Data Mining Techniques, Issues, Applications-Chapter 1

By Anupama Kota
Computer Science Dept

Data Mining Techniques

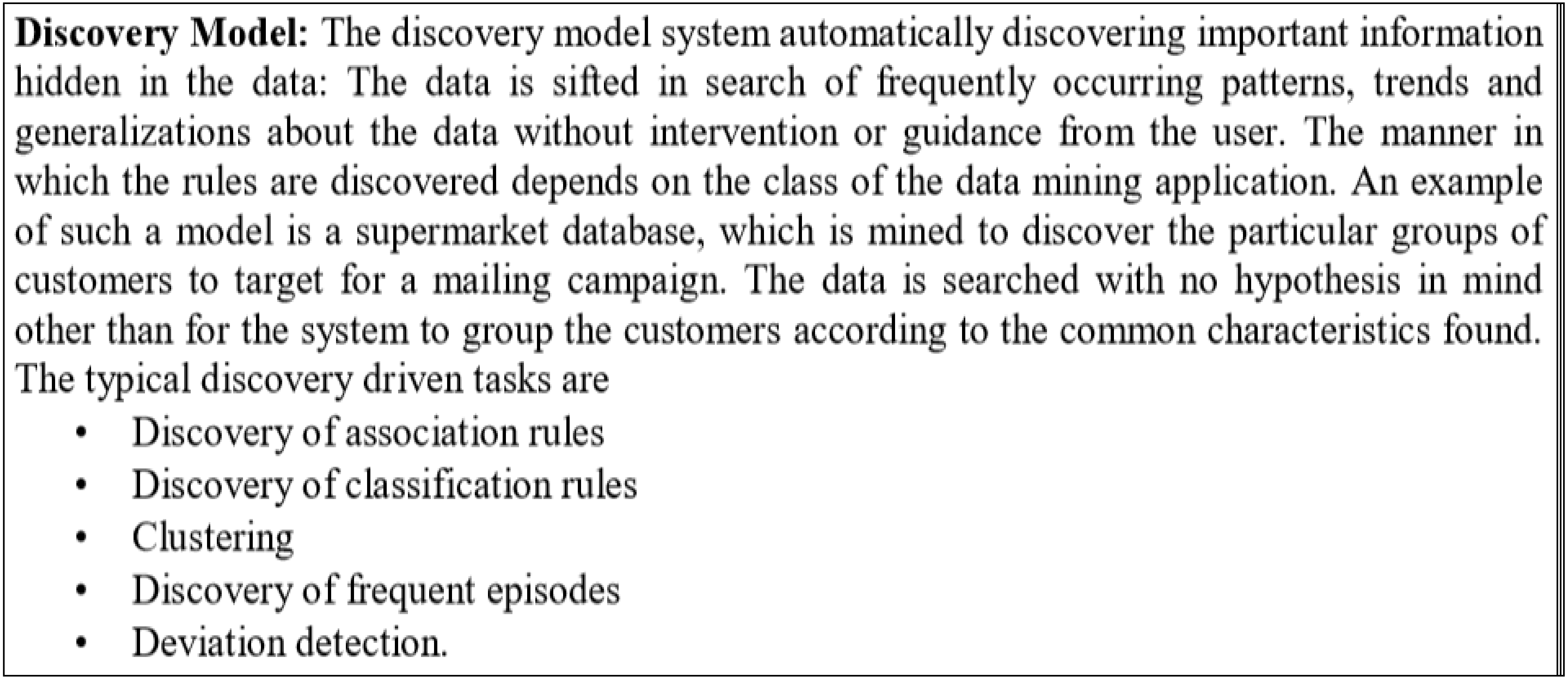
- Two Fundamental goals of Data Mining:
  - Prediction – predict unknown or future values of interest
  - Description – Focuses on finding patterns describing the data and the subsequent presentation for user interpretation
- Some of the DM techniques fulfilling these objectives
  - Associations – find all associations such that presence of one set of items in a transaction implies the other items
  - Classifications – Develop profile of different groups
  - Sequential Patterns-Identify the patterns with user defined minimum constraint
  - Clustering-Segment a database into subsets

Data Mining Techniques

- These DM techniques are classified as
  - User guided or verification-driven data mining
  - Discovery-Driven or automatic discovery of rules

  - **Verification Model:** The **verification model** takes an hypothesis from the user and tests the validity of it against the **data**

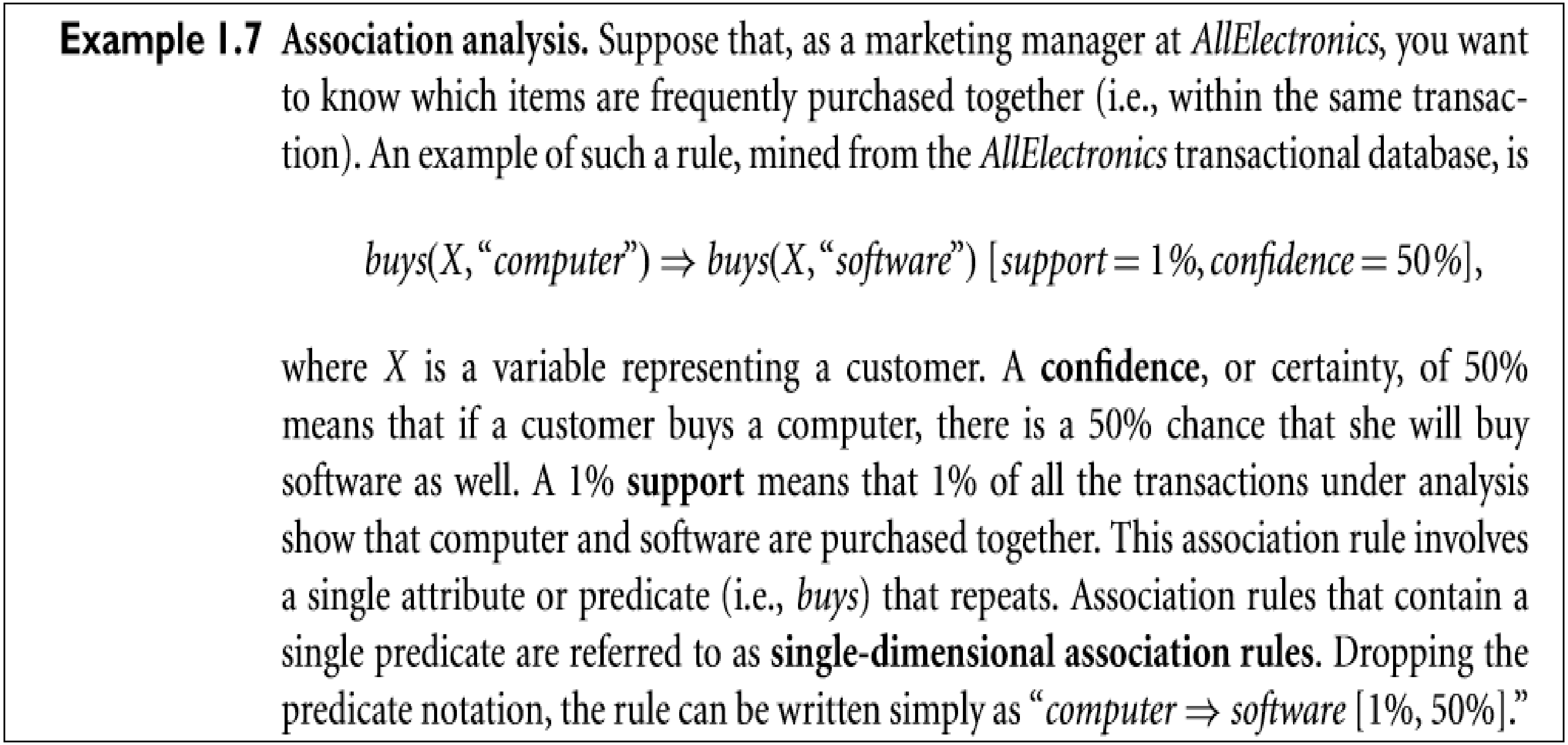
Data Mining Techniques



Discovery Model: The discovery model system automatically discovering important information hidden in the data: The data is sifted in search of frequently occurring patterns, trends and generalizations about the data without intervention or guidance from the user. The manner in which the rules are discovered depends on the class of the data mining application. An example of such a model is a supermarket database, which is mined to discover the particular groups of customers to target for a mailing campaign. The data is searched with no hypothesis in mind other than for the system to group the customers according to the common characteristics found. The typical discovery driven tasks are

- Discovery of association rules
- Discovery of classification rules
- Clustering
- Discovery of frequent episodes
- Deviation detection.

Discovery Model –Association rule

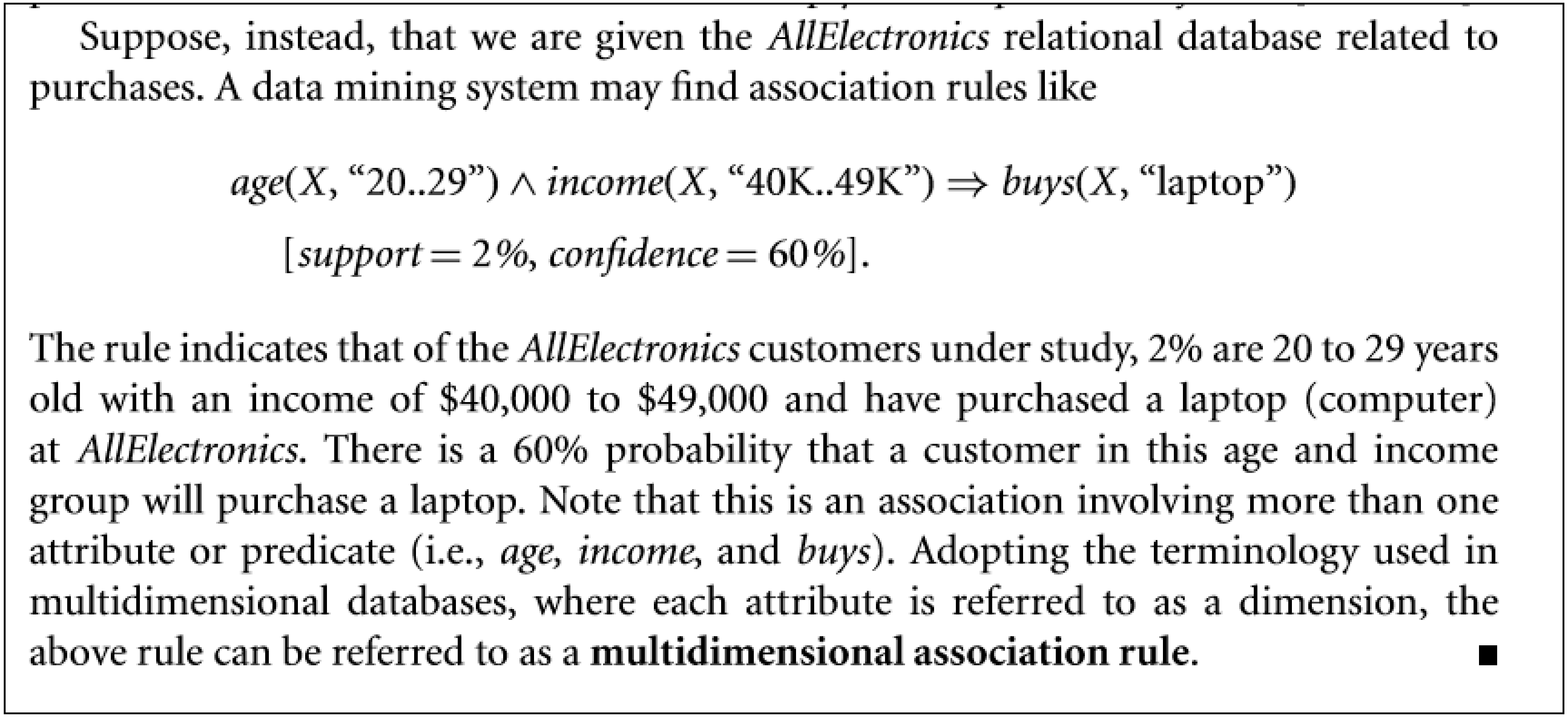


Example 1.7 **Association analysis.** Suppose that, as a marketing manager at *AllElectronics*, you want to know which items are frequently purchased together (i.e., within the same transaction). An example of such a rule, mined from the *AllElectronics* transactional database, is

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

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 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. This association rule involves a single attribute or predicate (i.e., *buys*) that repeats. Association rules that contain a single predicate are referred to as **single-dimensional association rules**. Dropping the predicate notation, the rule can be written simply as “$computer \Rightarrow software$ [1%, 50%].”

Discovery Model –Association rule

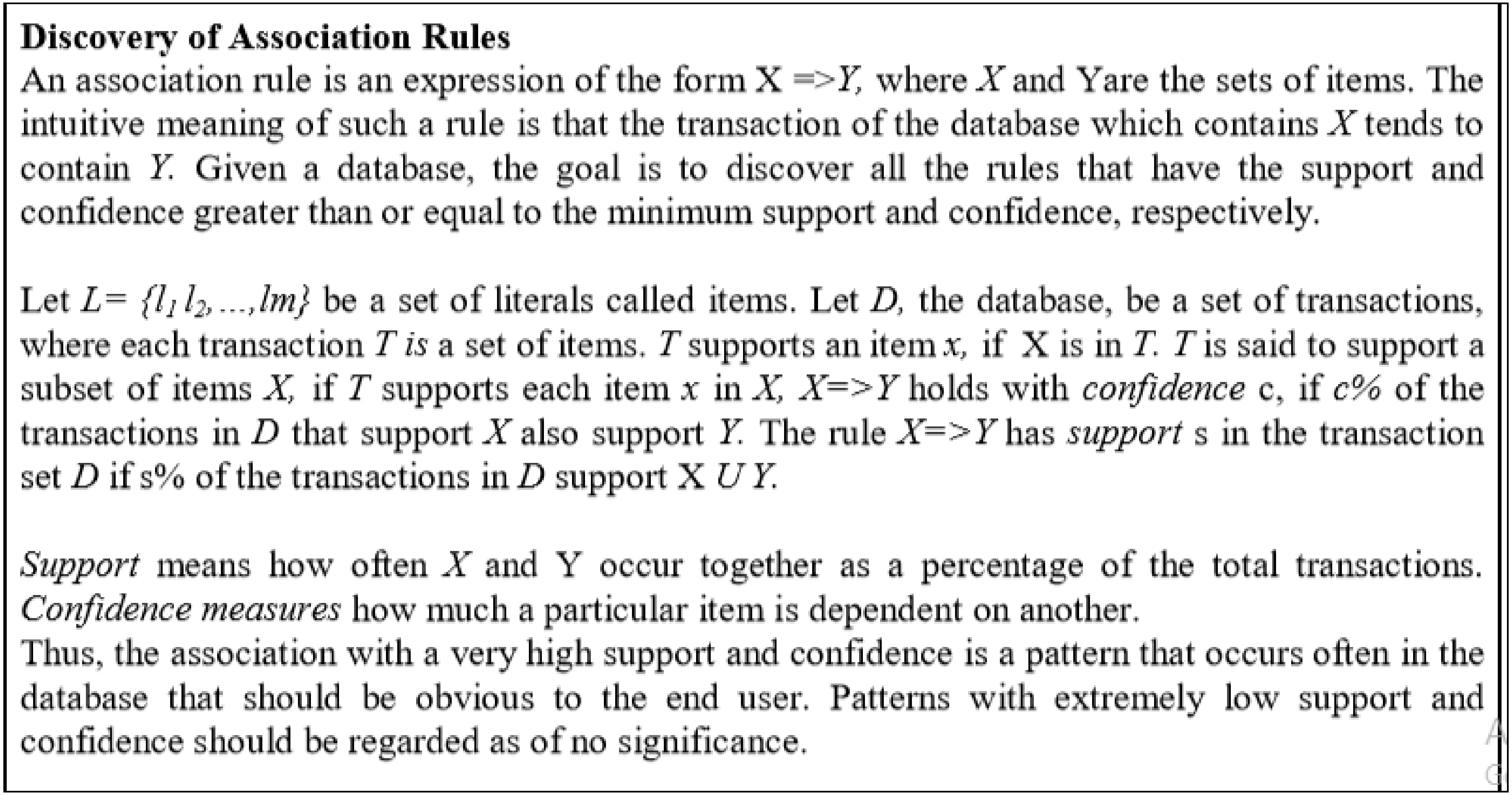


Suppose, instead, that we are given the *AllElectronics* relational database related to purchases. A data mining system may find association rules like

