Overview

Advancements in internet connection and data storage technology have enabled the storage of vast amounts of data at a relatively low cost in commercial or scientific research labs' data warehouses or elsewhere. Consequently, these advancements also imply that such data has been buried within its knowledge that can be vital to a company's growth. It can also be within a knowledge that could lead to significant discoveries in science, such as accurately predicting the weather and natural disasters and identifying the causes of and possible treatments for fatal illnesses. However, most of this data remains merely stored and unexamined - reflecting a phenomenon of 'data-rich but knowledge poor.

Knowledge discovery is generally the extraction of previously unknown and potentially useful information from databases. It is a process in which Data Mining constitutes the central element.

For this week's topic, we will look at Data Mining from several aspects, ranging from the core concept of Data Mining as a process to its applications.

1.1 What is Data Mining

Our study of Data Mining focusses on the contexts surrounding this definition:

> Data Mining is a process to discover patterns and relationships in data that involves various techniques from multiple disciplines, such as database systems, machine learning, information science, visualization, and statistical methods.

We will look at the detailed explanation of Data Mining in the later sections.

1.1.1 The Origins of Data Mining

There is no universal agreement on the definition of data mining or even what constitutes data mining. For example, is data mining a machine learning technique, predictive modeling, or a form of statistics enhanced with learning theory, or is it a new concept?

Most data mining problems and corresponding solutions have roots in classical data analysis. However, data mining has its bases in various disciplines, of which the two essentials are statistics and machine learning. Statistics has its origins in mathematics; hence, there has been a focus on mathematical rigor, an inclination to establish that something is sensible on theoretical grounds before testing it in practice. In contrast, the machine learning community has its origins in computer practice. This base has led to a practical orientation to experiment with something to explore how well it performs without deferring for formal proof of significance. Suppose the establishment in mathematics and formalisations is one of the notable differences between statistical and machine learning approaches to data mining; another difference is their comparative emphasis on models and algorithms.

The notion or concept of a model almost entirely drives modern statistics. This notion is a hypothesized conceptual structure that could have led to the data. In terms of the statistical emphasis on models, machine learning emphasizes algorithms. Consequently, the word “learning” comprises the notion of a process where the logic flow is coded in an implicit algorithm.

Data Mining's basic modeling principles also have origins in control theory, applied mainly to engineering systems and industrial processes. In this theory, determining a mathematical model for an unknown (a.k.a. target) system by observing its input-output data is commonly known as system identification. The purposes of system identification are numerous. However, from the perspective of Data Mining, an essential purpose is to predict a system's behavior and explain the interrelationships between the attributes or variables.

System identification is not a one-directional process: Structure and parameter identification must be performed repeatedly until a satisfactory model is encountered. If data analysts do not have any a priori knowledge about the target system, structure identification becomes challenging, consequently selecting the structure by trial and error.

In most target systems, when applying a data-mining approach, these structures are unknown, or they are so complex that it is impossible to obtain an acceptable mathematical model. Therefore, new techniques were developed for parameter identification, and they are a part of the spectra of Data Mining. Ultimately, we can

differentiate between how the terms "model" and "pattern" are interpreted in Data Mining. A model is a large-scale structure representing relationships over many or all cases of the data. In contrast, a pattern is a local structure of the model, satisfied by a few cases or in a small part of a data space.

1.1.2 What is NOT Data Mining?

While data mining covers a broad set of computing techniques and disciplines, not all data analytical methods are considered data mining processes. Data mining is usually applied (but is not limited to) to large data sets scenarios. Data mining also undergoes a standardised process of data exploration, pre-processing, modeling, and evaluation for knowledge discovery.

Some commonly used techniques to derive data patterns are not considered data mining, even though they use large data sets. The following lists what Data Mining is not.

- Computing algorithm or program

  Writing or using a computing program to derive data patterns or prediction results is not Data Mining.

- Descriptive statistics.

  Counting mean, standard deviation, and other descriptive statistics compute the aggregate structure of a data set. This computation is vital to understanding any data set, but computing these statistics is not considered Data Mining. Regardless, they are performed in the exploration phase of a Data Mining process.

- Hypothesis testing.

