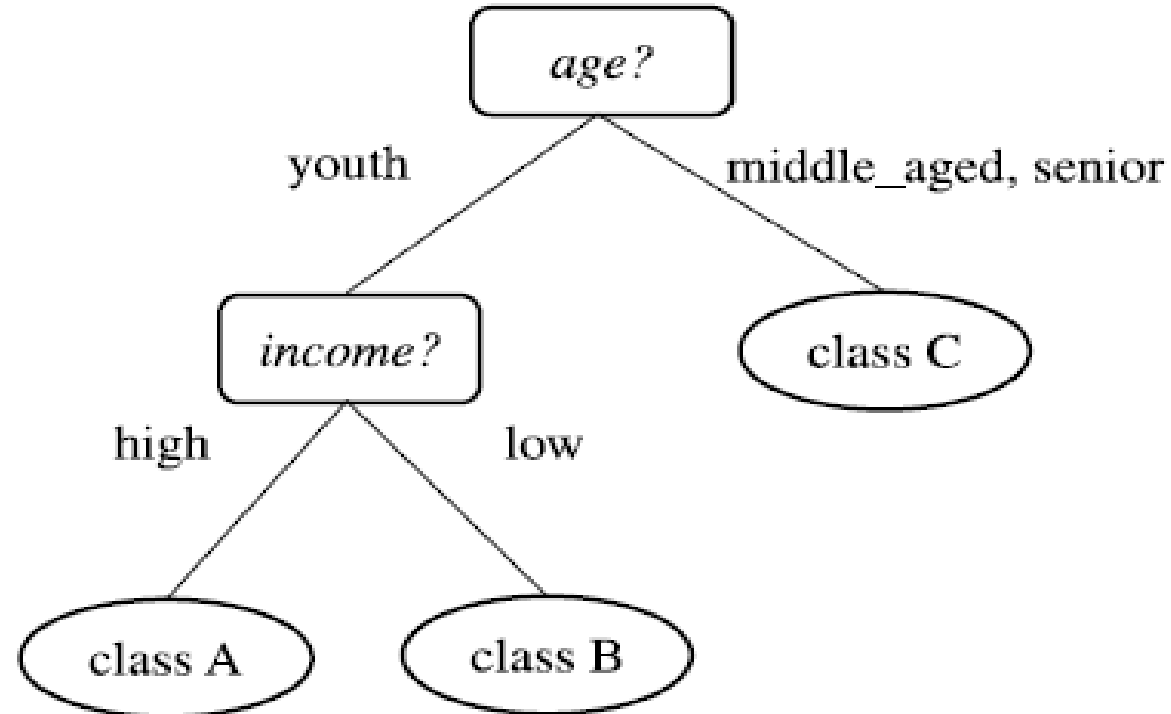
age(X, "youth") AND income(X, "high") \longrightarrow class(X, "A")

age(X, "youth") AND income(X, "low") \longrightarrow class(X, "B")