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

The rule indicates that of the *AllElectronics* customers under study, 2% are 20 to 29 years old with an income of $40,000 to $49,000 and have purchased a laptop (computer) at *AllElectronics*. There is a 60% probability that a customer in this age and income group will purchase a laptop. Note that this is an association involving more than one attribute or predicate (i.e., *age*, *income*, and *buys*). Adopting the terminology used in multidimensional databases, where each attribute is referred to as a dimension, the above rule can be referred to as a **multidimensional association rule.** ■

Discovery Model –Association rule



Discovery of Association Rules

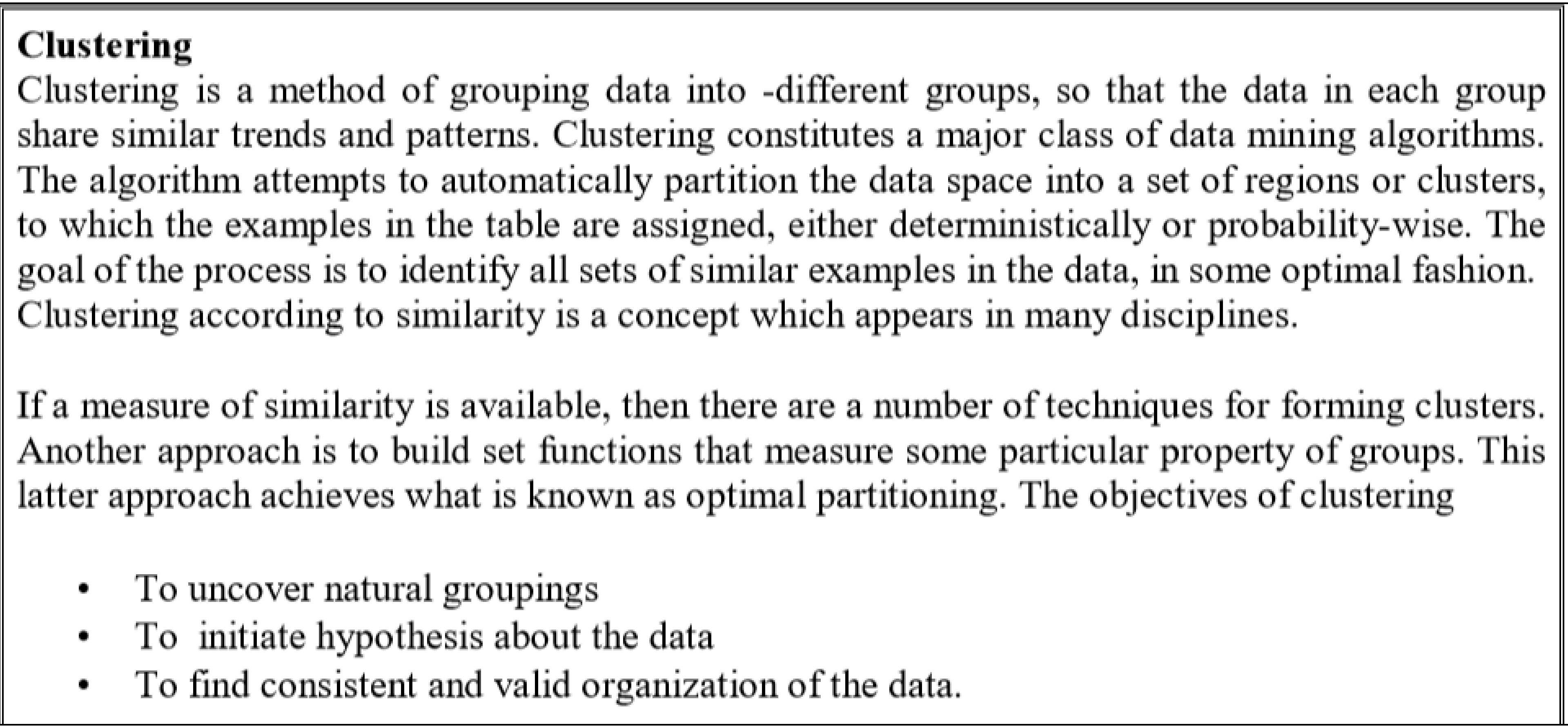
An association rule is an expression of the form X =>*Y*, where *X* and Yare the sets of items. The intuitive meaning of such a rule is that the transaction of the database which contains *X* tends to contain *Y*. Given a database, the goal is to discover all the rules that have the support and confidence greater than or equal to the minimum support and confidence, respectively.

Let $L= \{l_1 l_2, \ldots, lm\}$ be a set of literals called items. Let *D*, the database, be a set of transactions, where each transaction *T is* a set of items. *T* supports an item *x*, if X is in *T*. *T* is said to support a subset of items *X*, if *T* supports each item *x* in *X*, *X*=>*Y* holds with *confidence* c, if *c%* of the transactions in *D* that support *X* also support *Y*. The rule *X*=>*Y* has *support* s in the transaction set *D* if s% of the transactions in *D* support X *U Y*.

Support means how often *X* and Y occur together as a percentage of the total transactions. *Confidence measures* how much a particular item is dependent on another.

Thus, the association with a very high support and confidence is a pattern that occurs often in the database that should be obvious to the end user. Patterns with extremely low support and confidence should be regarded as of no significance.

Discovery Model –Clustering



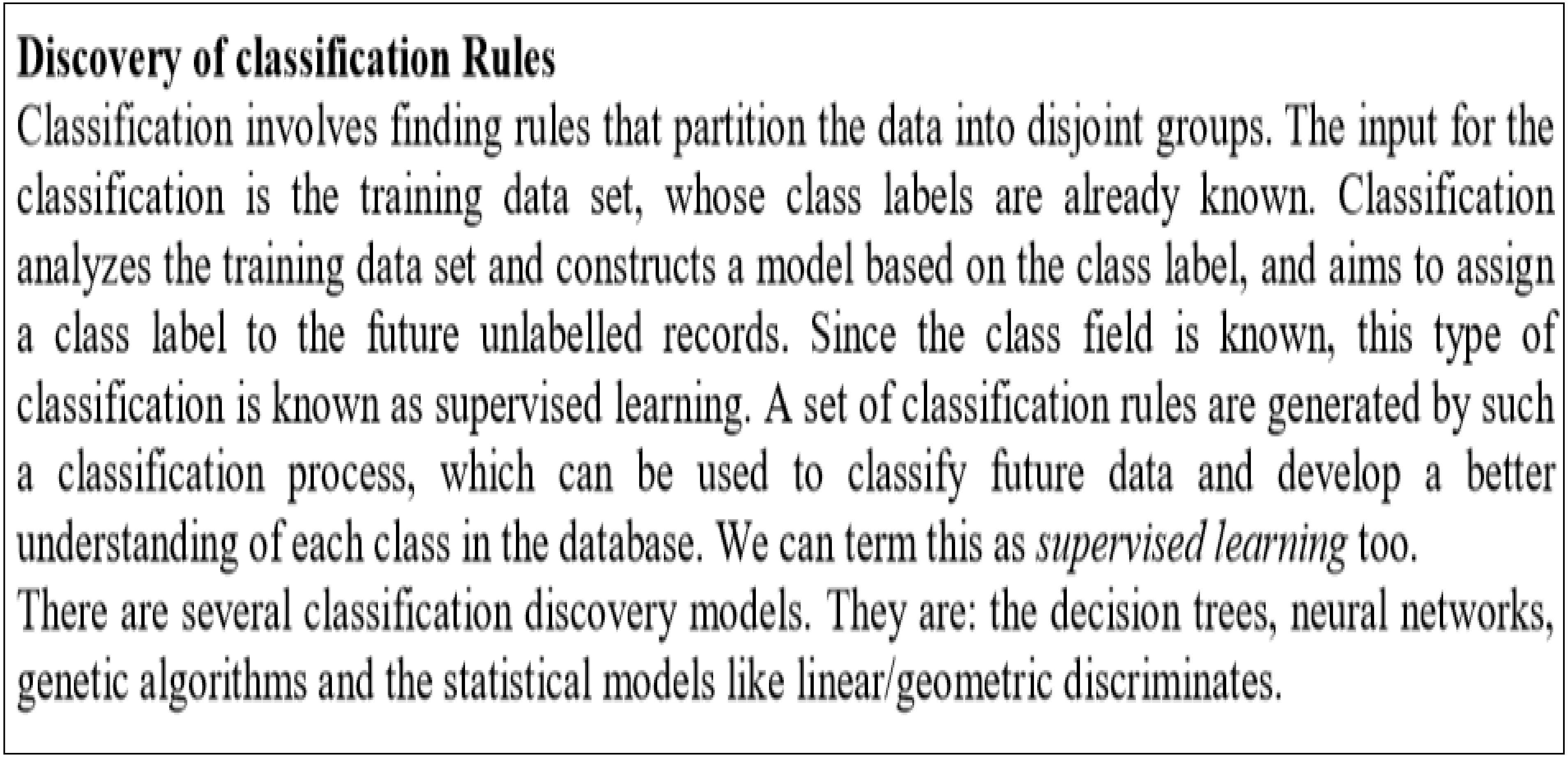
Clustering

Clustering is a method of grouping data into -different groups, so that the data in each group share similar trends and patterns. Clustering constitutes a major class of data mining algorithms. The algorithm attempts to automatically partition the data space into a set of regions or clusters, to which the examples in the table are assigned, either deterministically or probability-wise. The goal of the process is to identify all sets of similar examples in the data, in some optimal fashion. Clustering according to similarity is a concept which appears in many disciplines.

If a measure of similarity is available, then there are a number of techniques for forming clusters. Another approach is to build set functions that measure some particular property of groups. This latter approach achieves what is known as optimal partitioning. The objectives of clustering

- To uncover natural groupings
- To initiate hypothesis about the data
- To find consistent and valid organization of the data.

Discovery Model –Classification rule

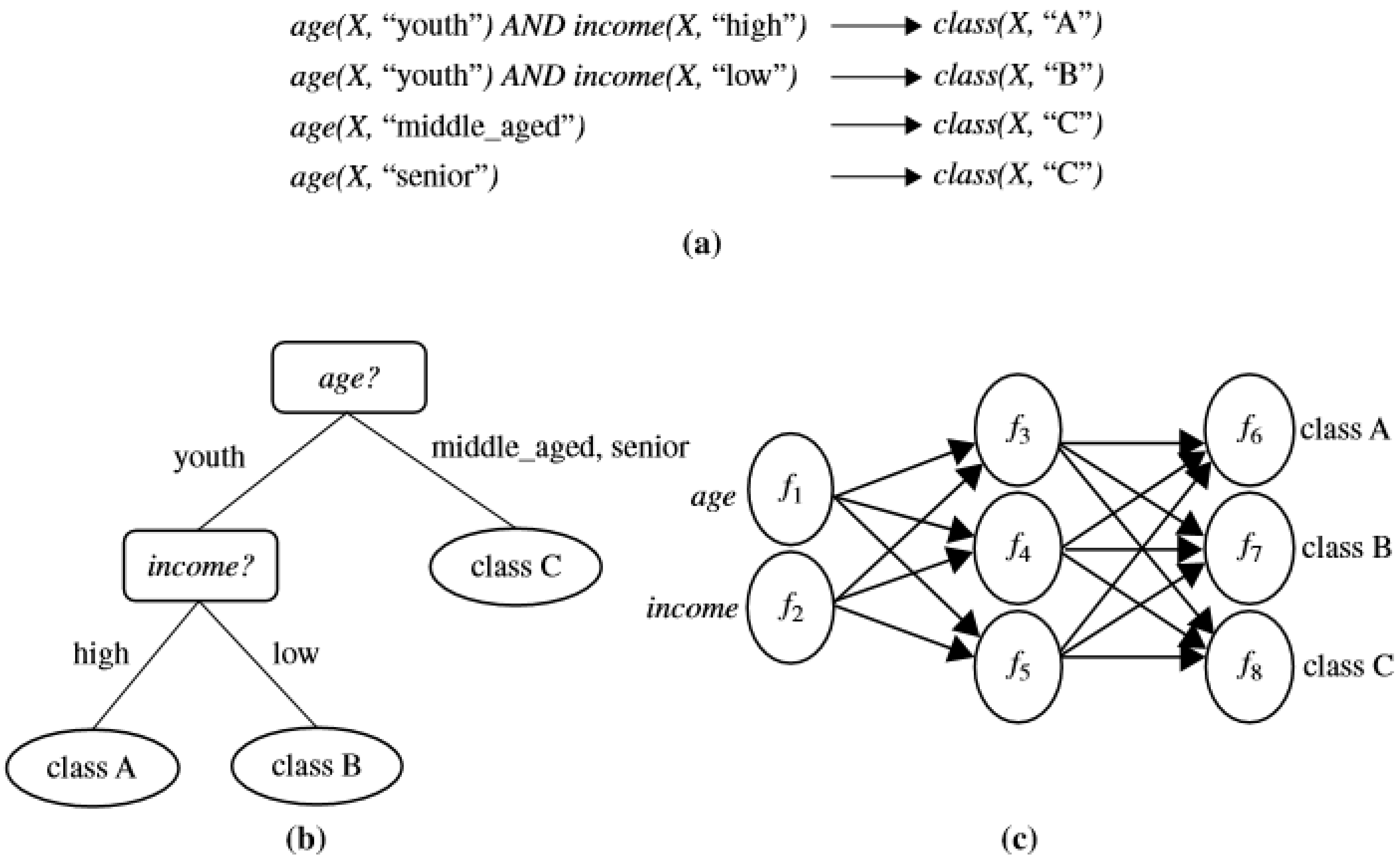


Discovery of classification Rules

Classification involves finding rules that partition the data into disjoint groups. The input for the classification is the training data set, whose class labels are already known. Classification analyzes the training data set and constructs a model based on the class label, and aims to assign a class label to the future unlabelled records. Since the class field is known, this type of classification is known as supervised learning. A set of classification rules are generated by such a classification process, which can be used to classify future data and develop a better understanding of each class in the database. We can term this as *supervised learning* too.

There are several classification discovery models. They are: the decision trees, neural networks, genetic algorithms and the statistical models like linear/geometric discriminates.