  In confirmatory data analysis, data is tested if a hypothesis has significant evidence to support it. On the other hand, data mining is a process where many hypotheses may be developed and tested based on collected data sets. Since the Data Mining phases are iterative, users can refine the testing in each phase. That means it is one of the data analysis methods in modelling phase of Data Mining.

- Queries

Information retrieval systems may use data modeling techniques like clustering to index large databases. However, querying and generating the result is not considered Data Mining. Instead, the data query techniques (either on structured SQL or unstructured NoSQL databases) are used in the phases of a data mining process.

- Exploratory visualisation

  A computerized reporting system outputting data patterns with exploratory visualization capable of displaying interactive charts, tables, or trends is not Data Mining. Describing data using a visualization application enables users to discover data patterns and relationships in large data sets. Like descriptive statistics, they are essential in the data understanding and preprocessing phases in Data Mining process.

- Machine learning technique

  Performing data analysis using a supervised or unsupervised machine learning method(s), for example, Decision Tree, Clustering, or Neural Network, is not Data Mining.

- Dimensional slicing

  Business intelligence and online analytical processing (OLAP) applications, predominant in business environments, primarily deliver data analysis via dimensional slicing, filtering, and pivoting. OLAP data analysis is supported by a data warehousing schema (i.e., data warehouse) design where the data is structured as dimensions and quantitative facts for measurement. While this design technique is useful in providing patterns in data, it is considered information retrieval but not Data Mining.

1.1.3 Fallacies of What Data Mining Can Accomplish

Based on the definitive context given in previous sections, generally knowing what data mining is, we can further understand the common fallacies of what data mining can accomplish as follows.

Fallacy 1. There are data mining tools that we can automate the processing of data sets and find answers to problems.

Reality: There are no automatic data mining tools that will mechanically solve problems without human intervention. Instead, data mining is a process.

Fallacy 2. The data mining process is autonomous, requiring little or no human oversight.

Reality: Without skilled human oversight, data mining tools' arbitrary use will likely generate incorrect solutions to the inappropriate question applied to the inaccurate data. Further, the erroneous analysis brings more harm than no analysis since it leads to decision-making for deployments or policy recommendations that will probably be expensive failures. Finally, even after the model is deployed, the introduction of new data often requires updating the model and domain expert(s) to guide the updates. Thus, human analysts must evaluate continuous quality monitoring and other assessment measures.

Fallacy 3. Data mining can recognise the causes of business or research problems.

Reality: The knowledge discovery process will help businesses or researchers to find patterns of data behaviour. However, it depends on humans to determine the causes based on the patterns identified.

Fallacy 4. Data mining will automatically clean up and process a dirty data set.

Reality: Instead, data pre-processing and cleaning up will not be performed automatically. In addition, as a preliminary phase in data mining, data pre-processing often deals with historical data that has not been inspected or utilized in an extended period. Therefore, organizations beginning a new data-mining process will often face the problem of data that is of bad quality for analysis and needs considerable fixing and updating.

Fallacy 5. Data mining provides positive and useful results for application deployments or policy recommendations.

Reality: There is no guarantee of positive and useful results when mining data for actionable knowledge. Data mining is not a cure-all for problem-solving. However, when used appropriately by whom understand the problem domain

and models involved, the data requirements, and the overall project objectives, data mining can certainly provide actionable and fruitful results.

1.1.4 The Case for Data Mining

In past decades, we have witnessed an enormous collection of data with the progress of information technology, internet networks, and business models it enables. Moreover, the online applications built on these advancements like social networking and mobile services unleash enormous complex and heterogeneous data waiting to be explored. However, unfortunately, conventional data analysis techniques such as hypothesis testing, dimensional slicing, and descriptive statistics can only bring users limited knowledge discovery. This limitation implies the need for a paradigm:

- To handle a massive volume of data, explore the patterns and interrelationships of thousands of magnitudes as attributes or variables, and deploy data modeling techniques to derive useful insights from the data sets.
- A standardized process, tools, and techniques to systematically and intelligently assist humans to process these data and extracting useful information.

1.1.4.1 Data Volume

The enormous amount of data captured by organizations is exponentially increasing due to the internet and information technology advancements. Furthermore, the rapid decrease in storage and data processing costs and businesses' need to extract data for possible leverage forms a solid motivation to store more additional data. A rapid growth in the volume of data exposes the limits of current data analysis methodologies. Usually, in implementations, the time to construct a generalization model is crucial, and data volume plays a significant factor in deciding the time to development and deployment.