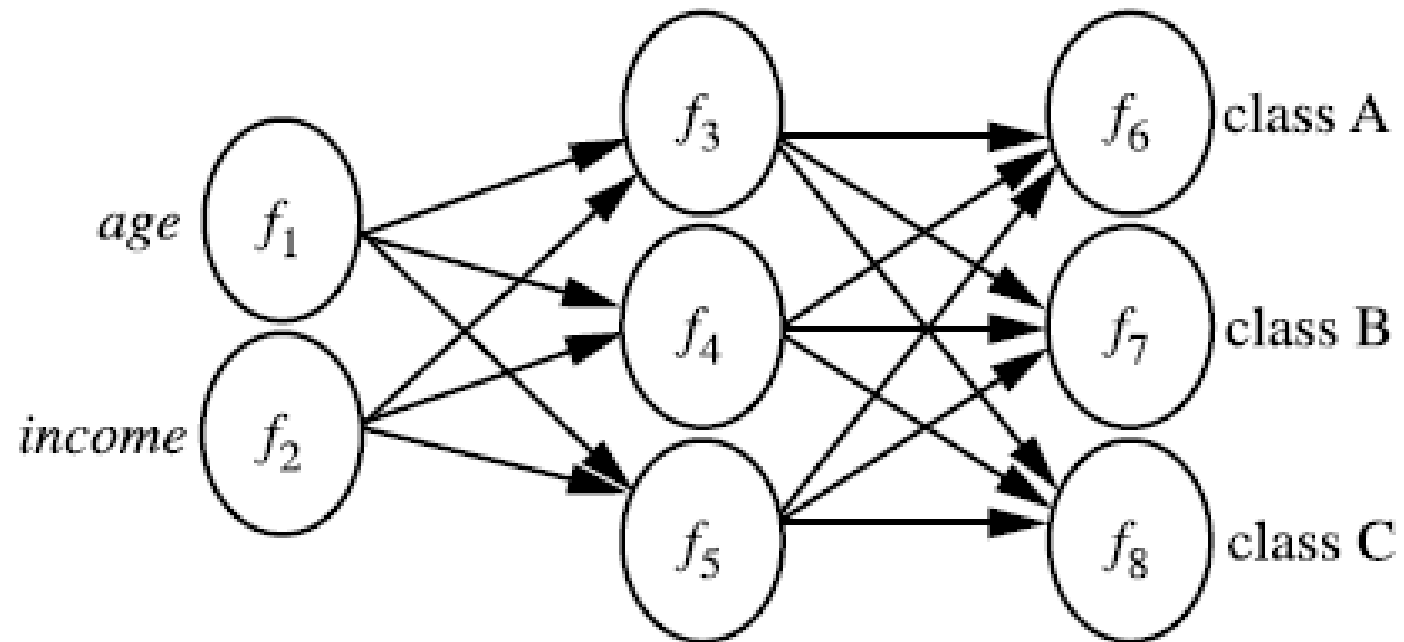
age(X, "middle_aged") \longrightarrow class(X, "C")

age(X, "senior") \longrightarrow class(X, "C")

Decision Tree



Neural Network



Prediction

- Whereas classification predicts **categorical (discrete, unordered)** labels, prediction models predict continuous-valued functions.
 - Predicting missing or unavailable numerical data values rather than class labels.
 - Types of predictions:
 - Predict some unavailable data values or pending trends,
 - Predict a class label for some data.
 - Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well,
- Classification** and **prediction** may need to be preceded by **relevance analysis**,
→ attempts to identify attributes that do not contribute to the classification or prediction process.

Example

- Suppose, as a sales manager of the Electronics shop, you would like to classify a large set of items in the store, based on three kinds of responses to a sales campaign: **good response**, **mild response**, and **no response**.
- You would like to derive a model for each of these three classes based on the descriptive features of the items, such as **price, brand, place made, type, and category**.
- The resulting classification distinguishes each class from the others, presenting an organized picture of the data set.
- **Resulting classification can be expressed in the form of a decision tree.**

Example

- The decision tree, may identify price as being the single factor that best distinguishes the three classes and may also reveal that, after price, other features that help further distinguish objects of each class from another include brand and place made.



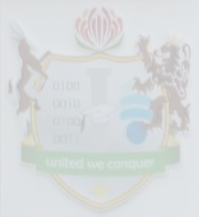
Faculty of
Information
Technology
**BELGIUM
CAMPUS**
ITVERSITY



