Data Mining (DM)-Chapter 1

Why Data Mining?

What is Data Mining?

Applications

KDD Vs DM

Database Vs DM

Other Related Info

By Anupama Kota
Computer Science Dept

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 Data Mining?

Credit ratings/targeted marketing:

Given a database of 100,000 names, which persons are the least likely to default on their credit cards?

Identify likely responders to sales promotions

Fraud detection

Which types of transactions are likely to be fraudulent, given the demographics and transactional history of a particular customer?

Customer relationship management:

Which of my customers are likely to be the most loyal, and which are most likely to leave for a competitor? :

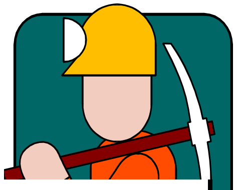
What Is Data Mining?

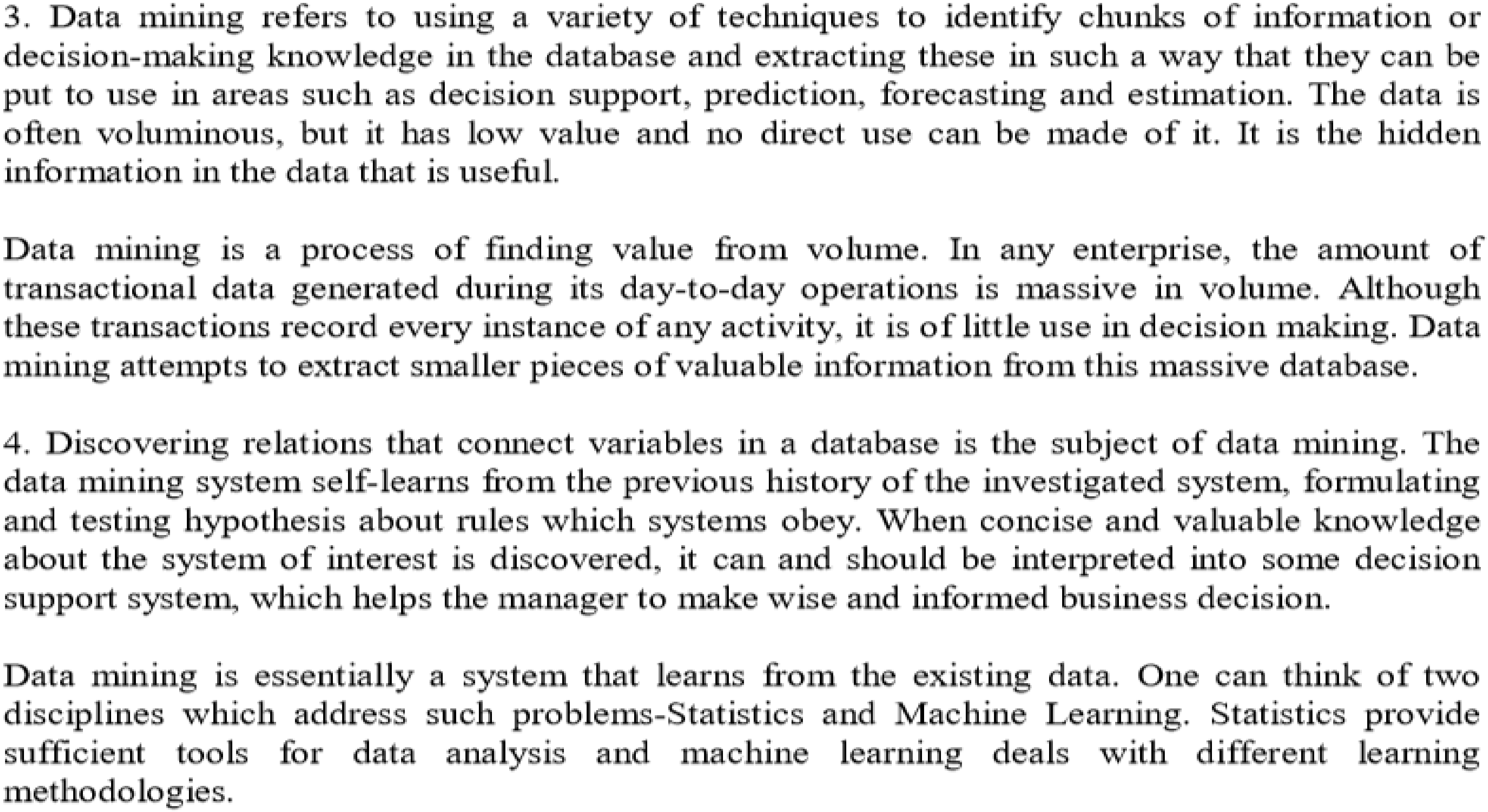
- Data Mining is defined as the procedure of extracting information from huge sets of data. In other words, we can say that data mining is mining knowledge from data.
- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) understandable patterns or knowledge from huge amount of data
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

What Is Data Mining?

- Process of semi-automatically analyzing large databases to find patterns that are:
  - valid: hold on new data with some certainty
  - novel: non-obvious to the system
  - useful: should be possible to act on the item
  - understandable: humans should be able to interpret the pattern

What Is Data Mining?