Discovery Model –Classification rule



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")

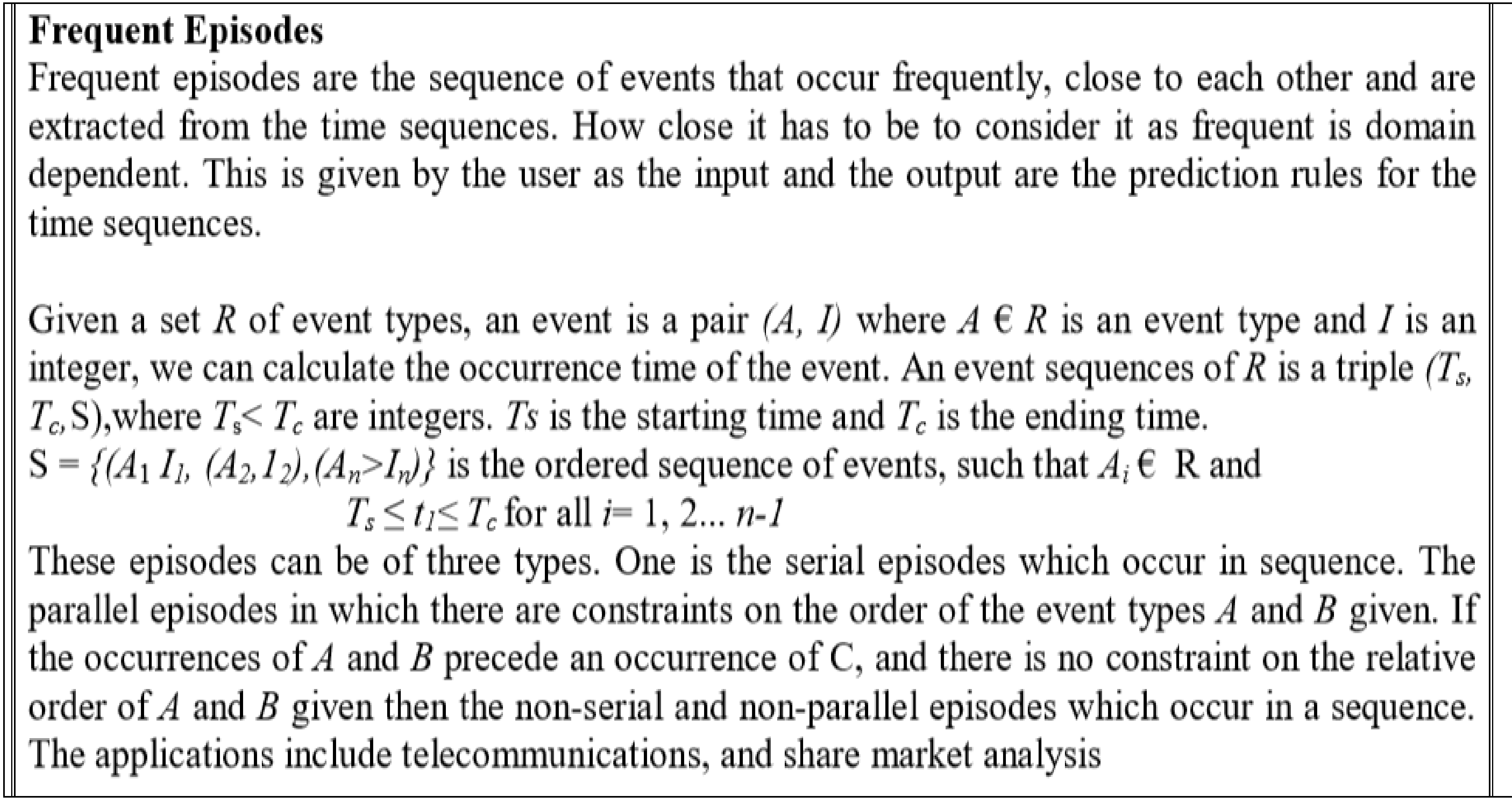
(a)

(b)

(c)

igure 1.9 A classification model can be represented in various forms: (a) IF-THEN rules, (b) a decision tree, or (c) a neural network.

Discovery Model –Frequent Episode



Frequent Episodes

Frequent episodes are the sequence of events that occur frequently, close to each other and are extracted from the time sequences. How close it has to be to consider it as frequent is domain dependent. This is given by the user as the input and the output are the prediction rules for the time sequences.

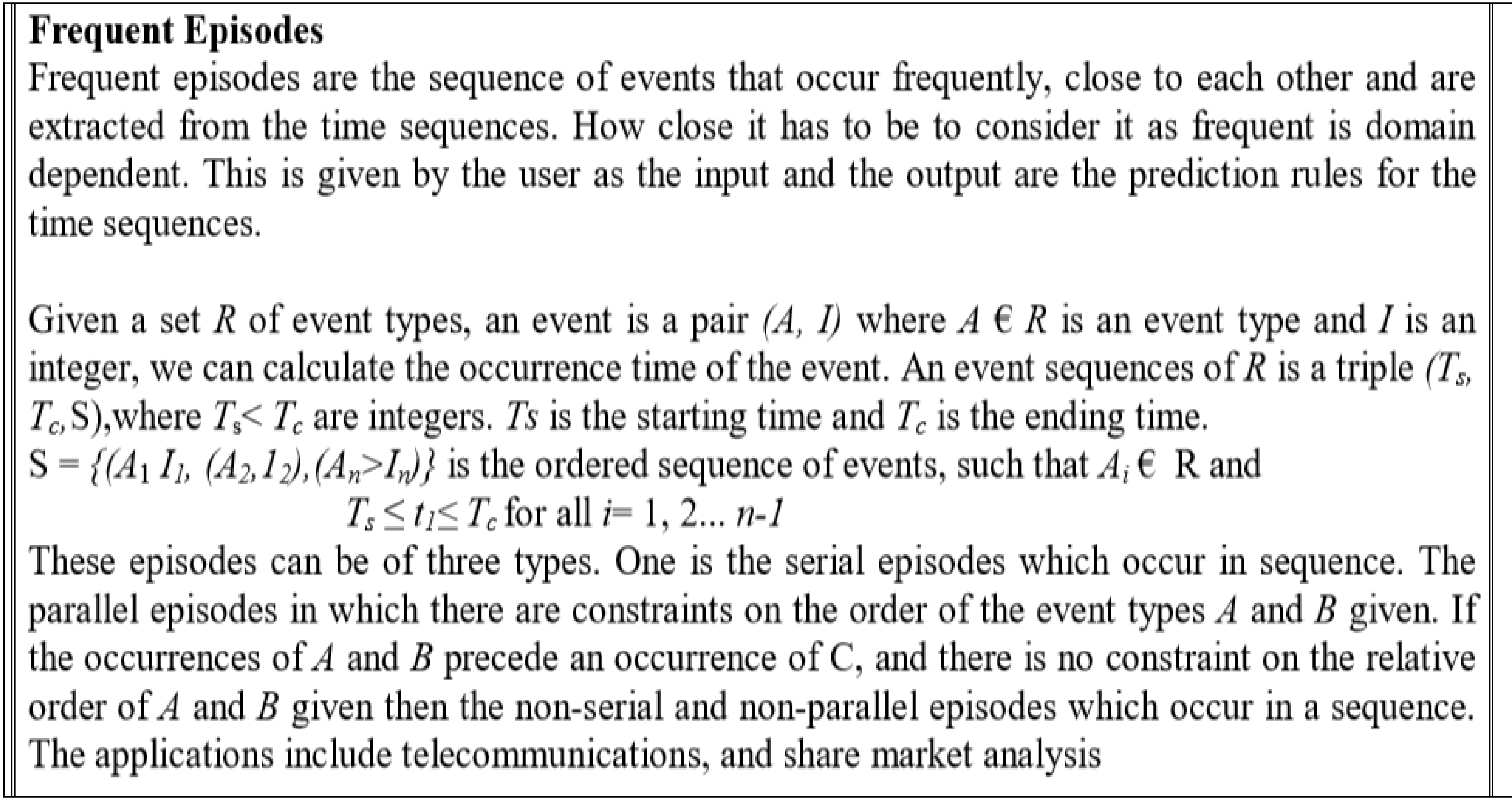
Given a set R of event types, an event is a pair (A, I) where A € R is an event type and I is an integer, we can calculate the occurrence time of the event. An event sequences of R is a triple $(T_s, T_c$, S),where $T_s < T_c$ are integers. Ts is the starting time and T_c is the ending time.

S = $\{(A_1 I_1, (A_2,I_2),(A_n>I_n)\}$ is the ordered sequence of events, such that A_i € R and

$$T_s \leq t_1 \leq T_c \text{ for all } i = 1, 2 \ldots n\text{-}1$$

These episodes can be of three types. One is the serial episodes which occur in sequence. The parallel episodes in which there are constraints on the order of the event types A and B given. If the occurrences of A and B precede an occurrence of C, and there is no constraint on the relative order of A and B given then the non-serial and non-parallel episodes which occur in a sequence. The applications include telecommunications, and share market analysis

Discovery Model –Frequent Episode



Frequent Episodes

Frequent episodes are the sequence of events that occur frequently, close to each other and are extracted from the time sequences. How close it has to be to consider it as frequent is domain dependent. This is given by the user as the input and the output are the prediction rules for the time sequences.

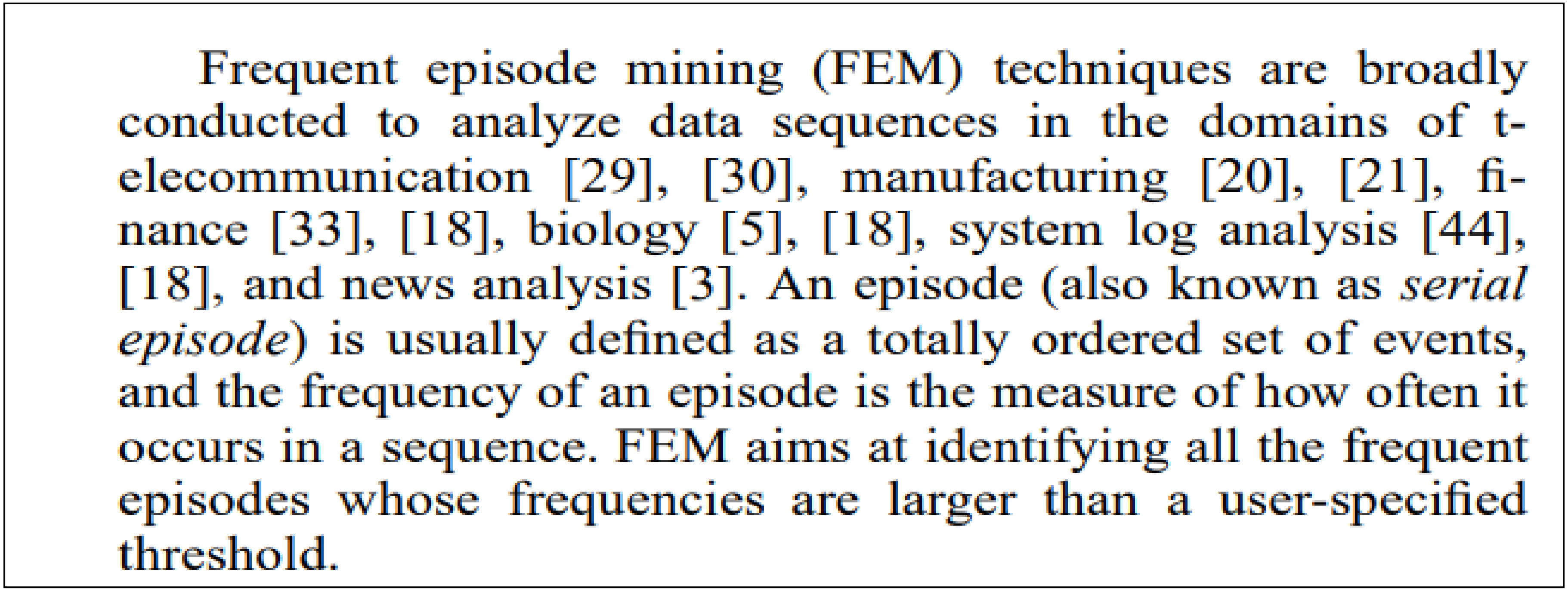
Given a set R of event types, an event is a pair (A, I) where $A \in R$ is an event type and I is an integer, we can calculate the occurrence time of the event. An event sequences of R is a triple $(T_s, T_c,$ S),where $T_s < T_c$ are integers. Ts is the starting time and T_c is the ending time.

S = $\{(A_1 I_1, (A_2,I_2),(A_n>I_n)\}$ is the ordered sequence of events, such that $A_i \in$ R and

$$T_s \leq t_1 \leq T_c \text{ for all } i = 1, 2\ldots n\text{-}1$$

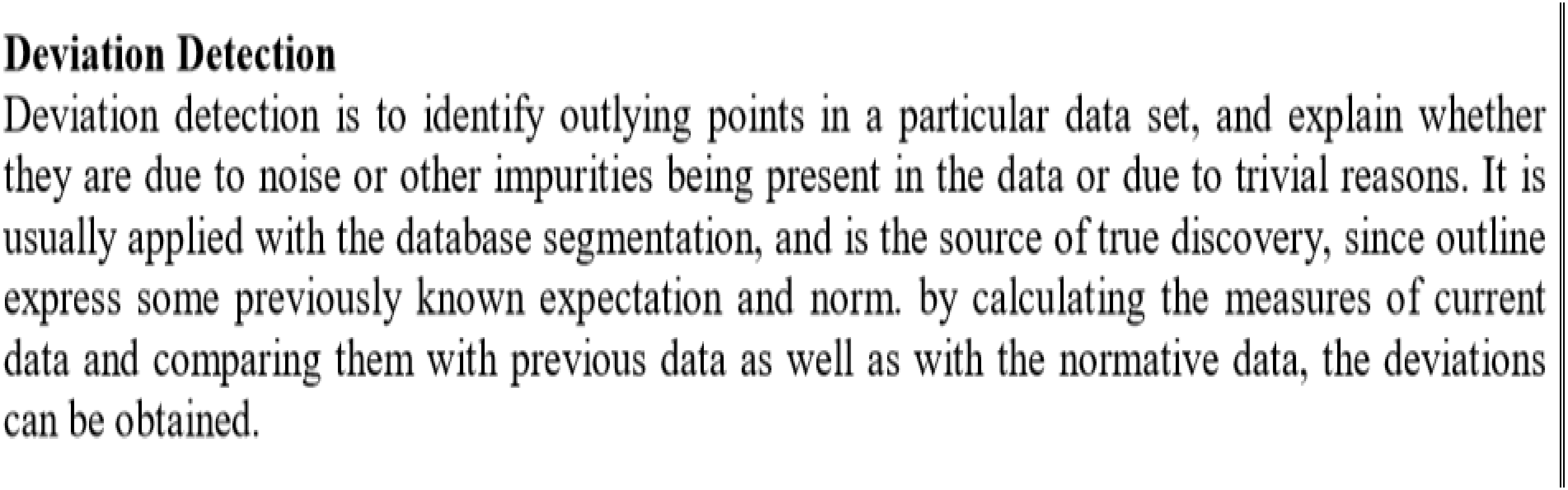
These episodes can be of three types. One is the serial episodes which occur in sequence. The parallel episodes in which there are constraints on the order of the event types A and B given. If the occurrences of A and B precede an occurrence of C, and there is no constraint on the relative order of A and B given then the non-serial and non-parallel episodes which occur in a sequence. The applications include telecommunications, and share market analysis

Discovery Model –Frequent Episode



Frequent episode mining (FEM) techniques are broadly conducted to analyze data sequences in the domains of telecommunication [29], [30], manufacturing [20], [21], finance [33], [18], biology [5], [18], system log analysis [44], [18], and news analysis [3]. An episode (also known as *serial episode*) is usually defined as a totally ordered set of events, and the frequency of an episode is the measure of how often it occurs in a sequence. FEM aims at identifying all the frequent episodes whose frequencies are larger than a user-specified threshold.