Business Intelligence

G. Mudare

HAG
6L168



Data Mining Functionalities

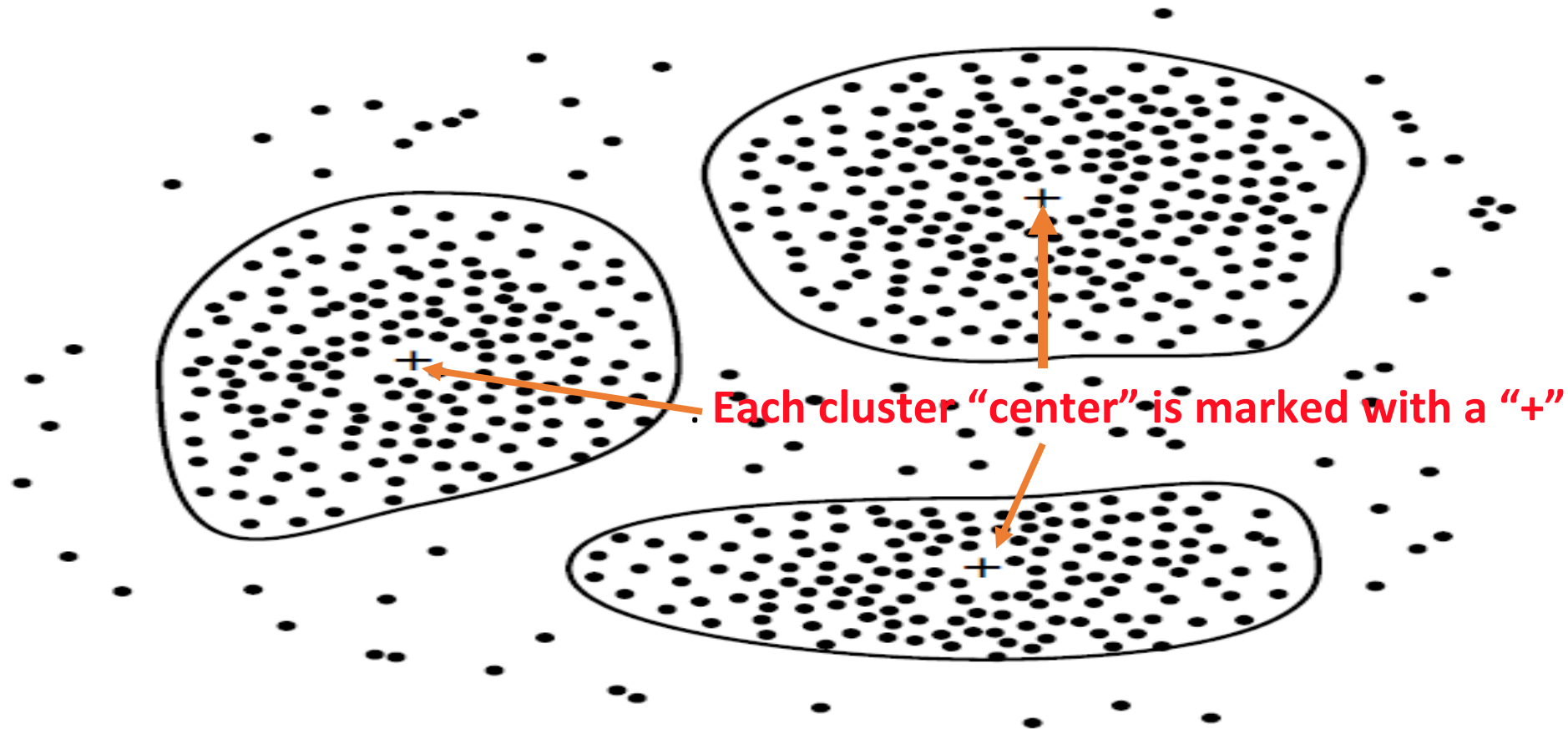
Clustering

- Similar to classification, clustering is the organization of data in classes.
- Unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes.
- Clustering is also called **unsupervised classification**, because the classification is not dictated by given class labels.
- Clustering approaches are based on the principle of
 - ❑ **Maximizing the similarity** between objects in the same class (**intra-class similarity**)
 - ❑ **minimizing the similarity** between objects of different classes (**inter-class similarity**).

Example

- **Cluster analysis.**
- Performed on Electronics customer data to identify homogeneous subpopulations of customers.
- To identify target groups for marketing

Customer data with respect to customer locations in a city,



Outlier analysis:

- Outliers are data elements that cannot be grouped in a given class or cluster. (*exceptions or surprises*)
- Although outliers can be considered noise and discarded in some applications, they can reveal important knowledge in other domains, and thus can be very significant and their analysis valuable. *eg In fraud detection*
- The analysis of outlier data is referred to as *outlier mining*.

Example

- **Outlier analysis.** 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.
- Outlier values may also be **detected with respect to the location and type of purchase, or the purchase frequency.**

Evolution Analysis

- Data evolution analysis describes and models regularities or trends for objects whose behaviour changes over time.
- May include **characterization, discrimination, association and correlation analysis, classification, prediction, or clustering of time related data,**
- Distinct features of analysis include:
 - Time-series data analysis,
 - Sequence or periodicity pattern matching,
 - similarity-based data analysis.

Example

- Suppose that you have the major stock market (time-series) data of the last several years available from the a Stock Exchange and you would like to invest in shares of high-tech industrial companies.
- A data mining study of stock exchange data may identify stock evolution regularities for overall stocks and for the stocks of particular companies.
- Such regularities may help predict future trends in stock market prices, contributing to your decision making regarding stock investments.

Are all of the Patterns interesting?

- A data mining system has the potential to generate thousands or even millions of patterns, but only a small fraction of the patterns potentially generated would actually be of interest to any given user.