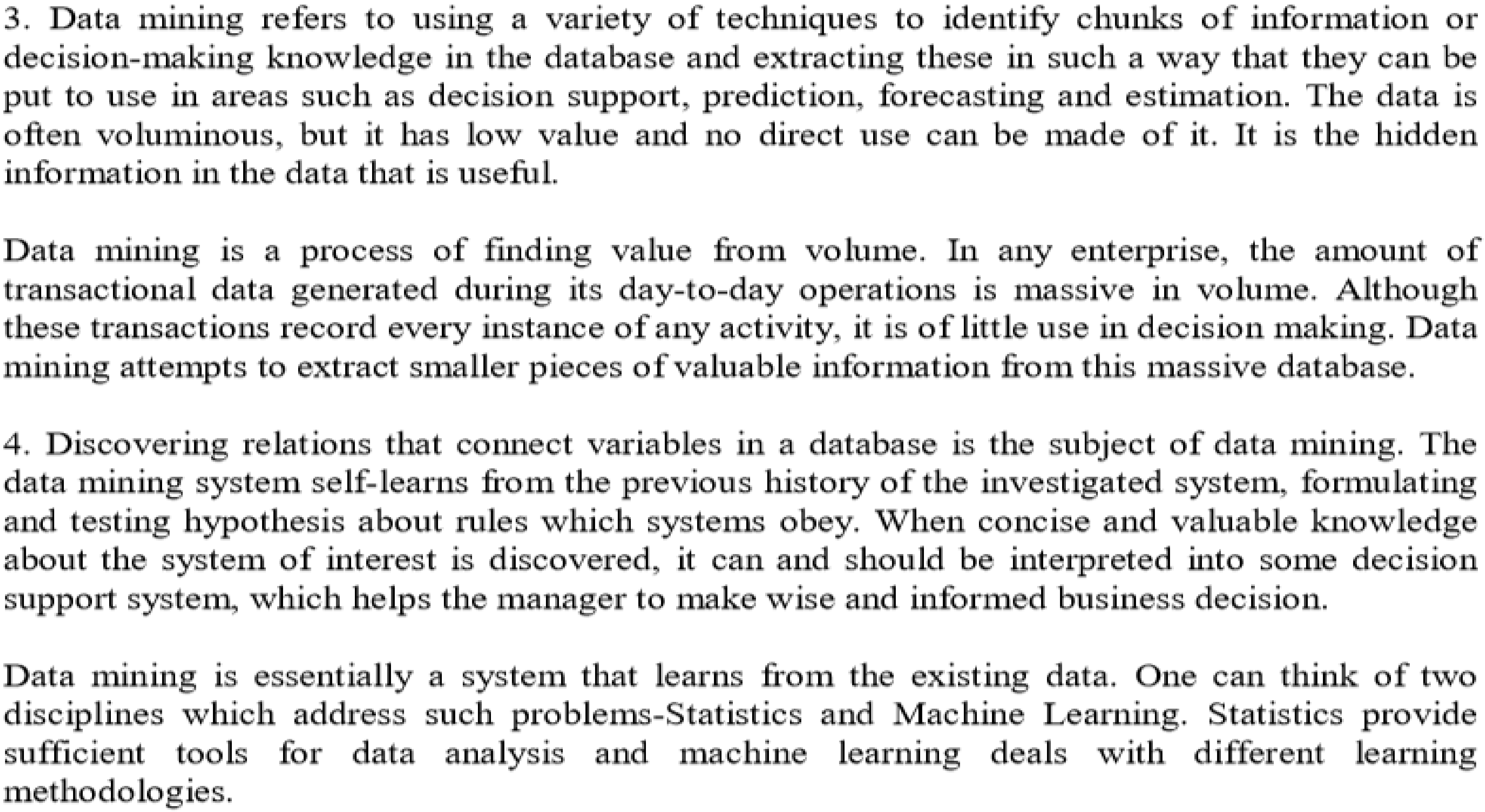
3. Data mining refers to using a variety of techniques to identify chunks of information or decision-making knowledge in the database and extracting these in such a way that they can be put to use in areas such as decision support, prediction, forecasting and estimation. The data is often voluminous, but it has low value and no direct use can be made of it. It is the hidden information in the data that is useful.

Data mining is a process of finding value from volume. In any enterprise, the amount of transactional data generated during its day-to-day operations is massive in volume. Although these transactions record every instance of any activity, it is of little use in decision making. Data mining attempts to extract smaller pieces of valuable information from this massive database.

4. Discovering relations that connect variables in a database is the subject of data mining. The data mining system self-learns from the previous history of the investigated system, formulating and testing hypothesis about rules which systems obey. When concise and valuable knowledge about the system of interest is discovered, it can and should be interpreted into some decision support system, which helps the manager to make wise and informed business decision.

Data mining is essentially a system that learns from the existing data. One can think of two disciplines which address such problems-Statistics and Machine Learning. Statistics provide sufficient tools for data analysis and machine learning deals with different learning methodologies.

What Is Data Mining?



3. Data mining refers to using a variety of techniques to identify chunks of information or decision-making knowledge in the database and extracting these in such a way that they can be put to use in areas such as decision support, prediction, forecasting and estimation. The data is often voluminous, but it has low value and no direct use can be made of it. It is the hidden information in the data that is useful.

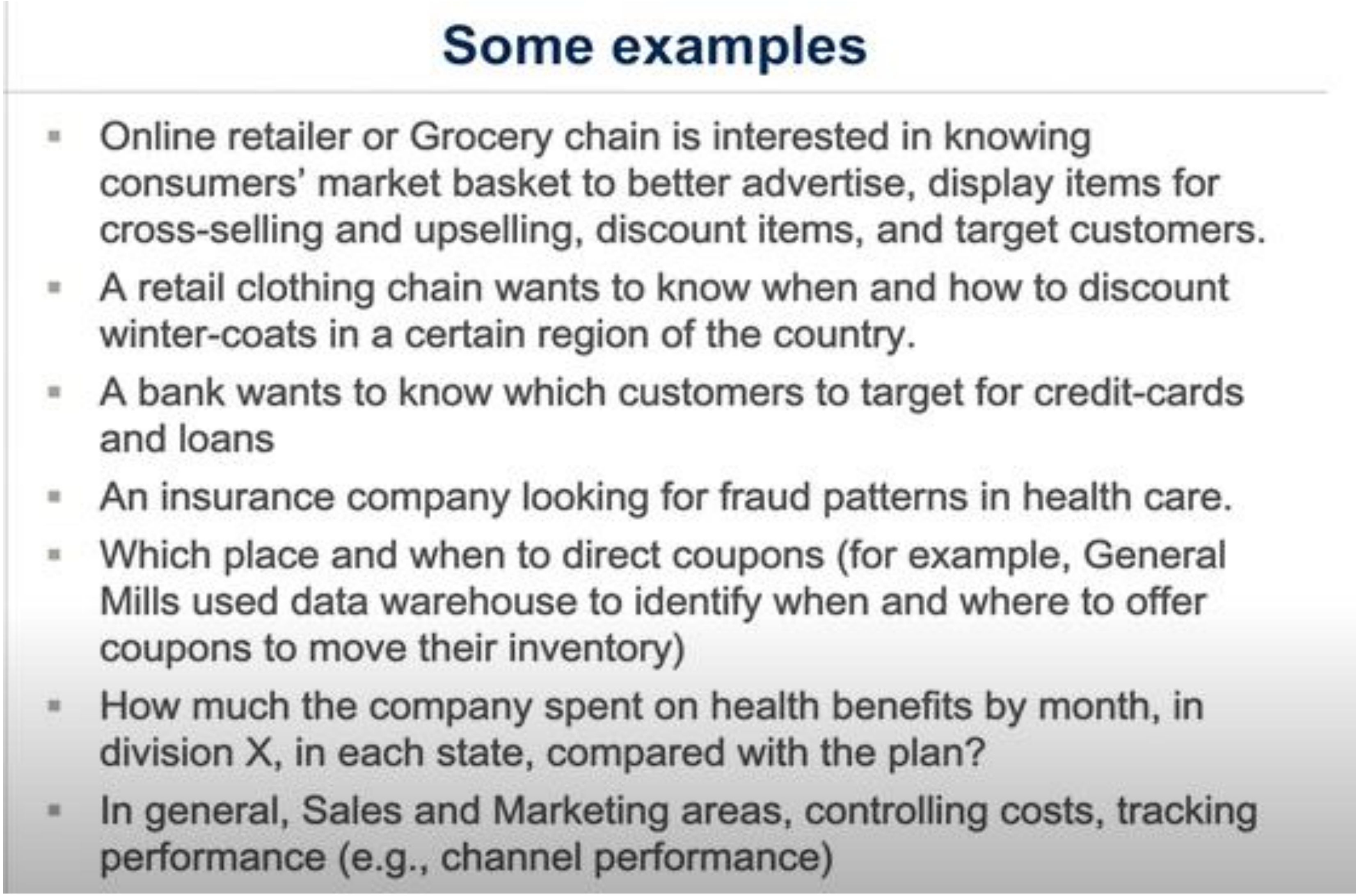
Data mining is a process of finding value from volume. In any enterprise, the amount of transactional data generated during its day-to-day operations is massive in volume. Although these transactions record every instance of any activity, it is of little use in decision making. Data mining attempts to extract smaller pieces of valuable information from this massive database.