Discovery Model –Deviation Detection



Deviation Detection

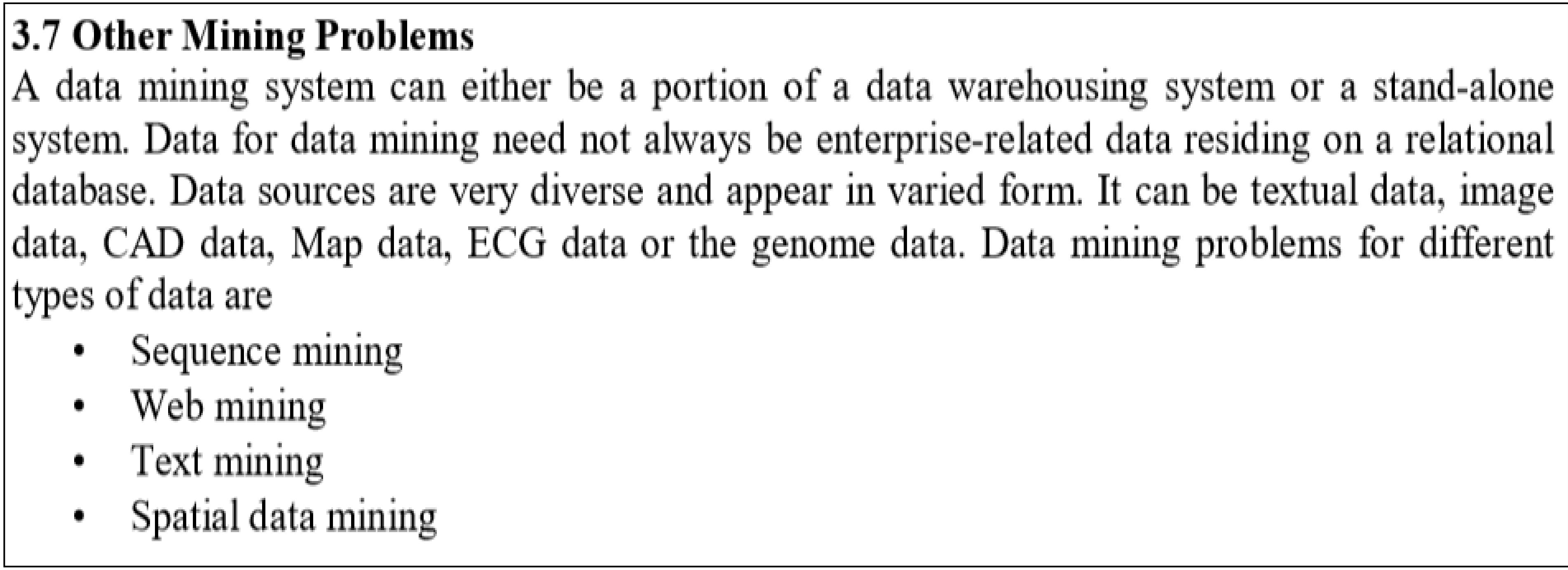
Deviation detection is to identify outlying points in a particular data set, and explain whether they are due to noise or other impurities being present in the data or due to trivial reasons. It is usually applied with the database segmentation, and is the source of true discovery, since outline express some previously known expectation and norm. by calculating the measures of current data and comparing them with previous data as well as with the normative data, the deviations can be obtained.

Discovery Model –Neural Network

A **neural network** is a series of algorithms that endeavors to recognize underlying relationships in a set of **data** through a process that mimics the way the human brain operates. ... **Neural networks** can adapt to changing input; so the **network** generates the best possible result without needing to redesign the output criteria.

Neural networks are designed to work just like the human brain does. In the case of recognizing handwriting or facial recognition, the brain very quickly makes some decisions. For **example**, in the case of facial recognition, the brain might start with "It is female or male?

Other Mining Problems

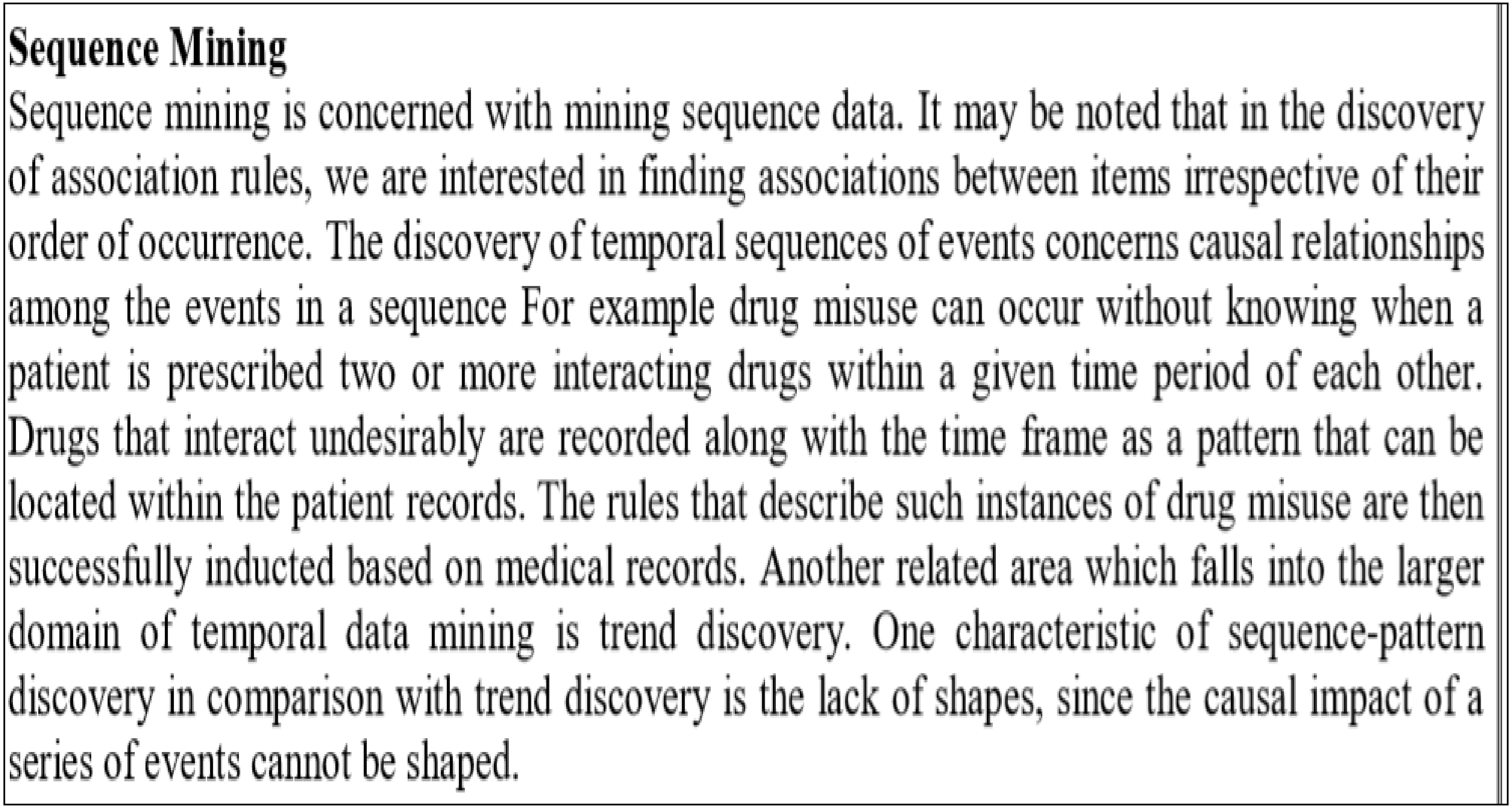


3.7 Other Mining Problems

A data mining system can either be a portion of a data warehousing system or a stand-alone system. Data for data mining need not always be enterprise-related data residing on a relational database. Data sources are very diverse and appear in varied form. It can be textual data, image data, CAD data, Map data, ECG data or the genome data. Data mining problems for different types of data are

- Sequence mining
- Web mining
- Text mining
- Spatial data mining

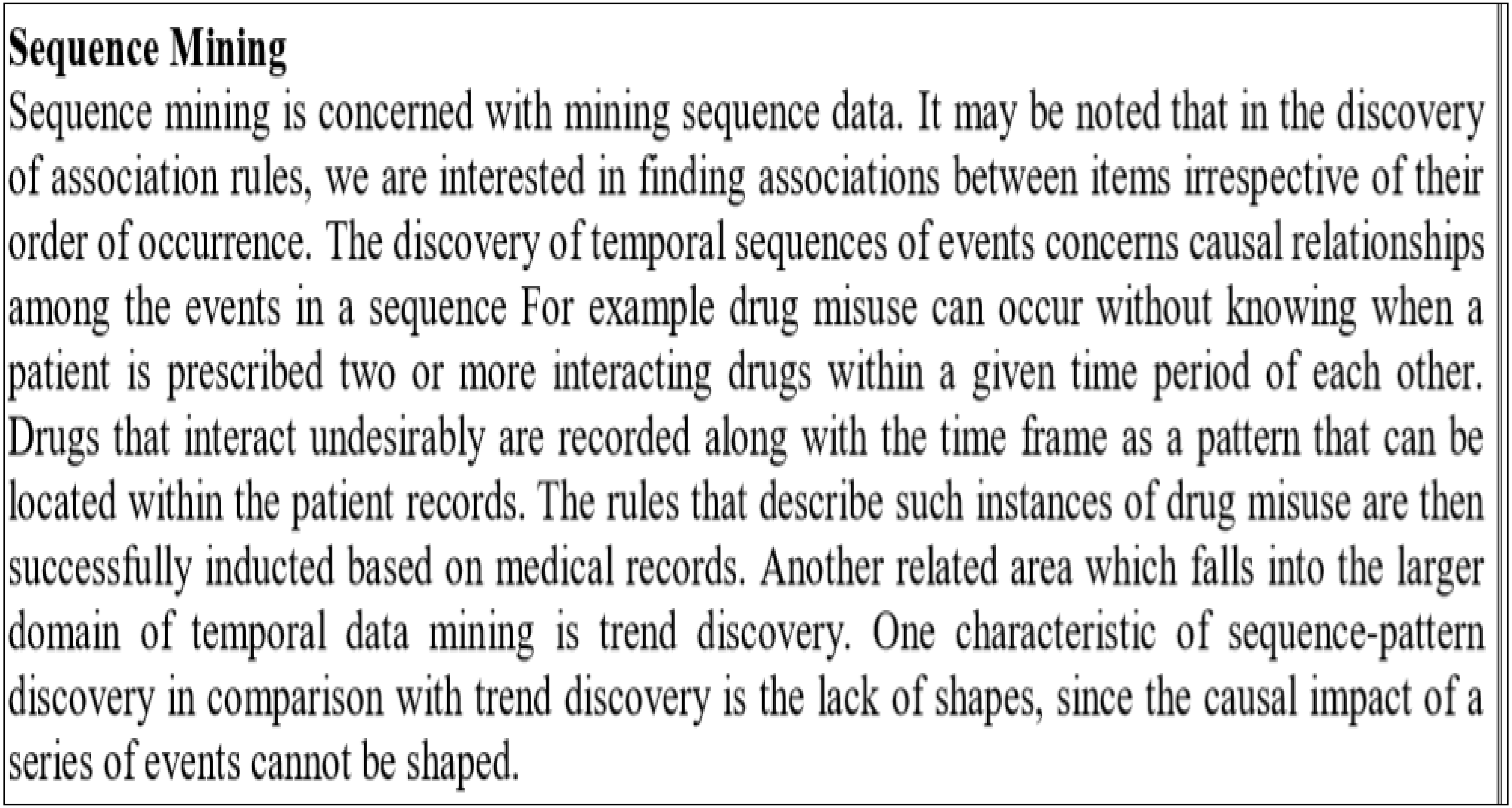
Other Mining Problems



Sequence Mining

Sequence mining is concerned with mining sequence data. It may be noted that in the discovery of association rules, we are interested in finding associations between items irrespective of their order of occurrence. The discovery of temporal sequences of events concerns causal relationships among the events in a sequence For example drug misuse can occur without knowing when a patient is prescribed two or more interacting drugs within a given time period of each other. Drugs that interact undesirably are recorded along with the time frame as a pattern that can be located within the patient records. The rules that describe such instances of drug misuse are then successfully inducted based on medical records. Another related area which falls into the larger domain of temporal data mining is trend discovery. One characteristic of sequence-pattern discovery in comparison with trend discovery is the lack of shapes, since the causal impact of a series of events cannot be shaped.

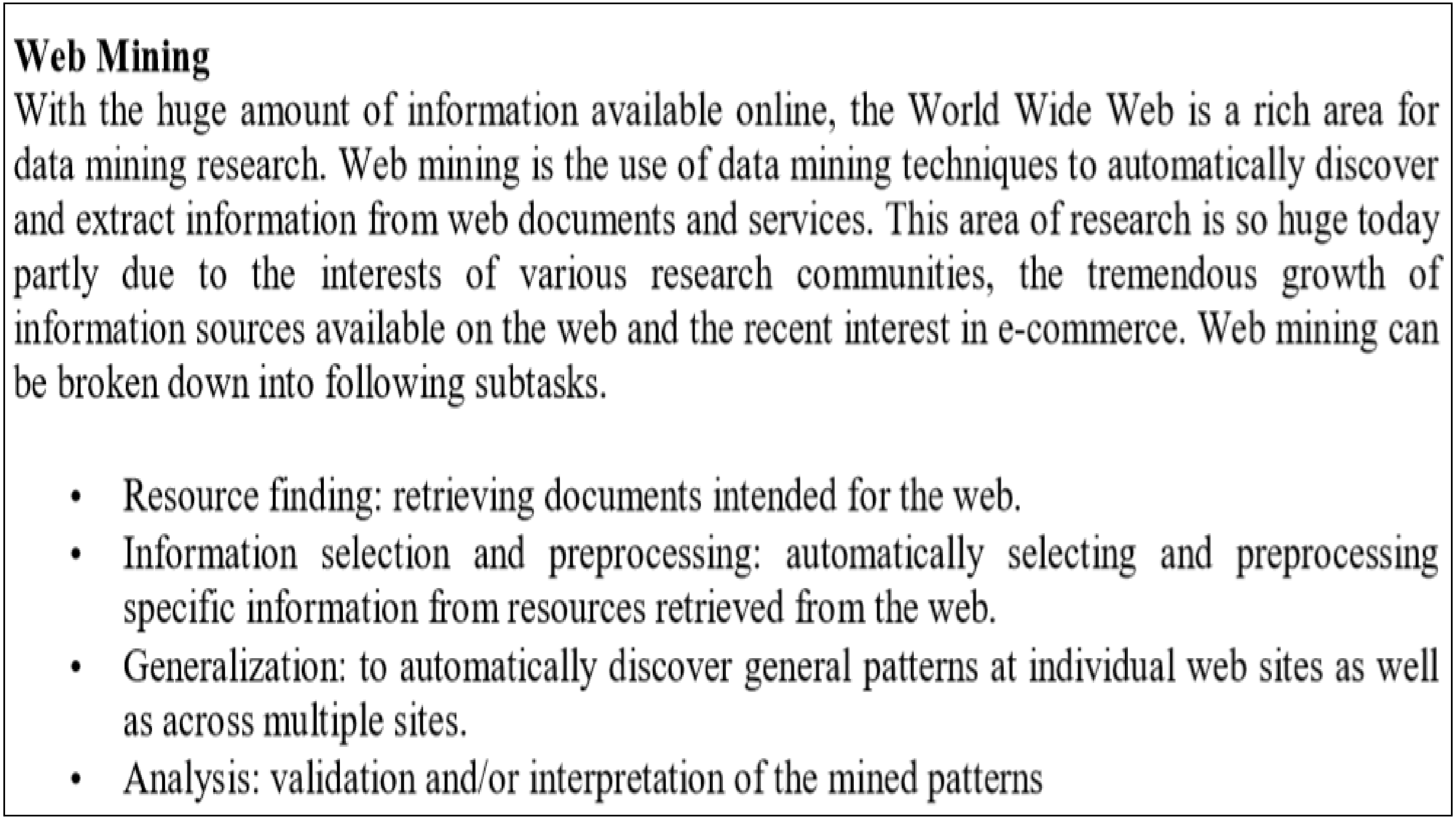
Other Mining Problems



Sequence Mining

Sequence mining is concerned with mining sequence data. It may be noted that in the discovery of association rules, we are interested in finding associations between items irrespective of their order of occurrence. The discovery of temporal sequences of events concerns causal relationships among the events in a sequence For example drug misuse can occur without knowing when a patient is prescribed two or more interacting drugs within a given time period of each other. Drugs that interact undesirably are recorded along with the time frame as a pattern that can be located within the patient records. The rules that describe such instances of drug misuse are then successfully inducted based on medical records. Another related area which falls into the larger domain of temporal data mining is trend discovery. One characteristic of sequence-pattern discovery in comparison with trend discovery is the lack of shapes, since the causal impact of a series of events cannot be shaped.