What makes a pattern interesting?

- Easily understood by humans,
- Valid on new or test data with some degree of certainty,
- Potentially useful,
- Novel.
- A pattern is also interesting if it validates a hypothesis that the user sought to confirm.
- An interesting pattern represents knowledge.

- Identifying and measuring the interestingness of patterns and rules discovered, or to be discovered, is essential for the evaluation of the mined knowledge and the KDD process as a whole.
- **rule support**, of the form $X \rightarrow Y$ representing the percentage of transactions that the given rule satisfies taken to be the probability
 - $P(X \cup Y)$, where $XUY \rightarrow$ a transaction contains both X and Y , that is, the union of itemsets X and Y
 - **confidence**, which assesses the degree of certainty of the detected association.
 - This is conditional probability $P(Y/X)$, i.e. the probability that a transaction containing X also contains Y .

Measures of pattern interestingness

- , **support and confidence** are defined as

$$\text{support}(X \Rightarrow Y) = P(X \cup Y).$$

$$\text{confidence}(X \Rightarrow Y) = P(Y|X).$$

- Each interestingness measure is associated with a threshold, which may be controlled by the user.
- For example, rules that do not satisfy a confidence threshold of, say, 50% can be considered uninteresting.
- Rules below the threshold likely reflect noise, exceptions, or minority cases and are probably of less value.