4. Discovering relations that connect variables in a database is the subject of data mining. The data mining system self-learns from the previous history of the investigated system, formulating and testing hypothesis about rules which systems obey. When concise and valuable knowledge about the system of interest is discovered, it can and should be interpreted into some decision support system, which helps the manager to make wise and informed business decision.

Data mining is essentially a system that learns from the existing data. One can think of two disciplines which address such problems-Statistics and Machine Learning. Statistics provide sufficient tools for data analysis and machine learning deals with different learning methodologies.

Potential Applications

- Data analysis and decision support
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)
- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining
  - Bioinformatics and bio-data analysis

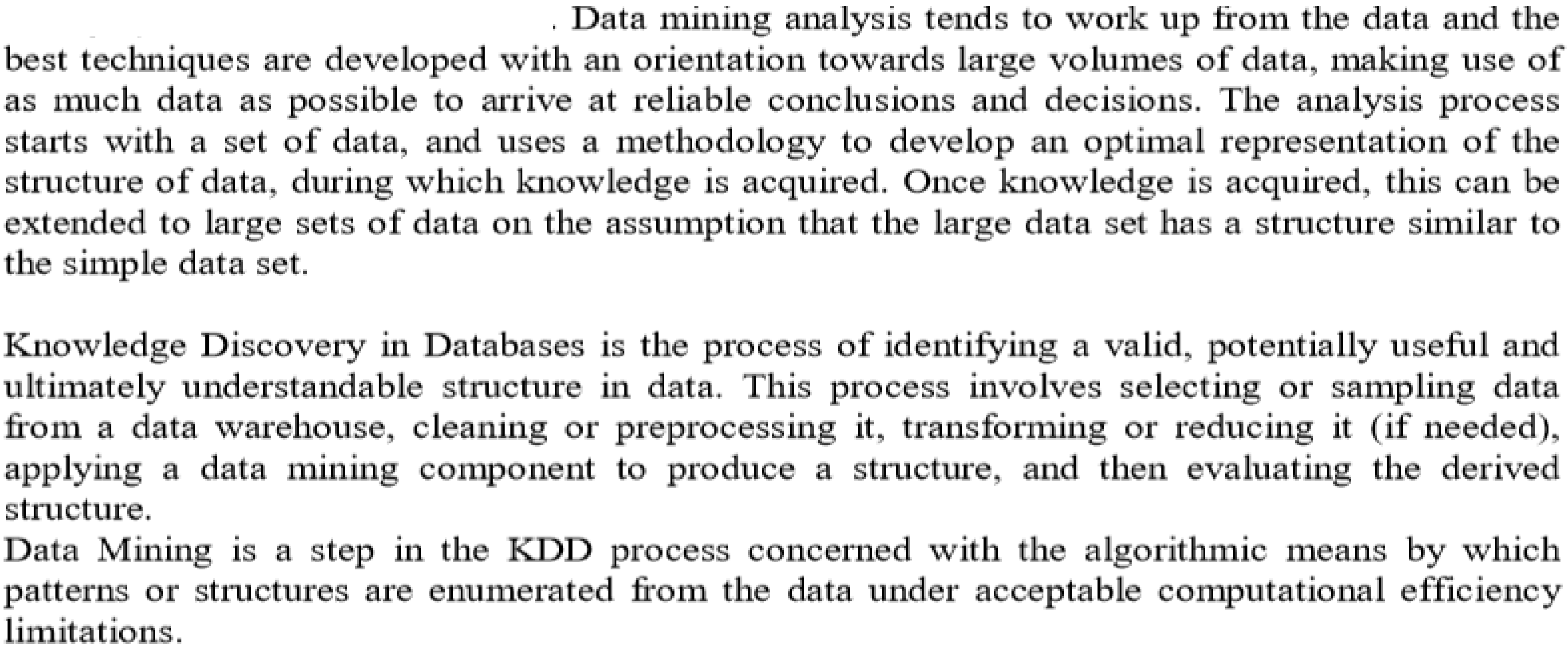


Some examples

- Online retailer or Grocery chain is interested in knowing consumers' market basket to better advertise, display items for cross-selling and upselling, discount items, and target customers.
- A retail clothing chain wants to know when and how to discount winter-coats in a certain region of the country.
- A bank wants to know which customers to target for credit-cards and loans
- An insurance company looking for fraud patterns in health care.
- Which place and when to direct coupons (for example, General Mills used data warehouse to identify when and where to offer coupons to move their inventory)
- How much the company spent on health benefits by month, in division X, in each state, compared with the plan?
- In general, Sales and Marketing areas, controlling costs, tracking performance (e.g., channel performance)

KDD Vs Data mining

- Knowledge Discovery in Databases (**KDD**) is the non-trivial extraction of implicit, previously unknown and potentially useful knowledge from **data**. **Data mining** is the exploration and analysis of large quantities of **data** in order to discover valid, novel, potentially useful, and ultimately understandable patterns in **data**.

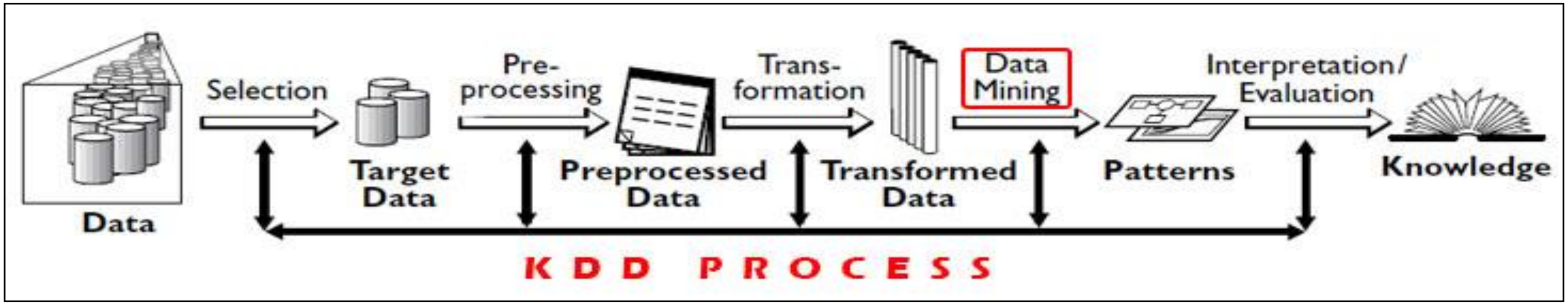


. Data mining analysis tends to work up from the data and the best techniques are developed with an orientation towards large volumes of data, making use of as much data as possible to arrive at reliable conclusions and decisions. The analysis process starts with a set of data, and uses a methodology to develop an optimal representation of the structure of data, during which knowledge is acquired. Once knowledge is acquired, this can be extended to large sets of data on the assumption that the large data set has a structure similar to the simple data set.