Other Mining Problems

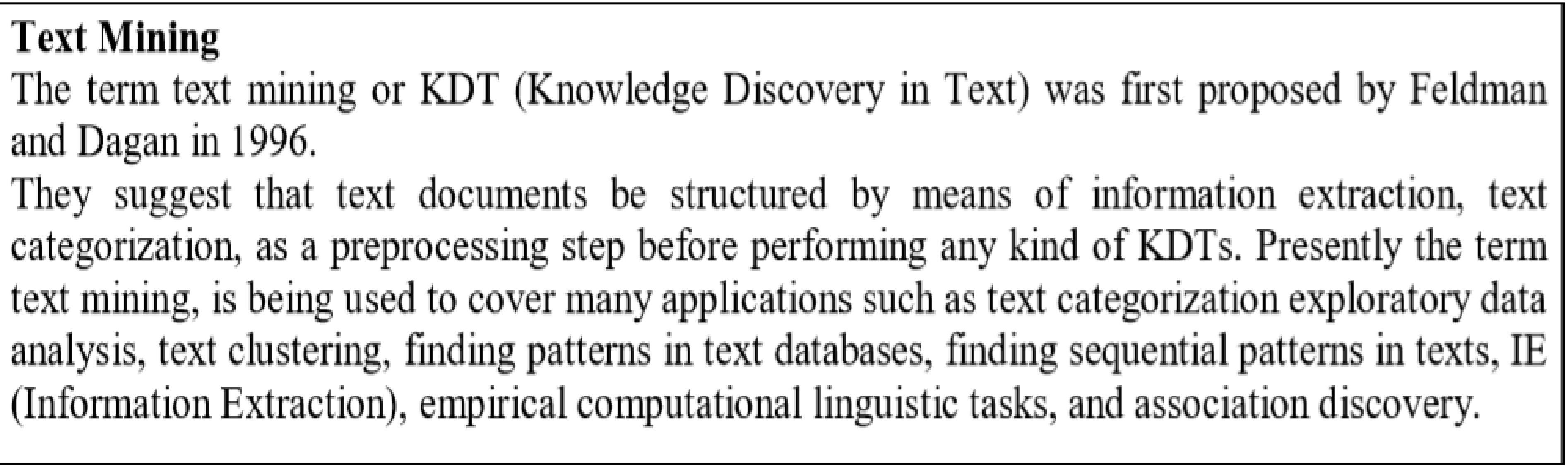


Web Mining

With the huge amount of information available online, the World Wide Web is a rich area for data mining research. Web mining is the use of data mining techniques to automatically discover and extract information from web documents and services. This area of research is so huge today partly due to the interests of various research communities, the tremendous growth of information sources available on the web and the recent interest in e-commerce. Web mining can be broken down into following subtasks.

- Resource finding: retrieving documents intended for the web.
- Information selection and preprocessing: automatically selecting and preprocessing specific information from resources retrieved from the web.
- Generalization: to automatically discover general patterns at individual web sites as well as across multiple sites.
- Analysis: validation and/or interpretation of the mined patterns

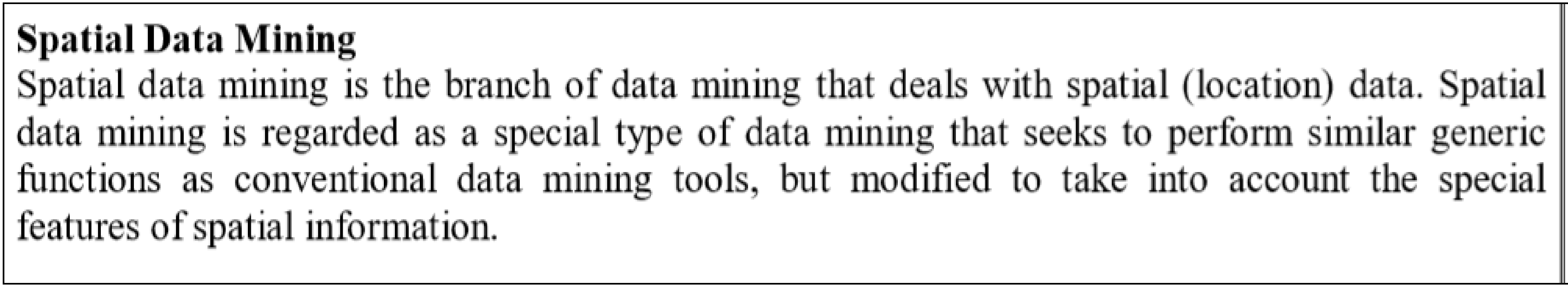
Other Mining Problems



Text Mining

The term text mining or KDT (Knowledge Discovery in Text) was first proposed by Feldman and Dagan in 1996.

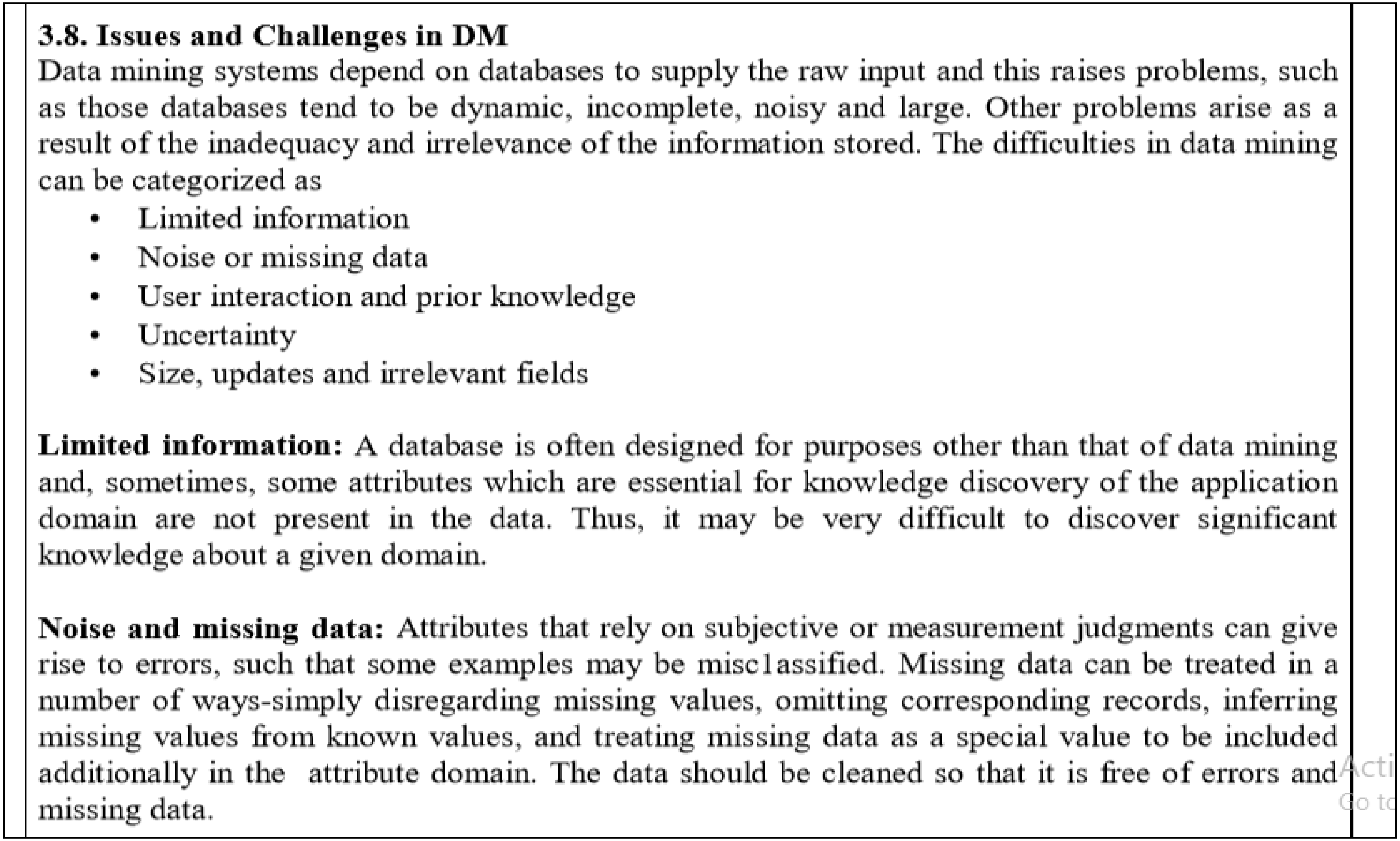
They suggest that text documents be structured by means of information extraction, text categorization, as a preprocessing step before performing any kind of KDTs. Presently the term text mining, is being used to cover many applications such as text categorization exploratory data analysis, text clustering, finding patterns in text databases, finding sequential patterns in texts, IE (Information Extraction), empirical computational linguistic tasks, and association discovery.



Spatial Data Mining

Spatial data mining is the branch of data mining that deals with spatial (location) data. Spatial data mining is regarded as a special type of data mining that seeks to perform similar generic functions as conventional data mining tools, but modified to take into account the special features of spatial information.

Issues and Challenges in DM



3.8. Issues and Challenges in DM

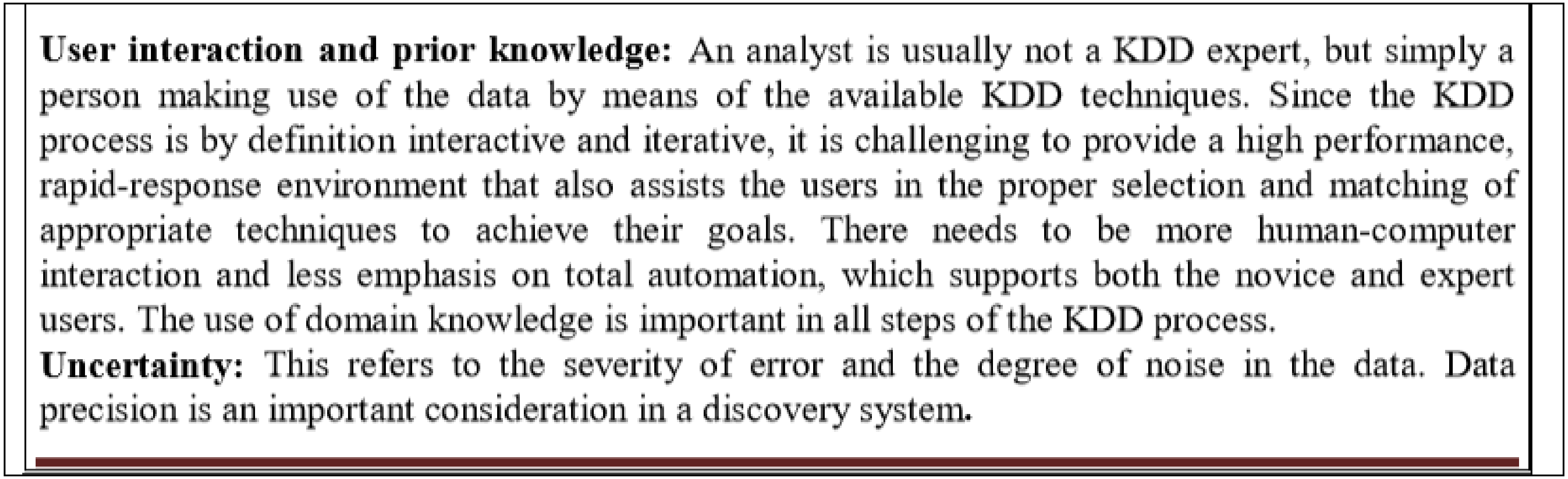
Data mining systems depend on databases to supply the raw input and this raises problems, such as those databases tend to be dynamic, incomplete, noisy and large. Other problems arise as a result of the inadequacy and irrelevance of the information stored. The difficulties in data mining can be categorized as

- Limited information
- Noise or missing data
- User interaction and prior knowledge
- Uncertainty
- Size, updates and irrelevant fields

Limited information: A database is often designed for purposes other than that of data mining and, sometimes, some attributes which are essential for knowledge discovery of the application domain are not present in the data. Thus, it may be very difficult to discover significant knowledge about a given domain.

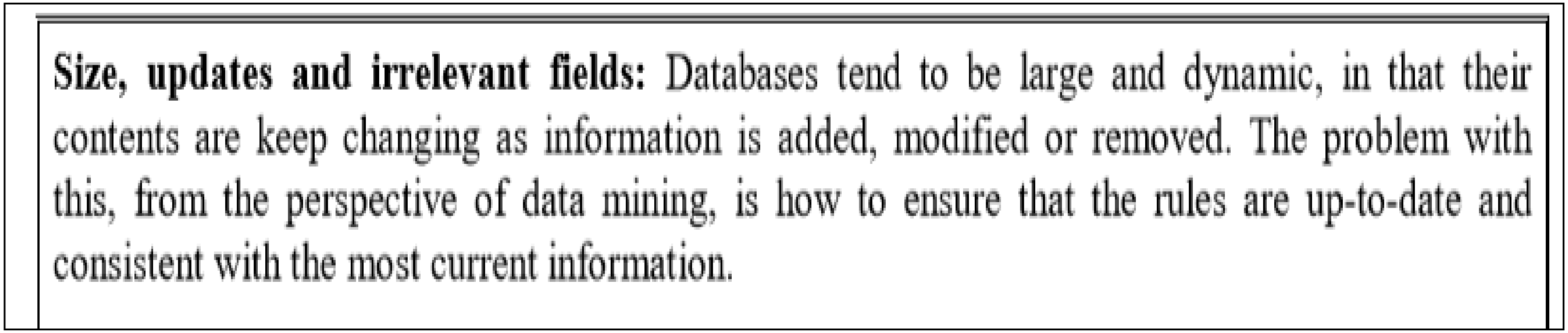
Noise and missing data: Attributes that rely on subjective or measurement judgments can give rise to errors, such that some examples may be misc1assified. Missing data can be treated in a number of ways-simply disregarding missing values, omitting corresponding records, inferring missing values from known values, and treating missing data as a special value to be included additionally in the attribute domain. The data should be cleaned so that it is free of errors and missing data.