1.1.4.2 Dimensions of Data

One of the primary characteristics of the Big Data phenomenon is high variety. A variety of data relates to multiple data types (numerical, categorical), data formats (audio files, video files), and data categories (demographics, location coordinates, graph data).

In addition, every data point contains numerous attributes or variables to provide context for the data. For example, every customer data of an eCommerce system may contain attributes, such as items purchased, customer demographics, items purchased, the quantity of purchase, and the clickstream. Therefore, deciding the most effective recommendation an eCommerce customer will respond to involves computing information and the other characteristic of the Big Data phenomenon, the high dimension.

Each attribute or variable is a dimension in the data space. For an instant, customer data has multiple attributes and can be viewed in multidimensional space. The addition of each dimension increases the complexity of data analysis techniques. As the dimension of the data increases, there is a need for an adaptable methodology that can work well with multiple data types and multiple attributes. Take an example of mining unstructured text content in a document. This text mining yields a data set where attributes range from a few hundred to hundreds of thousands of attributes because a document becomes a data point with each unique word as a dimension.

1.1.4.3 Complex Questions

The complexity of information that needs to be extracted increases as more complex data is available. To find the natural groupings in a data set with hundreds of dimensions, conventional analysis like hypothesis testing techniques is not scalable. A more automated approach, such as machine-learning algorithms, is needed to automate data exploration in the vast search space.

Conventional statistical analysis approaches a problem using a theoretical or stochastic model to predict a target variable based on input variables. Linear

regression and logistic regression analysis are classic examples of this technique where the model's parameters are estimated from the data. These techniques successfully modeled the relationships between target and input variables. Nevertheless, there is a need to extract information from big and complex data sets to explore and discover unknown data patterns, where the use of conventional statistical analysis techniques is finite.

Machine learning approaches the problem of modeling by attempting to find and select a model that can better characterize data or predict the output from input variables. The machine learning techniques are usually recursive and evaluate the output and "learn" from the modeled errors of previous steps in each cycle. This modeling route considerably aids in the exploratory data analysis since the approach here is not merely testing a hypothesis but developing many hypotheses for a given problem.

1.1.4.4. The Needed Paradigm

In today's challenge of analysing data with high volume and variety, it is essential to deploy an extensive methodology that can recursively utilise statistical and machine learning techniques to generate useful data patterns and relationships from large volumes and dimensions of data sets.

Data Mining is one such paradigm that can manage large data sets with many attributes and deploy complex modeling techniques, including machine learning and statistical analysis methods, to explore patterns from the data sets for complex questions.

Data analysts need to know what makes up a complete data mining process. In the next slide of "2. Data Mining as a Process", we will learn the core concept of Data Mining including the standard phases involved in the process.

1.2 The Concept of Data Mining

With no intention to attempt to cover all possible different views about data mining, our study focuses on the contexts surrounding this broad definition of Data Mining:

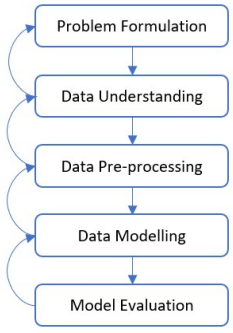
Data Mining is a process to discover patterns and relationships in data that involves various techniques from multiple disciplines, such as database systems, machine learning, information science, visualization, and statistical methods.
The word " process " is crucial in understanding Data Mining. As we have learned from the previous slides 1.1-1.5, Data Mining is not merely an application of statistical and machine learning methods. Instead, it is a thoughtfully planned and considered process of determining what will be most useful and actionable.

The general experimental process adapted to data-mining problems involves the following phases:

1. Problem Formulation
2. Data Understanding
3. Data Preparation (pre-processing)
4. Data Modelling
5. Model Evaluation

Figure 1 diagrammatically presents the general process flow of Data Mining. As shown in Figure 1, all phases, separately and the entire Data Mining process, are highly iterative.

Figure 1 A General Data Mining Process and Phases



All phases, separately and the entire data-mining process, are highly iterative, as shown in Figure 1. Therefore, a clear and adequate understanding of the whole process is vital for any successful data-mining application. For example, even though the machine learning or statistical techniques used in Phase 4 are powerful, the resulting model will not be valid and useful if data are wrongly used and pre-processed incorrectly or if the problem formulation is not precise or meaningful.