Knowledge Discovery in Databases is the process of identifying a valid, potentially useful and ultimately understandable structure in data. This process involves selecting or sampling data from a data warehouse, cleaning or preprocessing it, transforming or reducing it (if needed), applying a data mining component to produce a structure, and then evaluating the derived structure.

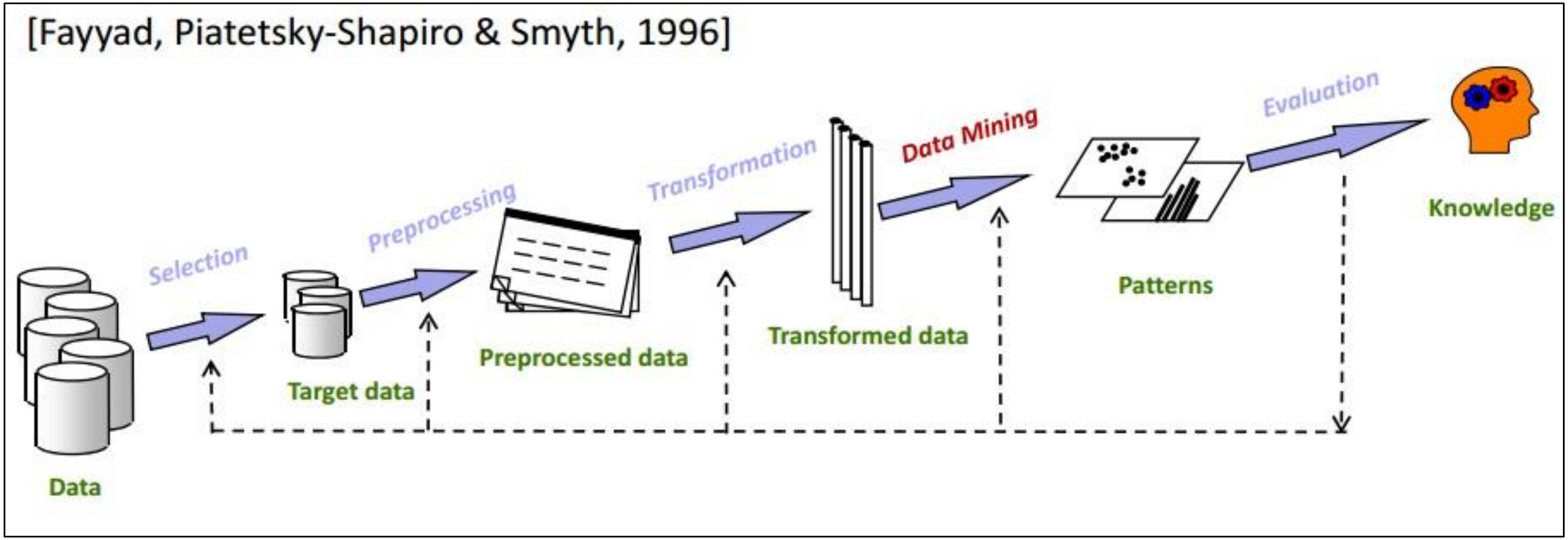
Data Mining is a step in the KDD process concerned with the algorithmic means by which patterns or structures are enumerated from the data under acceptable computational efficiency limitations.

KDD Vs Data mining



Data, in its raw form, is just a collection of things, where little information might be derived. Together with the development of information discovery methods(Data Mining and KDD), the value of the info is significantly improved.
Data mining is one among the steps of Knowledge Discovery in Databases(KDD) as can be shown by the image above. KDD is a multi-step process that encourages the conversion of data to useful information. Data mining is the pattern extraction phase of KDD. Data mining can take on several types, the option influenced by the desired outcomes.

KDD Vs Data mining (Stages of KDD)

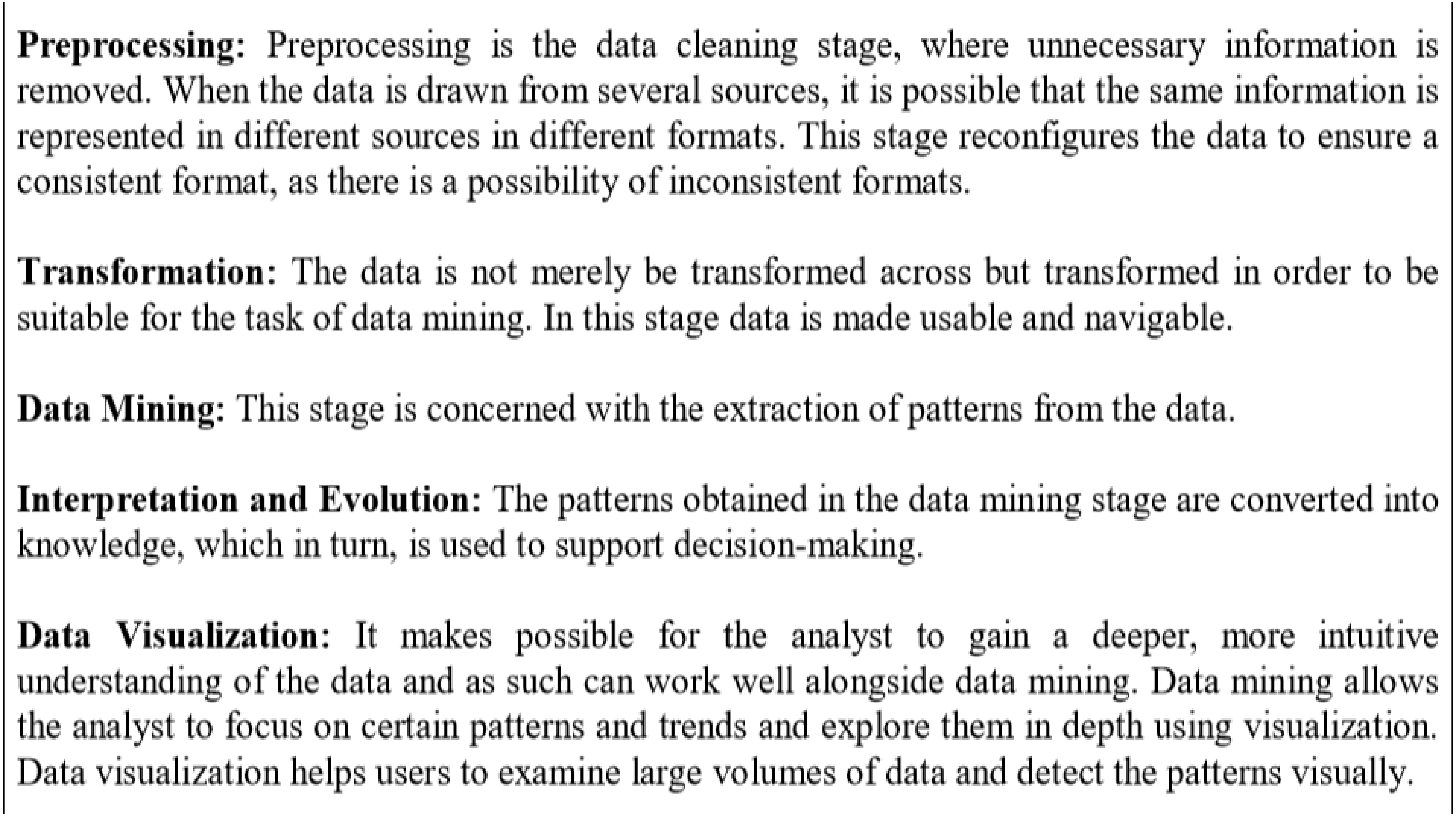


Stages of KDD

The stages of KDD, starting with the raw data and finishing with the extracted knowledge, are given below.