Issues and Challenges in DM-Contd



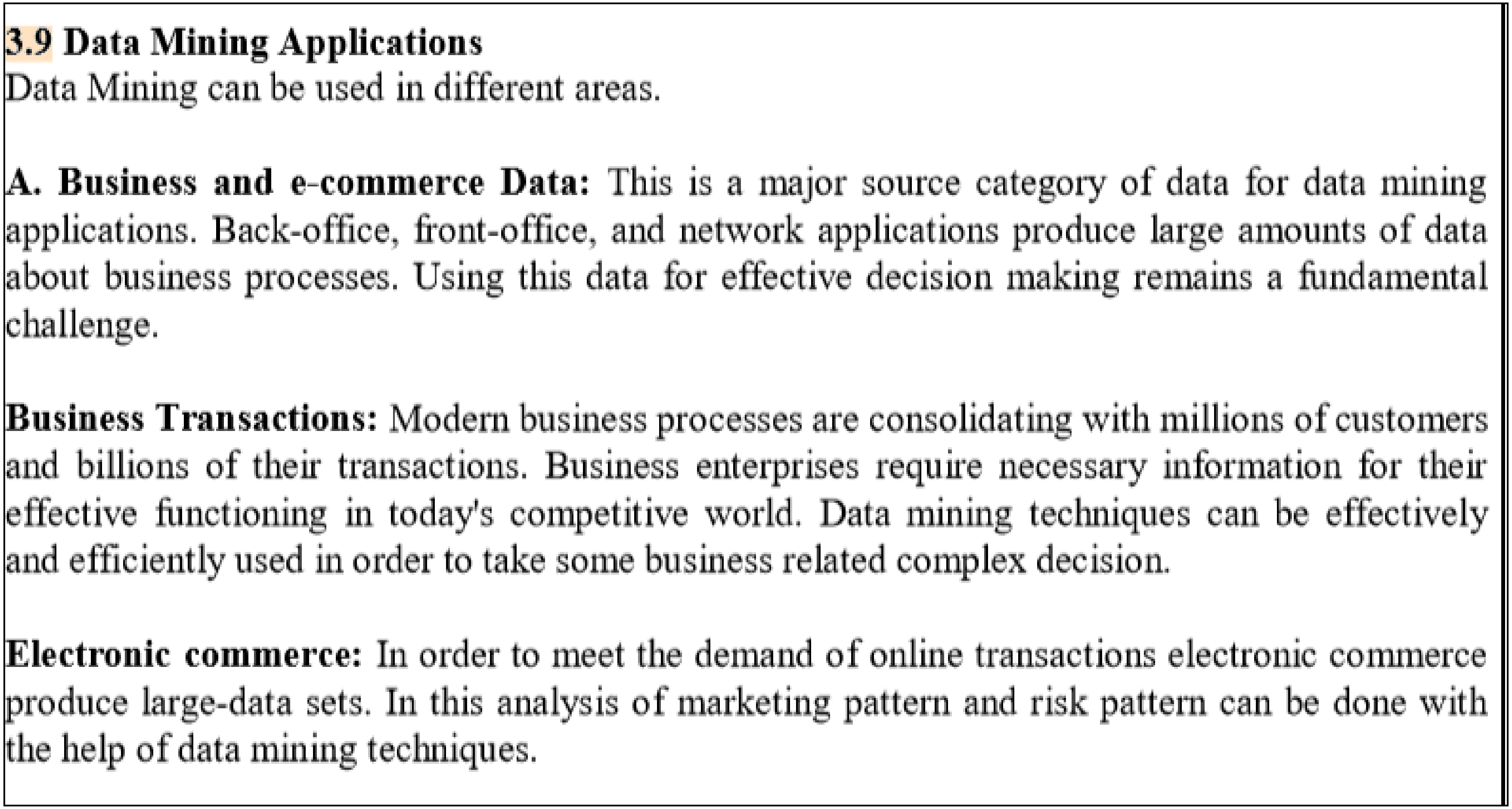
User interaction and prior knowledge: An analyst is usually not a KDD expert, but simply a person making use of the data by means of the available KDD techniques. Since the KDD process is by definition interactive and iterative, it is challenging to provide a high performance, rapid-response environment that also assists the users in the proper selection and matching of appropriate techniques to achieve their goals. There needs to be more human-computer interaction and less emphasis on total automation, which supports both the novice and expert users. The use of domain knowledge is important in all steps of the KDD process.

Uncertainty: This refers to the severity of error and the degree of noise in the data. Data precision is an important consideration in a discovery system.



Size, updates and irrelevant fields: Databases tend to be large and dynamic, in that their contents are keep changing as information is added, modified or removed. The problem with this, from the perspective of data mining, is how to ensure that the rules are up-to-date and consistent with the most current information.

Data Mining Applications



3.9 Data Mining Applications

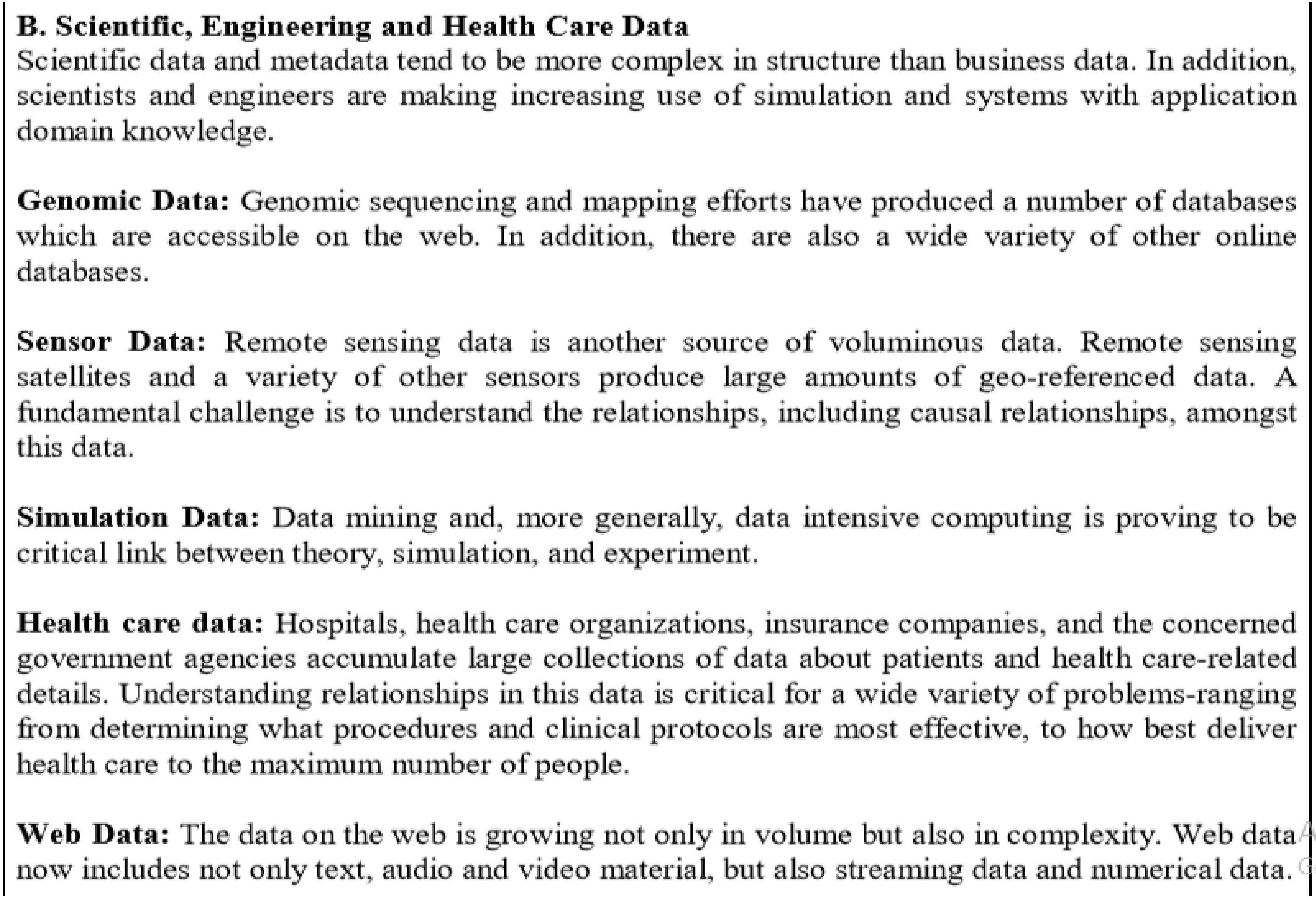
Data Mining can be used in different areas.

A. Business and e-commerce Data: This is a major source category of data for data mining applications. Back-office, front-office, and network applications produce large amounts of data about business processes. Using this data for effective decision making remains a fundamental challenge.

Business Transactions: Modern business processes are consolidating with millions of customers and billions of their transactions. Business enterprises require necessary information for their effective functioning in today's competitive world. Data mining techniques can be effectively and efficiently used in order to take some business related complex decision.

Electronic commerce: In order to meet the demand of online transactions electronic commerce produce large-data sets. In this analysis of marketing pattern and risk pattern can be done with the help of data mining techniques.

Data Mining Applications-CONTD



B. Scientific, Engineering and Health Care Data
Scientific data and metadata tend to be more complex in structure than business data. In addition, scientists and engineers are making increasing use of simulation and systems with application domain knowledge.

Genomic Data: Genomic sequencing and mapping efforts have produced a number of databases which are accessible on the web. In addition, there are also a wide variety of other online databases.

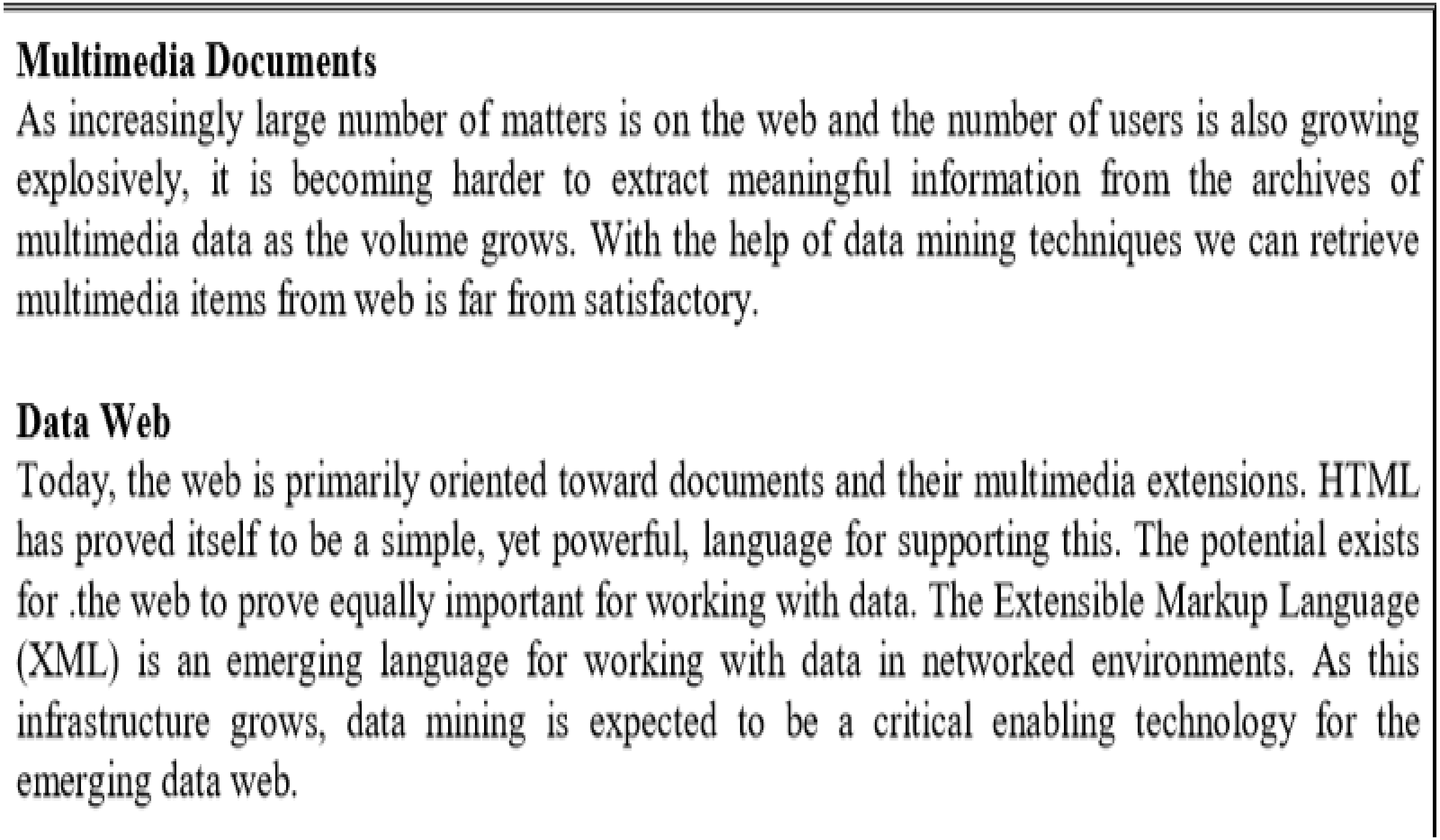
Sensor Data: Remote sensing data is another source of voluminous data. Remote sensing satellites and a variety of other sensors produce large amounts of geo-referenced data. A fundamental challenge is to understand the relationships, including causal relationships, amongst this data.

Simulation Data: Data mining and, more generally, data intensive computing is proving to be critical link between theory, simulation, and experiment.

Health care data: Hospitals, health care organizations, insurance companies, and the concerned government agencies accumulate large collections of data about patients and health care-related details. Understanding relationships in this data is critical for a wide variety of problems-ranging from determining what procedures and clinical protocols are most effective, to how best deliver health care to the maximum number of people.

Web Data: The data on the web is growing not only in volume but also in complexity. Web data now includes not only text, audio and video material, but also streaming data and numerical data.