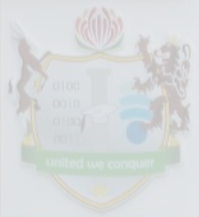


Faculty of
Information
Technology
**BELGIUM
CAMPUS**
ITVERSITY

Business Intelligence

G. Mudare

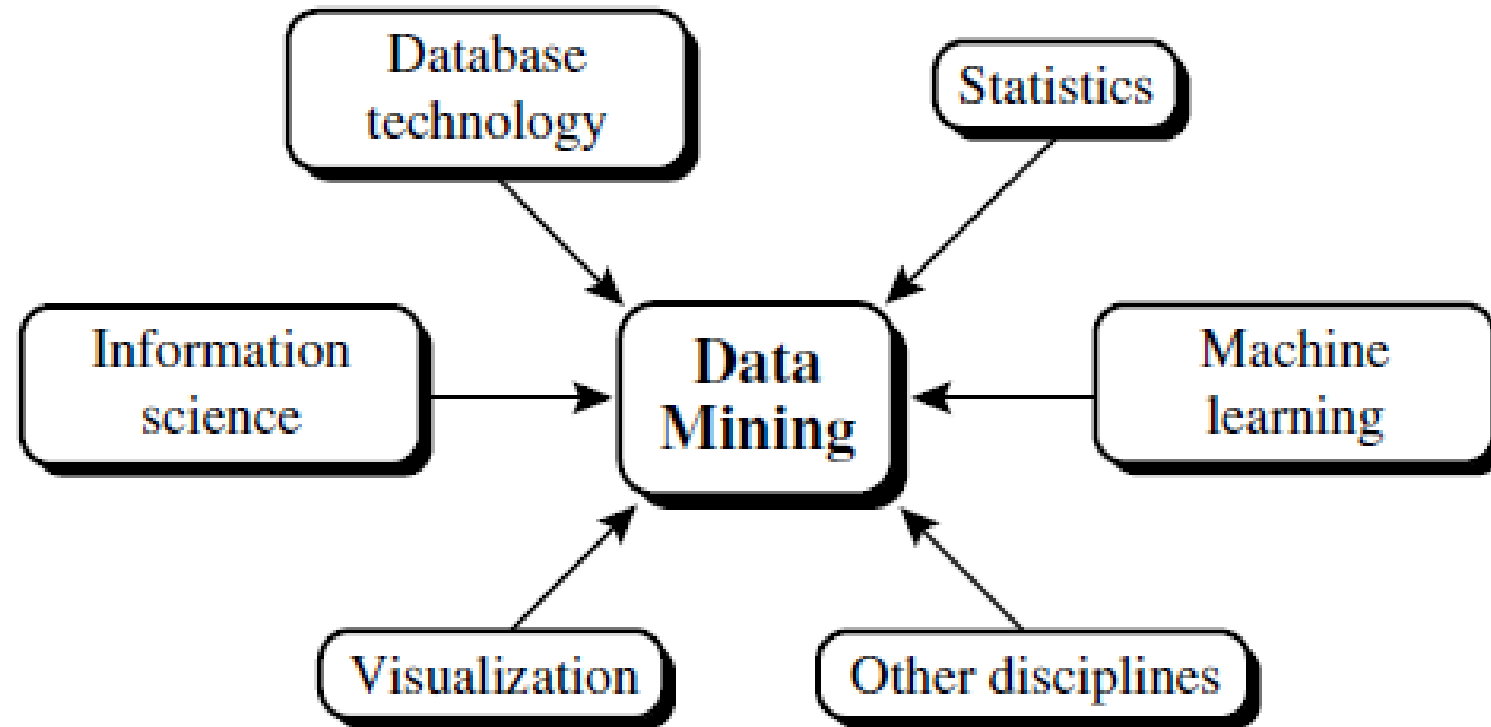
HAG
6L168



Data Mining Systems

Data Mining Systems

- Data mining is an interdisciplinary field, the confluence of a set of disciplines,



Categorizing data mining systems

- **Classification according to the type of data source mined:** relational, transactional, object-relational, or data warehouse mining system. With data such as spatial, time-series, text, stream data, multimedia data mining system, or a WorldWideWeb mining system
 - **Classification according to the kind of knowledge discovered:**
 - characterization, discrimination, association and correlation analysis, classification, prediction, clustering, outlier analysis, and evolution analysis.
 - **Classification according to mining techniques used or methods of data analysis employed**
 - (database-oriented or data warehouse-oriented techniques, machine learning): autonomous systems, interactive exploratory systems, query-driven systems
 - **Classification according to the applications adapted:**
 - Finance, telecommunications, DNA, stock markets, e-mail,
- ❑ A generic, all-purpose data mining system may not fit domain-specific mining tasks.

Data Mining Task Primitives

- A user normally has a data mining task in mind, i.e some form of data analysis that he or she would like to have performed
- A data mining task is in the form of a Data Mining Query defined in terms of data **mining task primitives**.
- **Data mining task primitives** allow the user to *interactively* communicate with the data mining system during discovery in order to direct the mining process, or examine the findings from different angles or depths

What is specified by Data Mining Task Primitives

- **The set of task-relevant data to be mined:** (*relevant attributes or dimensions*)
- **The kind of knowledge to be mined:** data mining functions to be performed, (*characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis, or evolution analysis.*)
- **The background knowledge to be used in the discovery process:** knowledge about the domain to be mined. (*Concept hierarchies eg age*)
- **The interestingness measures and thresholds for pattern evaluation:** evaluate the discovered patterns. (*support and confidence*)
- **The expected representation for visualizing the discovered patterns:** How patterns are to be displayed,, (*tables, charts, graphs, decision trees, and cubes.*)

Primitives for specifying a data mining task

- **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

Primitives for specifying a data mining task.

- **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

DATA MINING QUERY LANGUAGE

- A **data mining query language (DMQL)** can be designed to incorporate the primitives, allowing users to flexibly interact with data mining system
 - provides a foundation on which user-friendly graphical interfaces can be built.
 - facilitates a data mining system's communication with other information systems and its integration with the overall information processing environment.

Why is Designing a comprehensive data mining language is challenging?

- Each task has different requirements.
- The design of an effective data mining query language requires a deep understanding of the power, limitation, and underlying mechanisms of the various kinds of data mining tasks.

Major Issues in Data Mining

- Three Categorised exist:
 1. **Mining methodology and user interaction issues:** reflect the kinds of knowledge mined, the ability to mine knowledge at multiple granularities, the use of domain knowledge, ad hoc mining, and knowledge visualization.
 2. **Performance issues:** These include efficiency, scalability, and parallelization of data mining algorithms.
 3. **Issues relating to the diversity of database types:** *Handling complex types of data:*

Major Issues in Data Mining

1. Mining methodology and user interaction issues:
 - Mining different kinds of knowledge in database:
 - *Interactive mining of knowledge at multiple levels of abstraction:*
 - *Incorporation of background knowledge:*
2. Mining methodology and user interaction issues
 - *Handling noisy or incomplete data:*
 - *Pattern evaluation—the interestingness problem*
3. Performance issues
 - *Efficiency and scalability of data mining algorithms:*
 - *Parallel, distributed, and incremental mining algorithms:*

Major Issues in Data Mining

- Diversity of database types
 - Handling of relational and complex types of data
 - *Mining information from heterogeneous databases and global information systems*