Selection: This stage is concerned with selecting or segmenting the data that are relevant to some criteria. For example, for credit card customer profiling, we extract the type of transactions for each type of customers and we may not be interested in details of the shop where the transaction takes place.

KDD Vs Data mining (Stages of KDD)



Preprocessing: Preprocessing is the data cleaning stage, where unnecessary information is removed. When the data is drawn from several sources, it is possible that the same information is represented in different sources in different formats. This stage reconfigures the data to ensure a consistent format, as there is a possibility of inconsistent formats.

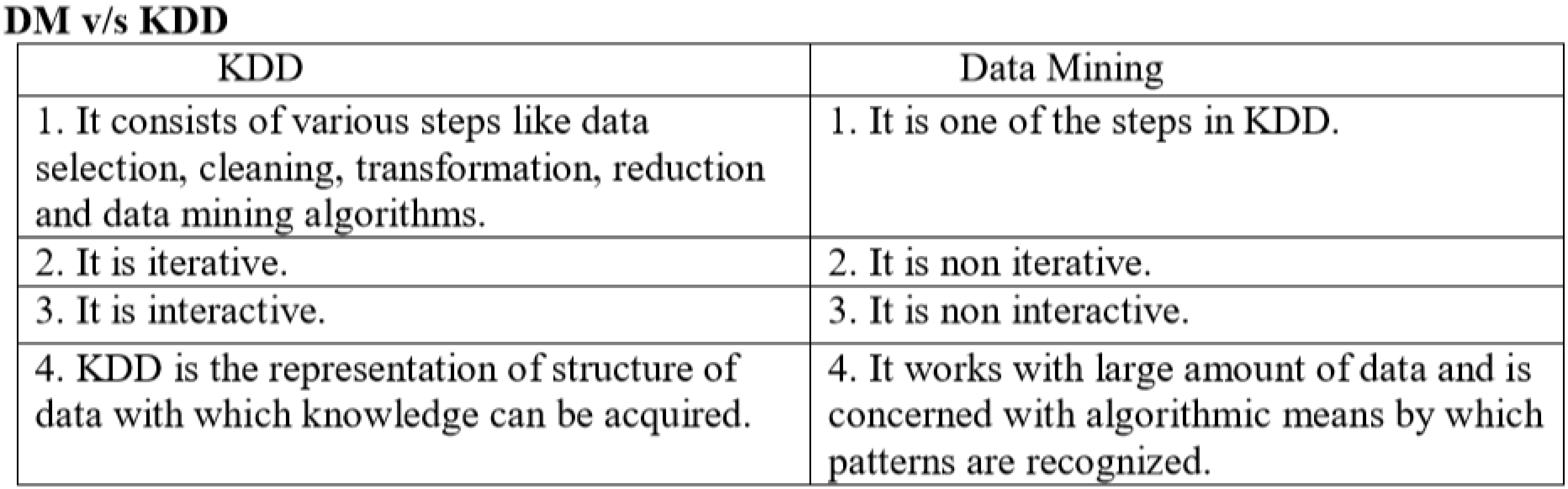
Transformation: The data is not merely be transformed across but transformed in order to be suitable for the task of data mining. In this stage data is made usable and navigable.

Data Mining: This stage is concerned with the extraction of patterns from the data.

Interpretation and Evolution: The patterns obtained in the data mining stage are converted into knowledge, which in turn, is used to support decision-making.

Data Visualization: It makes possible for the analyst to gain a deeper, more intuitive understanding of the data and as such can work well alongside data mining. Data mining allows the analyst to focus on certain patterns and trends and explore them in depth using visualization. Data visualization helps users to examine large volumes of data and detect the patterns visually.

KDD Vs Data mining (Stages of KDD)



DM v/s KDD

KDD	Data Mining
1. It consists of various steps like data selection, cleaning, transformation, reduction and data mining algorithms.	1. It is one of the steps in KDD.
2. It is iterative.	2. It is non iterative.
3. It is interactive.	3. It is non interactive.
4. KDD is the representation of structure of data with which knowledge can be acquired.	4. It works with large amount of data and is concerned with algorithmic means by which patterns are recognized.

Difference Between Database and Data Mining (DBMS Vs Data Mining)

Data Base:

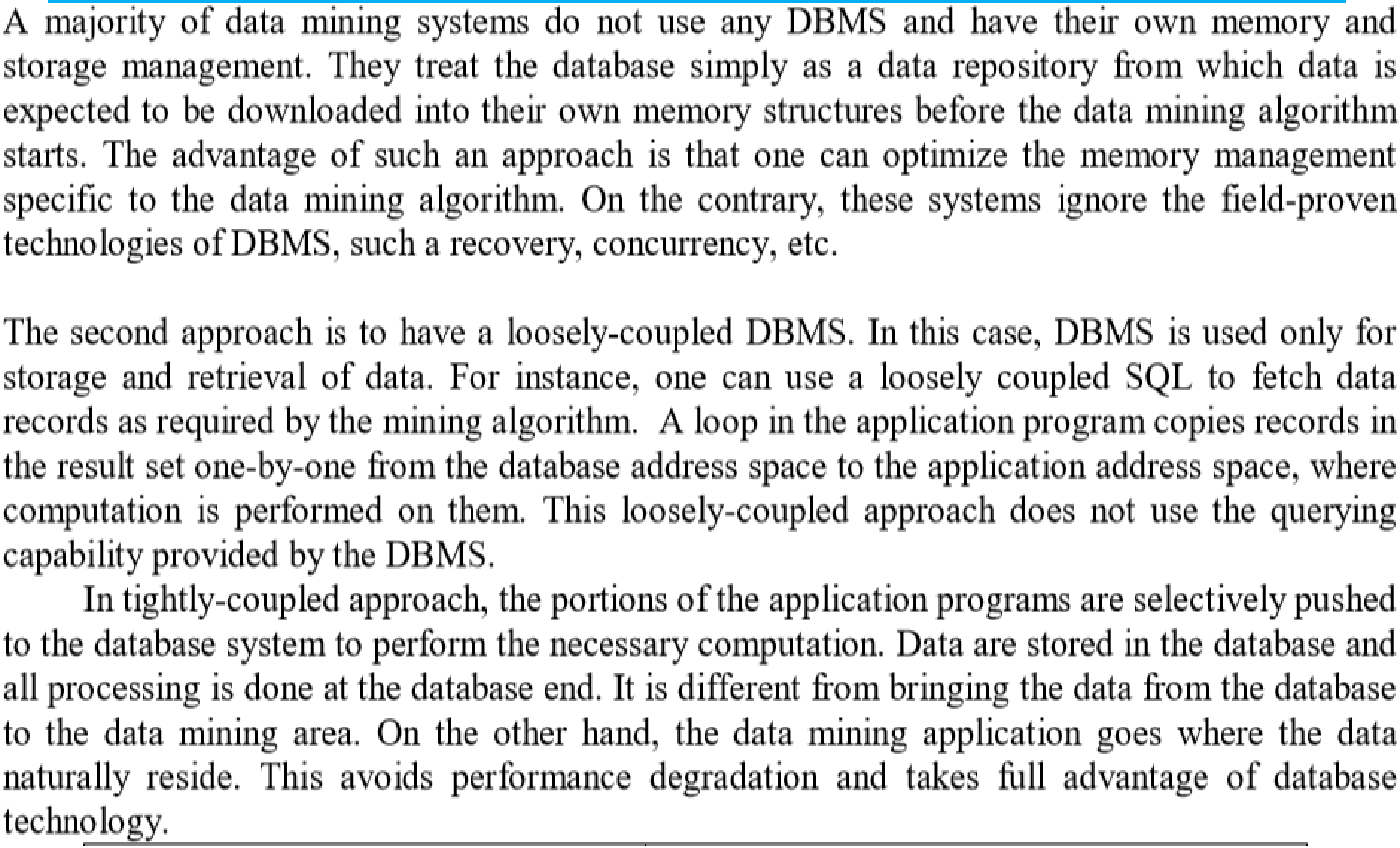
- The database is a collection of interrelated data and a set of programs to access those data.
- It is a software system that manages data stored in the database. It provides an effective method of defining, storing and retrieving the information contained in the database
- The primary goal of a DBMS is to provide an environment that is both convenient and efficient to use in retrieving and storing database information.
- Supports the query language, if we exactly know what information is needed a DBMS query can be used
- It provides users with information that they required. Some examples of DBMS packages are dBASE, FoxPro, FoxBase, Oracle, Ms-Access etc.

Difference Between Database and Data Mining (DBMS Vs Data Mining)

Data Mining:

- Data mining is the process of analyzing data from a different perspective and summarizing it into useful information – information that can be used to increase revenue cuts cost or both.
- Data mining the analysis step of the knowledge discovery in database process. For example, data mining software can help retail companies find customers with a common interest.
- Data mining deals with extracting useful and previously unknown information from raw data.

Difference Between Database and Data Mining (DBMS Vs Data Mining)



A majority of data mining systems do not use any DBMS and have their own memory and storage management. They treat the database simply as a data repository from which data is expected to be downloaded into their own memory structures before the data mining algorithm starts. The advantage of such an approach is that one can optimize the memory management specific to the data mining algorithm. On the contrary, these systems ignore the field-proven technologies of DBMS, such a recovery, concurrency, etc.

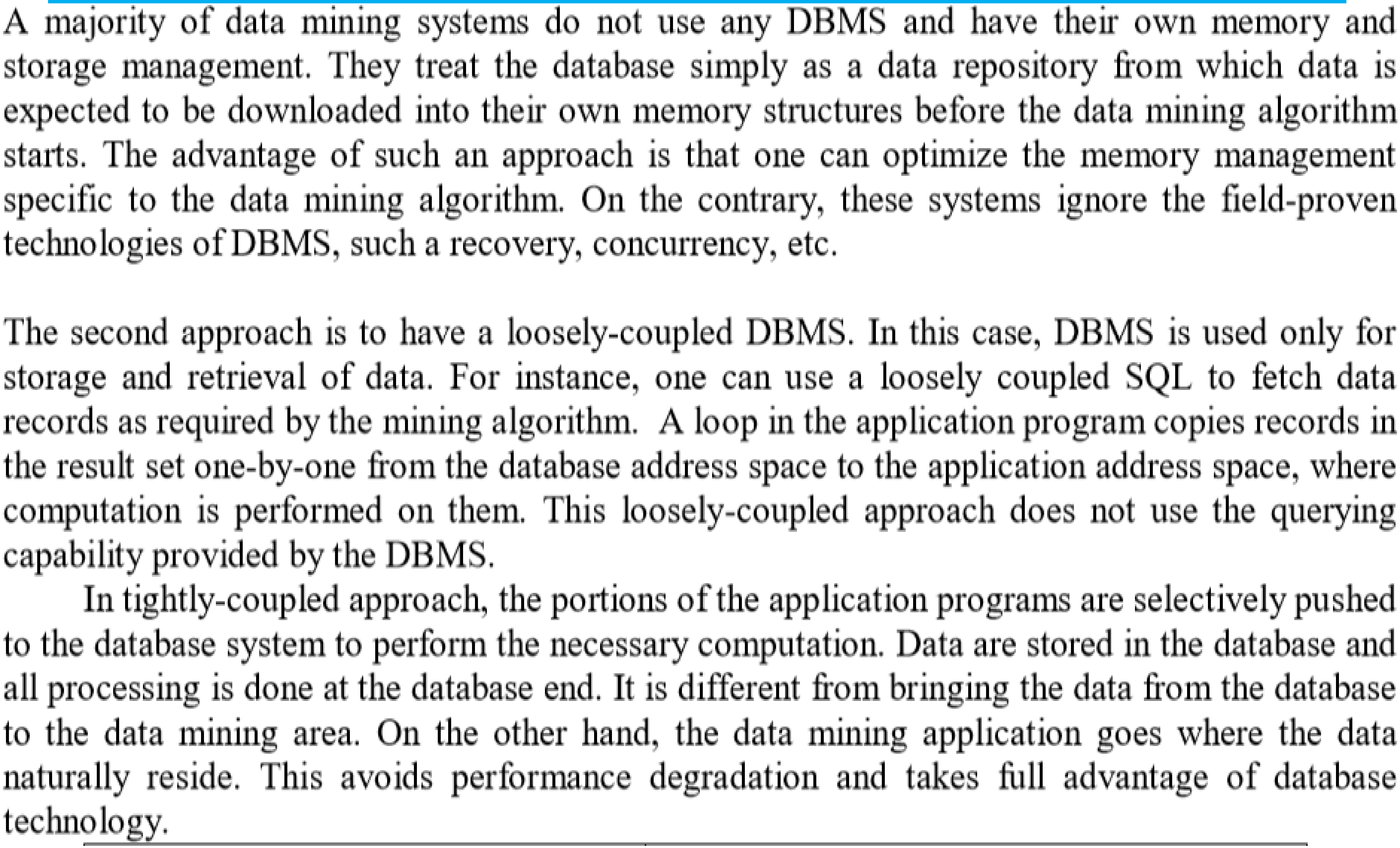
The second approach is to have a loosely-coupled DBMS. In this case, DBMS is used only for storage and retrieval of data. For instance, one can use a loosely coupled SQL to fetch data records as required by the mining algorithm. A loop in the application program copies records in the result set one-by-one from the database address space to the application address space, where computation is performed on them. This loosely-coupled approach does not use the querying capability provided by the DBMS.