Data Mining Applications-CONTD



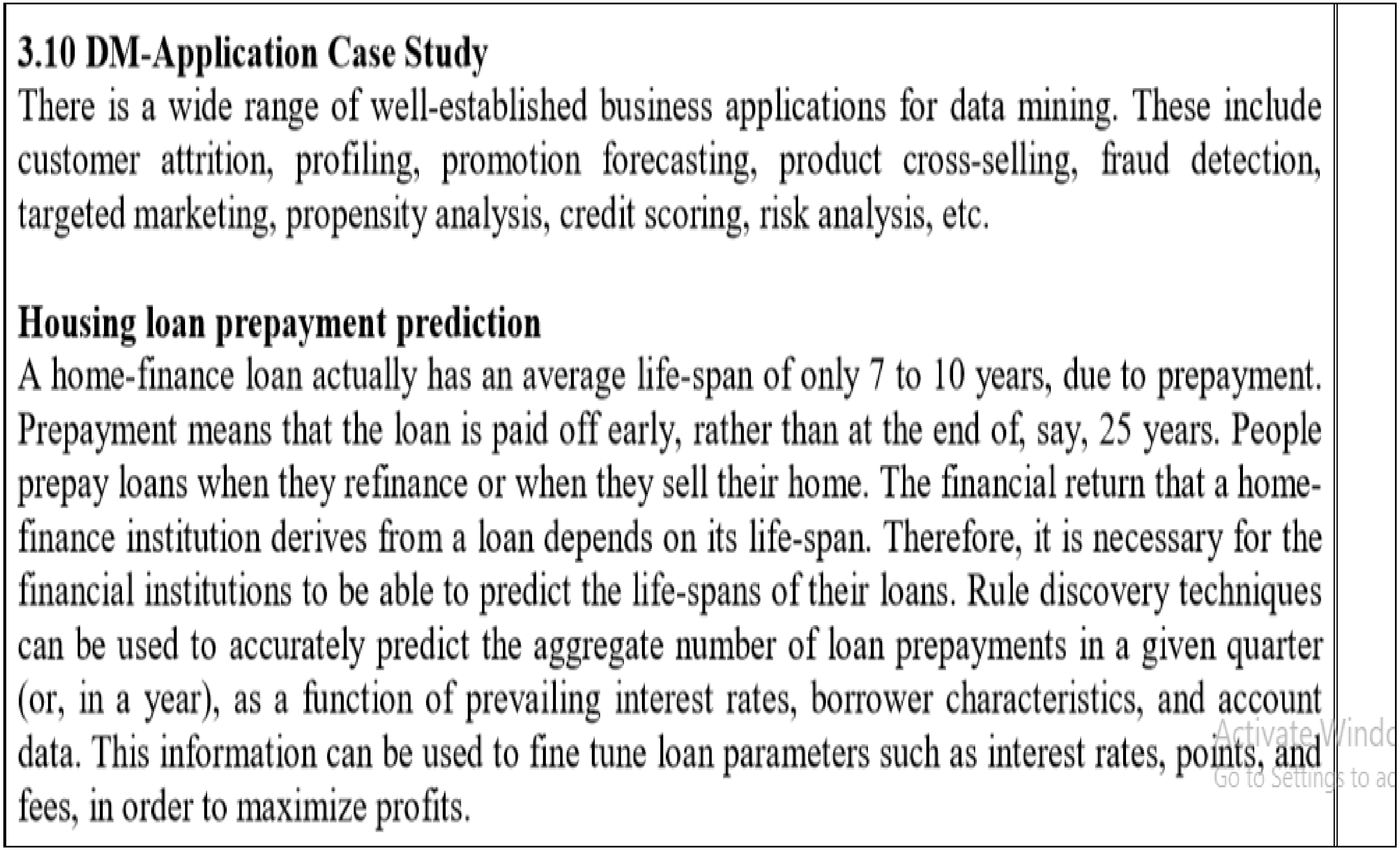
Multimedia Documents

As increasingly large number of matters is on the web and the number of users is also growing explosively, it is becoming harder to extract meaningful information from the archives of multimedia data as the volume grows. With the help of data mining techniques we can retrieve multimedia items from web is far from satisfactory.

Data Web

Today, the web is primarily oriented toward documents and their multimedia extensions. HTML has proved itself to be a simple, yet powerful, language for supporting this. The potential exists for .the web to prove equally important for working with data. The Extensible Markup Language (XML) is an emerging language for working with data in networked environments. As this infrastructure grows, data mining is expected to be a critical enabling technology for the emerging data web.

DM Applications-Case Study



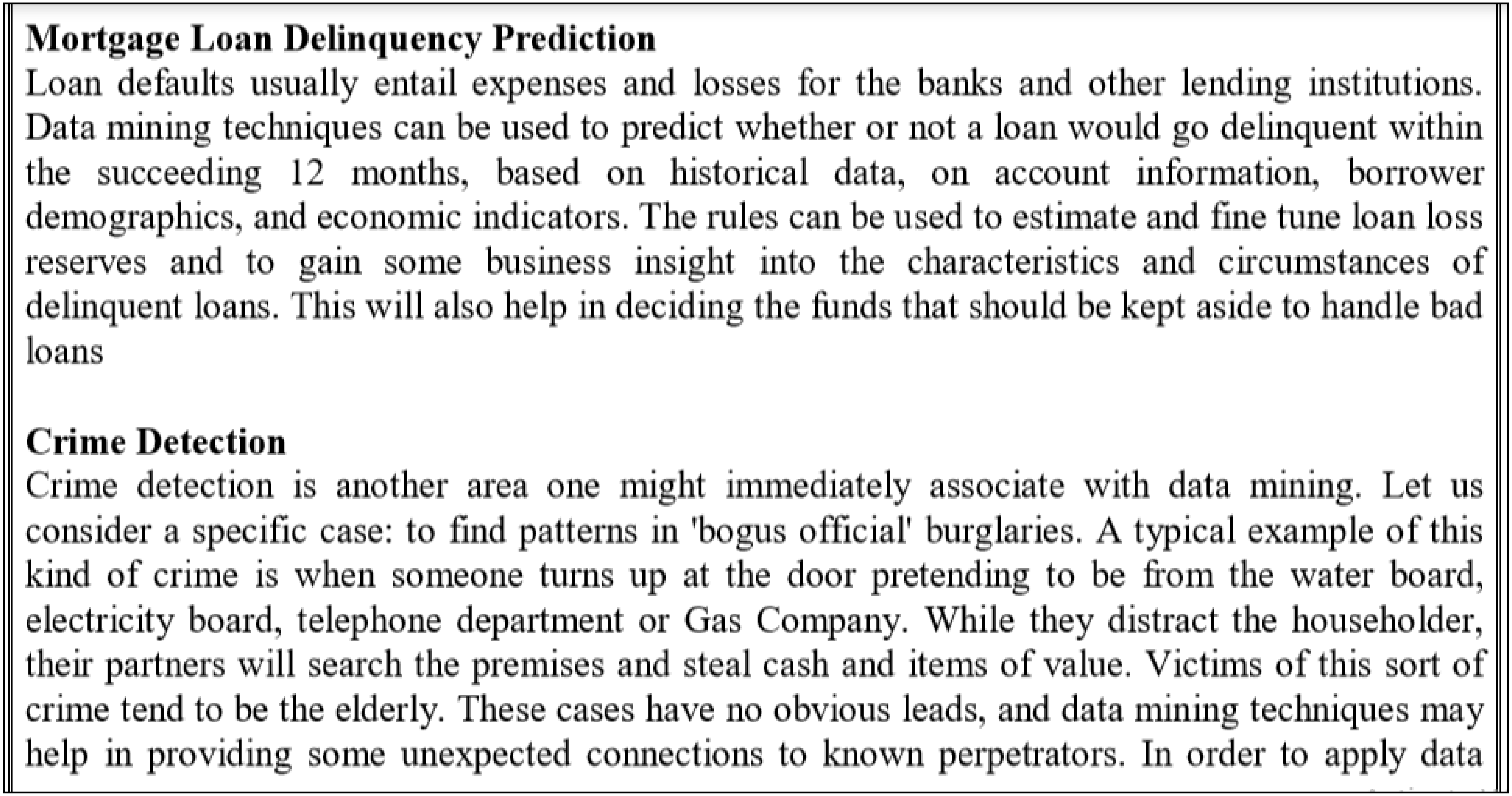
3.10 DM-Application Case Study

There is a wide range of well-established business applications for data mining. These include customer attrition, profiling, promotion forecasting, product cross-selling, fraud detection, targeted marketing, propensity analysis, credit scoring, risk analysis, etc.

Housing loan prepayment prediction

A home-finance loan actually has an average life-span of only 7 to 10 years, due to prepayment. Prepayment means that the loan is paid off early, rather than at the end of, say, 25 years. People prepay loans when they refinance or when they sell their home. The financial return that a home-finance institution derives from a loan depends on its life-span. Therefore, it is necessary for the financial institutions to be able to predict the life-spans of their loans. Rule discovery techniques can be used to accurately predict the aggregate number of loan prepayments in a given quarter (or, in a year), as a function of prevailing interest rates, borrower characteristics, and account data. This information can be used to fine tune loan parameters such as interest rates, points, and fees, in order to maximize profits.

DM Applications-Case Study-Contd



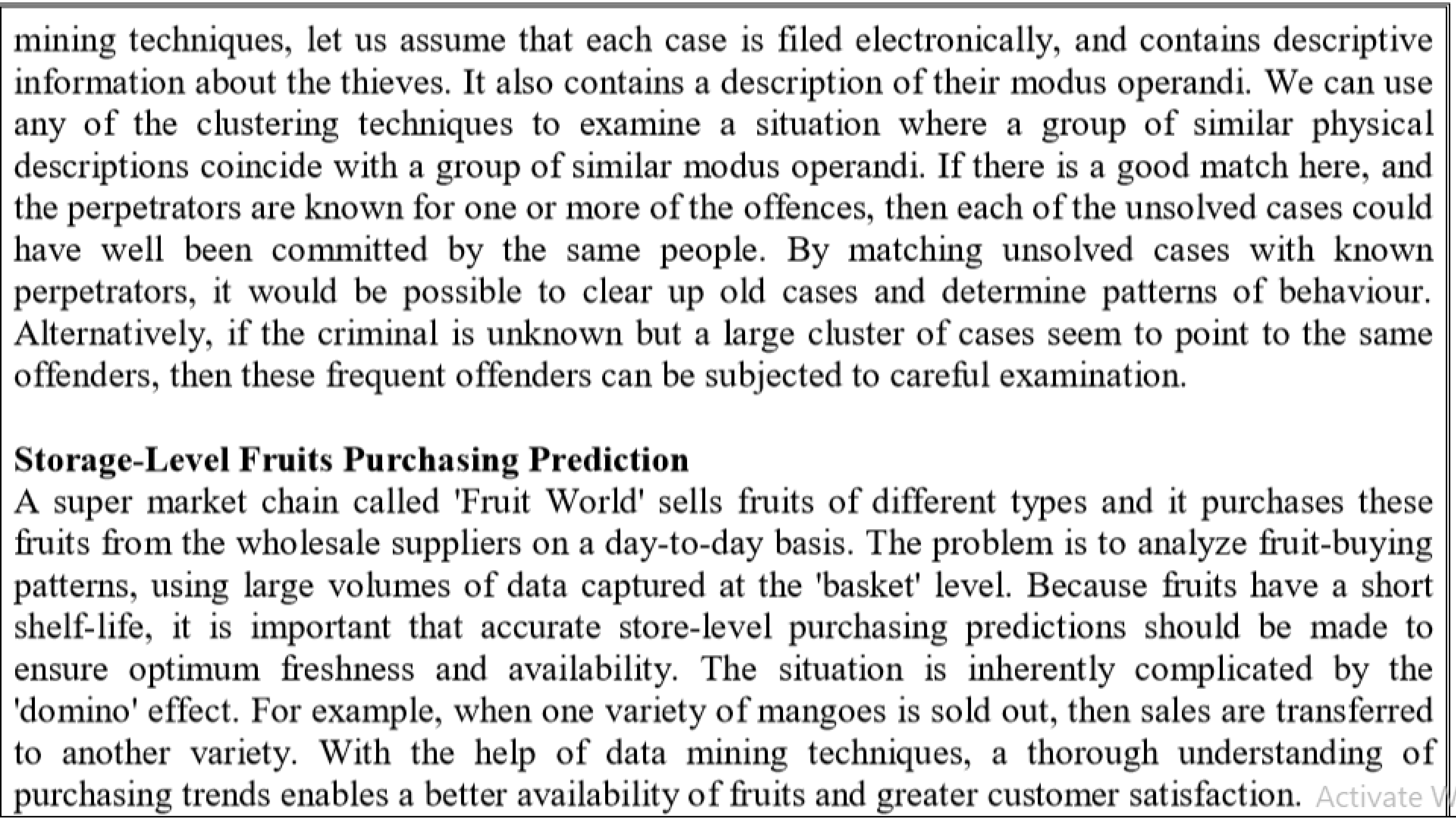
Mortgage Loan Delinquency Prediction

Loan defaults usually entail expenses and losses for the banks and other lending institutions. Data mining techniques can be used to predict whether or not a loan would go delinquent within the succeeding 12 months, based on historical data, on account information, borrower demographics, and economic indicators. The rules can be used to estimate and fine tune loan loss reserves and to gain some business insight into the characteristics and circumstances of delinquent loans. This will also help in deciding the funds that should be kept aside to handle bad loans

Crime Detection

Crime detection is another area one might immediately associate with data mining. Let us consider a specific case: to find patterns in 'bogus official' burglaries. A typical example of this kind of crime is when someone turns up at the door pretending to be from the water board, electricity board, telephone department or Gas Company. While they distract the householder, their partners will search the premises and steal cash and items of value. Victims of this sort of crime tend to be the elderly. These cases have no obvious leads, and data mining techniques may help in providing some unexpected connections to known perpetrators. In order to apply data

DM Applications-Case Study-Contd

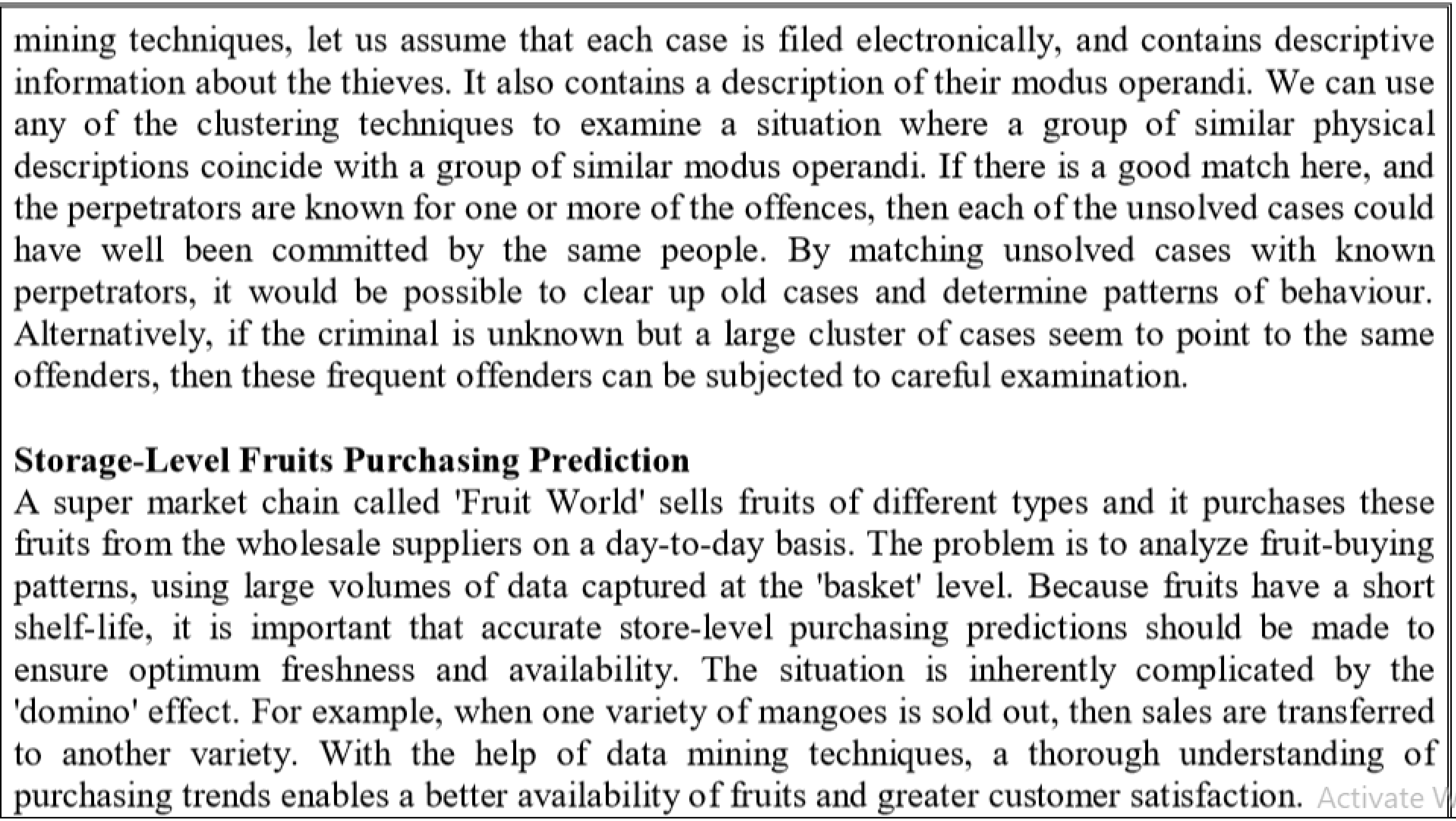


mining techniques, let us assume that each case is filed electronically, and contains descriptive information about the thieves. It also contains a description of their modus operandi. We can use any of the clustering techniques to examine a situation where a group of similar physical descriptions coincide with a group of similar modus operandi. If there is a good match here, and the perpetrators are known for one or more of the offences, then each of the unsolved cases could have well been committed by the same people. By matching unsolved cases with known perpetrators, it would be possible to clear up old cases and determine patterns of behaviour. Alternatively, if the criminal is unknown but a large cluster of cases seem to point to the same offenders, then these frequent offenders can be subjected to careful examination.