In tightly-coupled approach, the portions of the application programs are selectively pushed to the database system to perform the necessary computation. Data are stored in the database and all processing is done at the database end. It is different from bringing the data from the database to the data mining area. On the other hand, the data mining application goes where the data naturally reside. This avoids performance degradation and takes full advantage of database technology.

Data Mining vs. Database

- DB’s user knows what is looking for.
- DM’s user might/might not know what is looking for.
- DB’s answer to query is 100% accurate, if data correct.
- DM’s effort is to get the answer as accurate as possible.
- DB’s data are retrieved as stored.
- DM’s data need to be cleaned (some what) before producing results.
- DB’s results are subset of data.
- DM’s results are the analysis of the data.
- The meaningfulness of the results is not the concern of Database as
- it is the main issue in Data Mining.

Difference Between Database and Data Mining (DBMS Vs Data Mining)

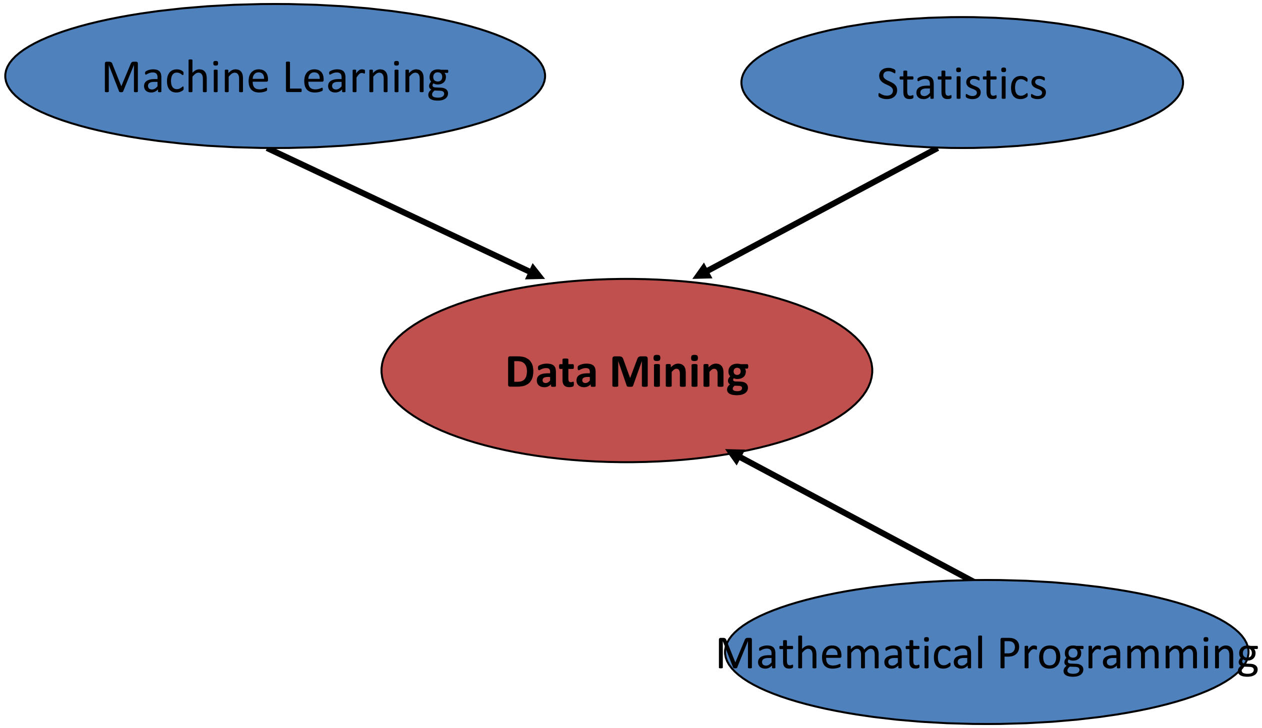


A majority of data mining systems do not use any DBMS and have their own memory and storage management. They treat the database simply as a data repository from which data is expected to be downloaded into their own memory structures before the data mining algorithm starts. The advantage of such an approach is that one can optimize the memory management specific to the data mining algorithm. On the contrary, these systems ignore the field-proven technologies of DBMS, such a recovery, concurrency, etc.

The second approach is to have a loosely-coupled DBMS. In this case, DBMS is used only for storage and retrieval of data. For instance, one can use a loosely coupled SQL to fetch data records as required by the mining algorithm. A loop in the application program copies records in the result set one-by-one from the database address space to the application address space, where computation is performed on them. This loosely-coupled approach does not use the querying capability provided by the DBMS.

In tightly-coupled approach, the portions of the application programs are selectively pushed to the database system to perform the necessary computation. Data are stored in the database and all processing is done at the database end. It is different from bringing the data from the database to the data mining area. On the other hand, the data mining application goes where the data naturally reside. This avoids performance degradation and takes full advantage of database technology.

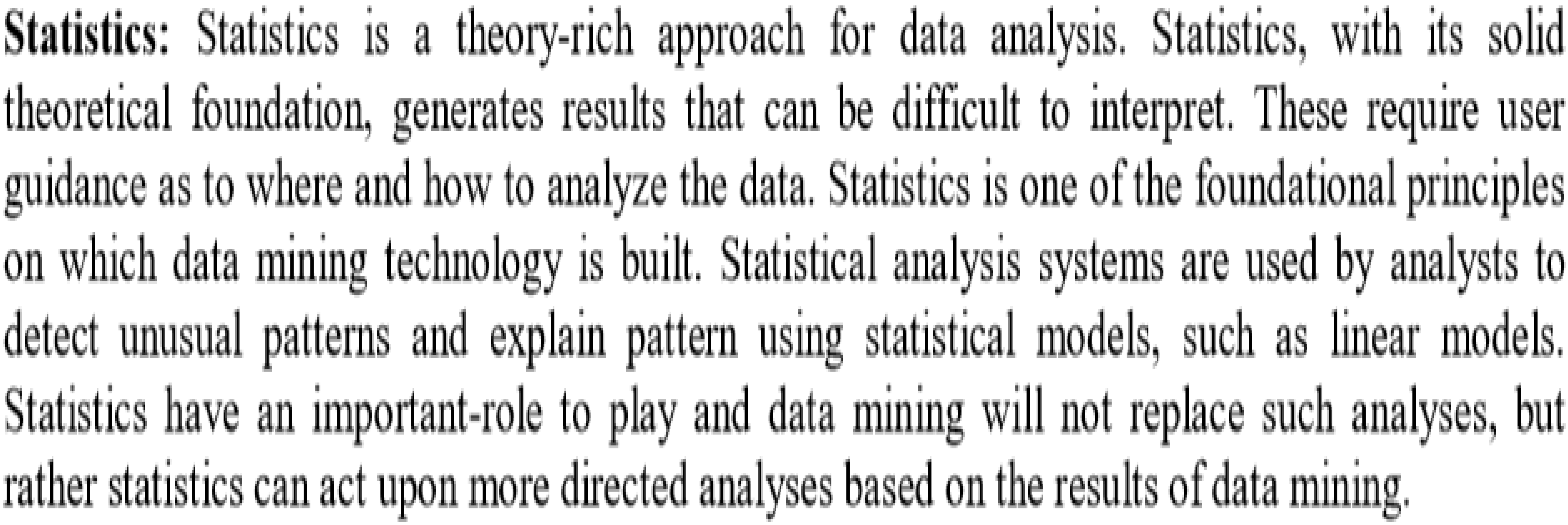
Other Related Info



Other Related Info

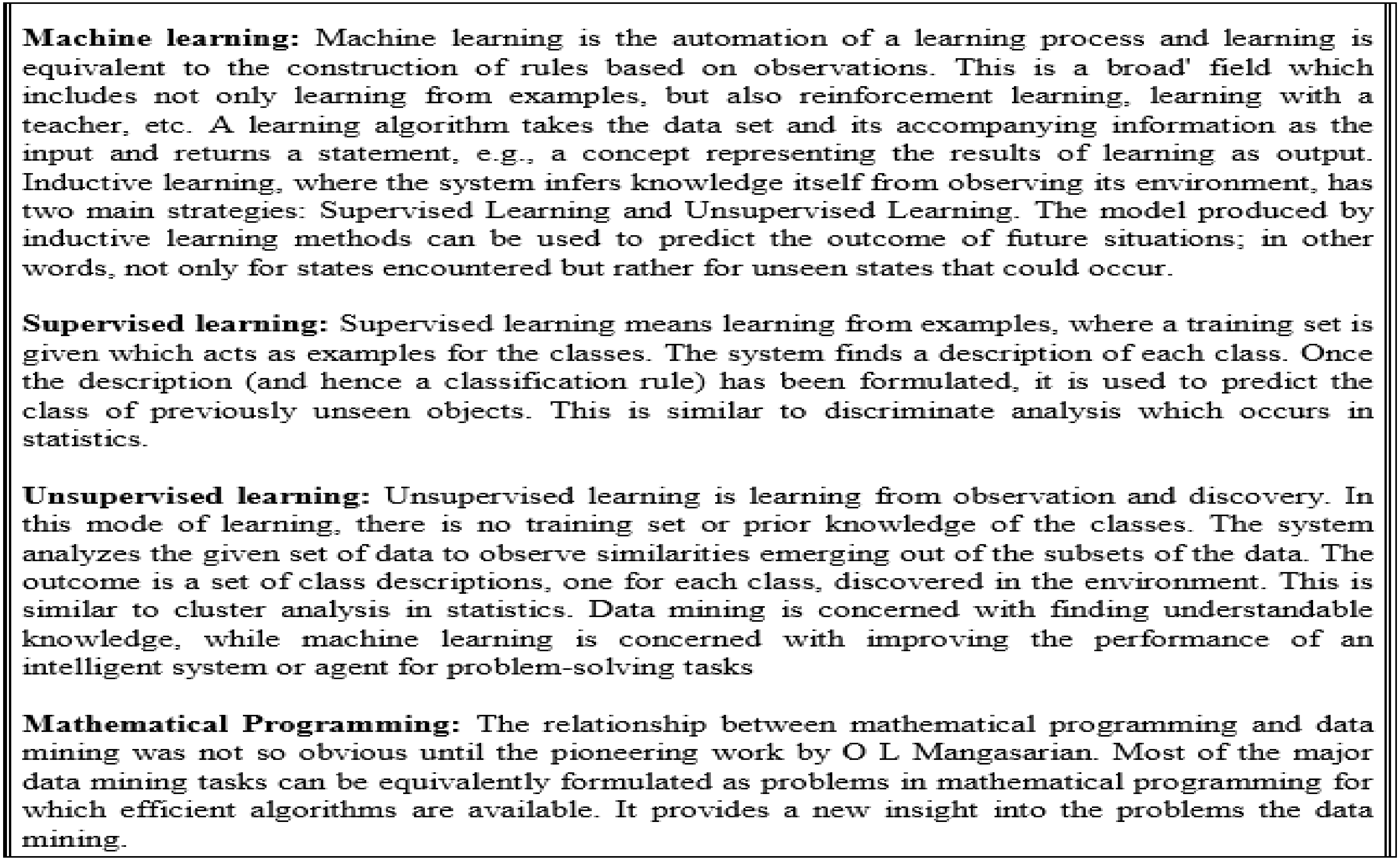
Statistics:

Statistics is a component of **data mining** that provides the tools and analytics techniques for dealing with large amounts of **data**. It is the science of learning from **data** and includes everything from collecting and organizing to analyzing and presenting **data**.



Statistics: Statistics is a theory-rich approach for data analysis. Statistics, with its solid theoretical foundation, generates results that can be difficult to interpret. These require user guidance as to where and how to analyze the data. Statistics is one of the foundational principles on which data mining technology is built. Statistical analysis systems are used by analysts to detect unusual patterns and explain pattern using statistical models, such as linear models. Statistics have an important-role to play and data mining will not replace such analyses, but rather statistics can act upon more directed analyses based on the results of data mining.

Other Related Info



Machine learning: Machine learning is the automation of a learning process and learning is equivalent to the construction of rules based on observations. This is a broad' field which includes not only learning from examples, but also reinforcement learning, learning with a teacher, etc. A learning algorithm takes the data set and its accompanying information as the input and returns a statement, e.g., a concept representing the results of learning as output. Inductive learning, where the system infers knowledge itself from observing its environment, has two main strategies: Supervised Learning and Unsupervised Learning. The model produced by inductive learning methods can be used to predict the outcome of future situations; in other words, not only for states encountered but rather for unseen states that could occur.

Supervised learning: Supervised learning means learning from examples, where a training set is given which acts as examples for the classes. The system finds a description of each class. Once the description (and hence a classification rule) has been formulated, it is used to predict the class of previously unseen objects. This is similar to discriminate analysis which occurs in statistics.

Unsupervised learning: Unsupervised learning is learning from observation and discovery. In this mode of learning, there is no training set or prior knowledge of the classes. The system analyzes the given set of data to observe similarities emerging out of the subsets of the data. The outcome is a set of class descriptions, one for each class, discovered in the environment. This is similar to cluster analysis in statistics. Data mining is concerned with finding understandable knowledge, while machine learning is concerned with improving the performance of an intelligent system or agent for problem-solving tasks

Mathematical Programming: The relationship between mathematical programming and data mining was not so obvious until the pioneering work by O L Mangasarian. Most of the major data mining tasks can be equivalently formulated as problems in mathematical programming for which efficient algorithms are available. It provides a new insight into the problems the data mining.