Storage-Level Fruits Purchasing Prediction

A super market chain called 'Fruit World' sells fruits of different types and it purchases these fruits from the wholesale suppliers on a day-to-day basis. The problem is to analyze fruit-buying patterns, using large volumes of data captured at the 'basket' level. Because fruits have a short shelf-life, it is important that accurate store-level purchasing predictions should be made to ensure optimum freshness and availability. The situation is inherently complicated by the 'domino' effect. For example, when one variety of mangoes is sold out, then sales are transferred to another variety. With the help of data mining techniques, a thorough understanding of purchasing trends enables a better availability of fruits and greater customer satisfaction.

DM Applications-Case Study-Contd

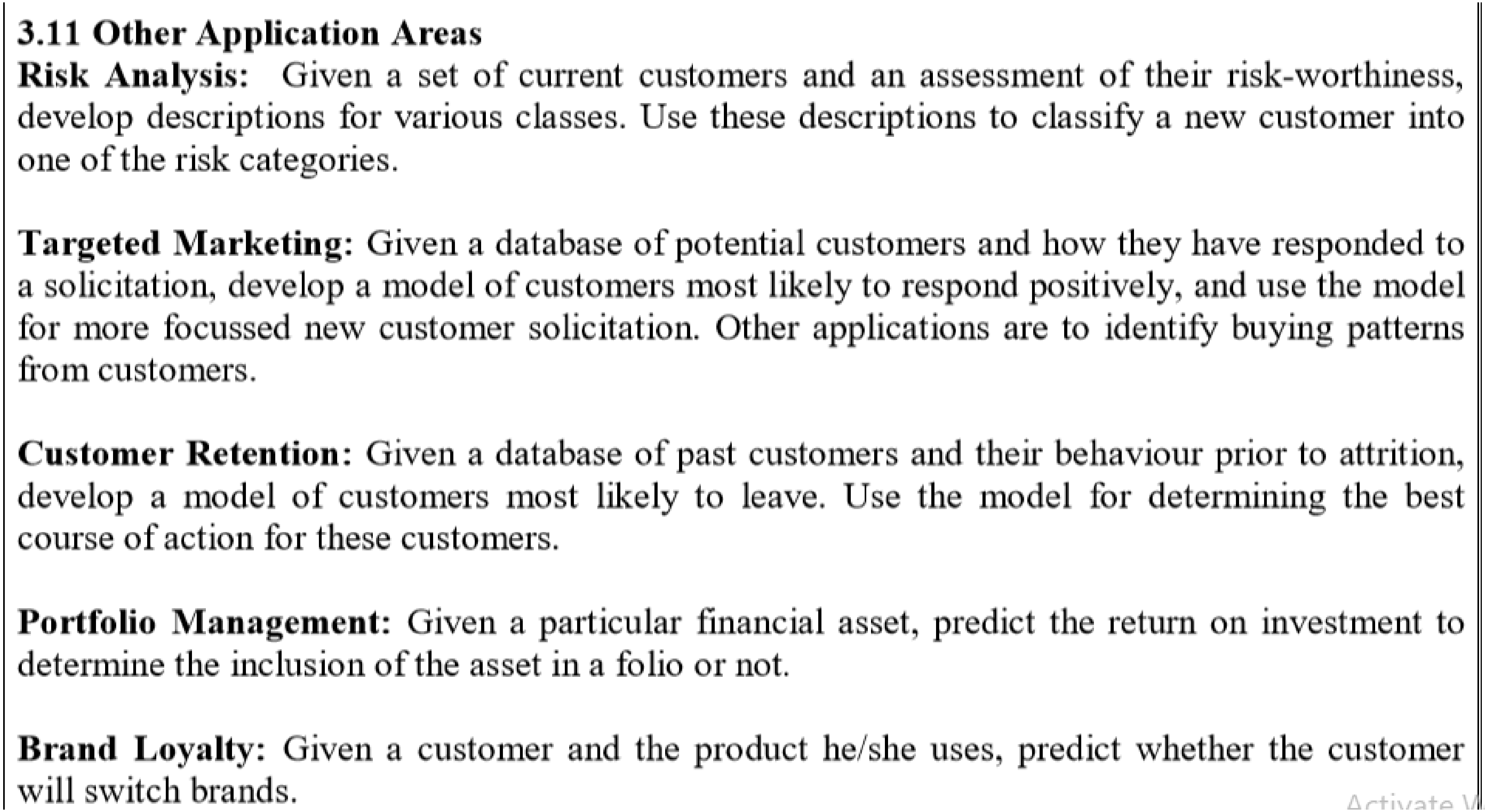


mining techniques, let us assume that each case is filed electronically, and contains descriptive information about the thieves. It also contains a description of their modus operandi. We can use any of the clustering techniques to examine a situation where a group of similar physical descriptions coincide with a group of similar modus operandi. If there is a good match here, and the perpetrators are known for one or more of the offences, then each of the unsolved cases could have well been committed by the same people. By matching unsolved cases with known perpetrators, it would be possible to clear up old cases and determine patterns of behaviour. Alternatively, if the criminal is unknown but a large cluster of cases seem to point to the same offenders, then these frequent offenders can be subjected to careful examination.

Storage-Level Fruits Purchasing Prediction

A super market chain called 'Fruit World' sells fruits of different types and it purchases these fruits from the wholesale suppliers on a day-to-day basis. The problem is to analyze fruit-buying patterns, using large volumes of data captured at the 'basket' level. Because fruits have a short shelf-life, it is important that accurate store-level purchasing predictions should be made to ensure optimum freshness and availability. The situation is inherently complicated by the 'domino' effect. For example, when one variety of mangoes is sold out, then sales are transferred to another variety. With the help of data mining techniques, a thorough understanding of purchasing trends enables a better availability of fruits and greater customer satisfaction.

Other Application Areas



3.11 Other Application Areas

Risk Analysis: Given a set of current customers and an assessment of their risk-worthiness, develop descriptions for various classes. Use these descriptions to classify a new customer into one of the risk categories.

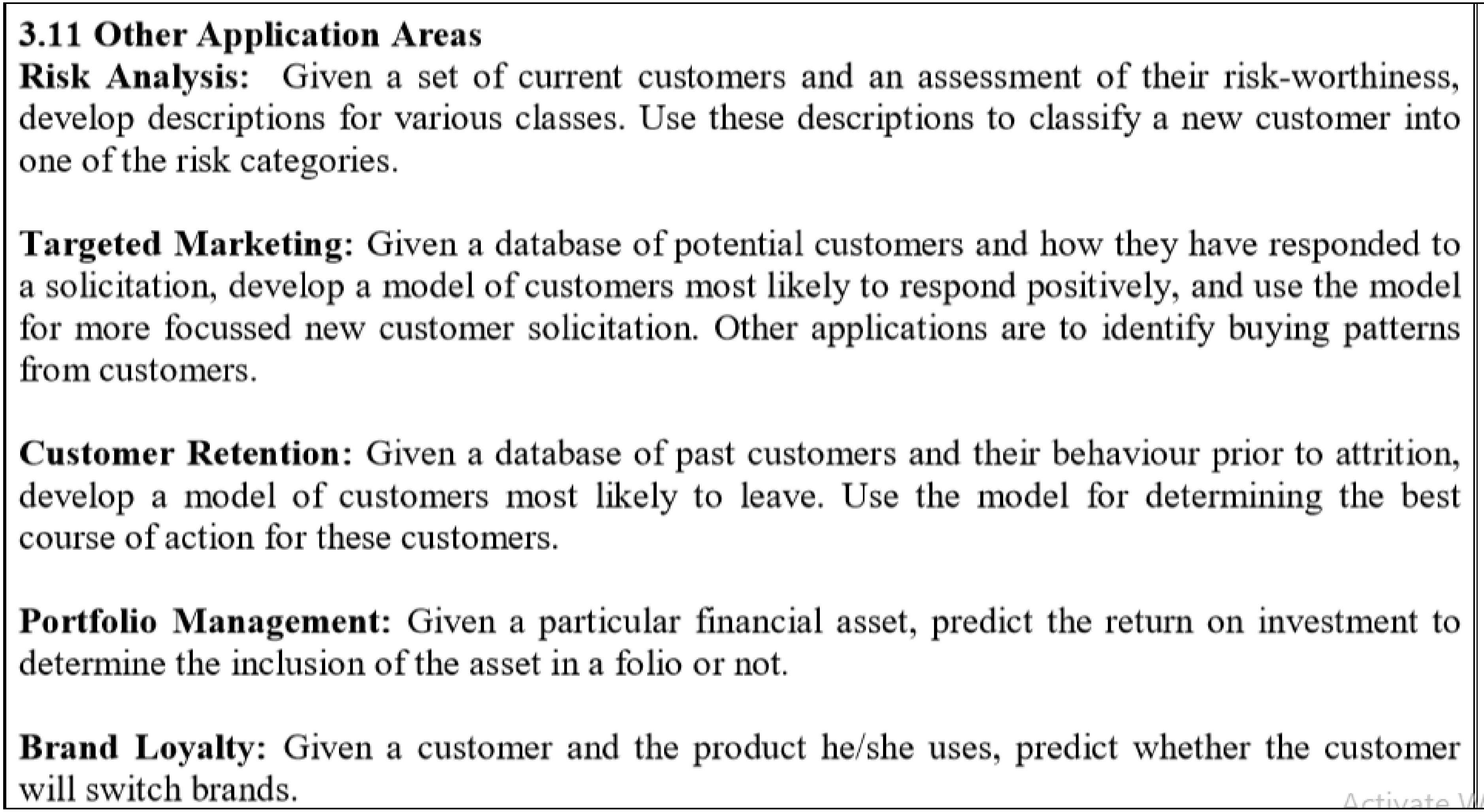
Targeted Marketing: Given a database of potential customers and how they have responded to a solicitation, develop a model of customers most likely to respond positively, and use the model for more focussed new customer solicitation. Other applications are to identify buying patterns from customers.

Customer Retention: Given a database of past customers and their behaviour prior to attrition, develop a model of customers most likely to leave. Use the model for determining the best course of action for these customers.

Portfolio Management: Given a particular financial asset, predict the return on investment to determine the inclusion of the asset in a folio or not.

Brand Loyalty: Given a customer and the product he/she uses, predict whether the customer will switch brands.

Other Application Areas



3.11 Other Application Areas

Risk Analysis: Given a set of current customers and an assessment of their risk-worthiness, develop descriptions for various classes. Use these descriptions to classify a new customer into one of the risk categories.

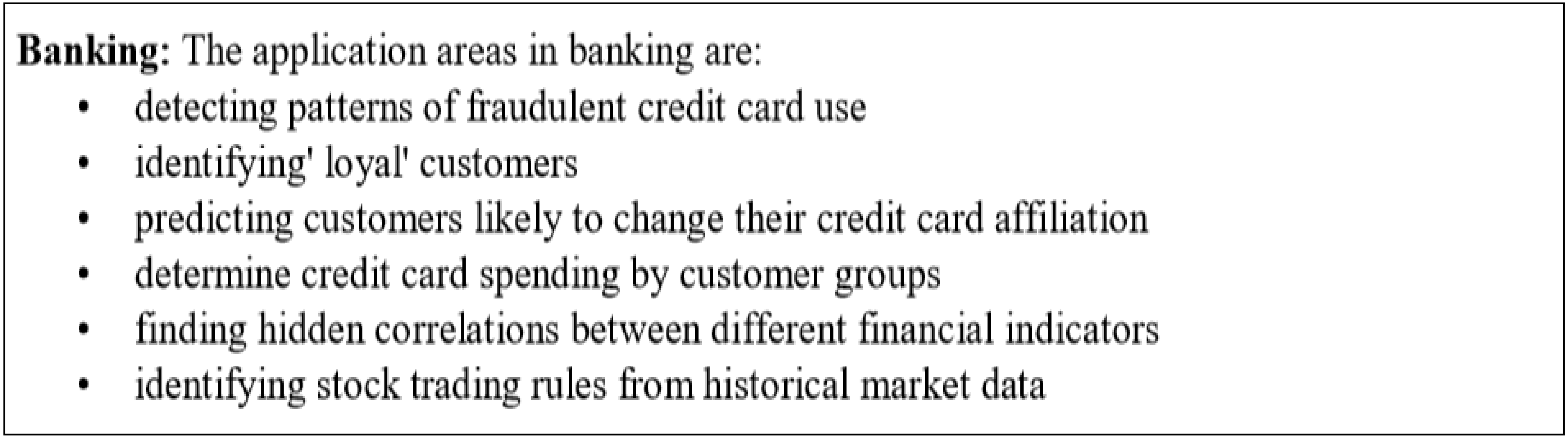
Targeted Marketing: Given a database of potential customers and how they have responded to a solicitation, develop a model of customers most likely to respond positively, and use the model for more focussed new customer solicitation. Other applications are to identify buying patterns from customers.

Customer Retention: Given a database of past customers and their behaviour prior to attrition, develop a model of customers most likely to leave. Use the model for determining the best course of action for these customers.

Portfolio Management: Given a particular financial asset, predict the return on investment to determine the inclusion of the asset in a folio or not.

Brand Loyalty: Given a customer and the product he/she uses, predict whether the customer will switch brands.

Other Application Areas-Contd



Banking: The application areas in banking are:

- detecting patterns of fraudulent credit card use
- identifying' loyal' customers
- predicting customers likely to change their credit card affiliation
- determine credit card spending by customer groups
- finding hidden correlations between different financial indicators
- identifying stock trading rules from historical market data