A data-mining process begins by developing a clear understanding of what problem(s) need to be solved in collaboration with experts in a particular application domain and data-mining (e.g., data analysts). This problem formulation phase includes interviewing domain experts using a (more or less formal) knowledge acquisition protocol. Subsequently, data analysts understand available data and prepare the data (evidence and hypotheses) for analysis. This data understanding phase involves inferring descriptive information in meta-data, basic data statistics (counts, ranges, distributions, visualizations), and data quality such as outliers, misclassification, duplication, and population imbalance. Based on the descriptive information, data are corrected if necessary and transformed for analysis. Then, appropriate data processing methods and applications are applied in the data pre-processing data phase to extract and enhance features for data modeling.

In the data modeling phase, appropriate machine learning or (and) statistical modeling technique(s) is selected (e.g., regression, rule-based system, decision tree, support vector machine, neural network). Then, several models may be built in this phase. The models built are then evaluated and optimized in light of the goals of the problem formulation, and hypotheses and plans are possibly adjusted. Finally, results are interpreted after determining the best model for deployment in an application.

Each phase involved in the Data Mining process will be further elaborated on in the following sections.

1.2.1 Problem Understanding and Formulation

Most data modeling studies are performed in a specific application domain. Consequently, domain-specific knowledge and experience are vital to developing an applicable and meaningful problem statement. Unfortunately, many application studies overlook the focus on the data-mining approach at the expense of a clear and precise problem statement.

In this phase, a data mining practitioner(s) usually identifies a set of attributes or variables for the unknown dependency and, if possible, a general form of this dependency as an initial inference as speculation to be tested or hypothesis. There may be several hypotheses formulated for a single problem at this phase.
This first phase of Data Mining requires the incorporated expertise of an application domain and a data-mining practitioner. This requirement implies a close interaction between the data-mining and the application experts in practice. However, the interaction and collaboration do not stop in this phase; it continues throughout the entire data-mining process.

A more detailed explanation of the problem formulation phase and the steps involved will be addressed in Week 2.

1.2.2 Data Understanding

Data understanding, also known as Exploratory Data Analysis (EDA) or data exploration, provides methods to understand the data. Basic understanding approaches involve computing descriptive statistics and visualization of data. These approaches can expose the data structure, the distribution of the data values, and the presence of extreme values, namely outliers, and highlight the associations within the data set.

Descriptive statistics like mean, median, mode, standard deviation, and range for each attribute or variable summarize the characteristics of the distribution of the

data. On the other hand, visual plots of data points instantly capture all the data points condensed into one chart.

In addition, this phase also involves understanding how data are generated and collected.

We will review more aspects of data understanding and the methods for exploring data in Weeks 2 and 3.

1.2.3 Data Preparation

Preparing the data sets to suit a data-mining process is the most time-consuming. This is because data rarely are available in the form required by the data modeling techniques. For example, most techniques would require data to be structured in a tabular format with records in rows and attributes (i.e., variables) in columns. If the data is in any other format, analysts need to transform the data by applying appropriate processing methods.

Furthermore, usually, data contain missing values, misclassification, duplication, or abnormal data distribution. These data quality problems require a thorough exploration of data to understand what data pre-processing techniques are appropriate for fixing the data. Besides, data analysts also need to pre-process the data into the types that suit the requirements of different data modeling techniques. For example, the Neural Network machine learning technique can only accept numerical data types.

A more detailed explanation of the data preparation phase and the techniques involved will be addressed in Week 3.

1.2.4 Data Modelling

A model is the abstract representation of the data and its relationships in a given data set. Hundreds of data modeling algorithms are developed today, derived from

statistics, machine learning, pattern recognition, and the computer science body of knowledge.

As data mining practitioners, it is essential to understand an algorithm before applying it. Specifically, the data mining practitioners must know how it works and decide what parameters need to be configured based on the problem and data understanding.

Data mining models built in the data modeling phase can be classified into the following categories: Predictive (a.k.a Supervised learning) and Descriptive (a.k.a. Unsupervised learning). Each category has many different algorithms with different approaches to solving the problem. Classification and estimation tasks are predictive techniques because they predict an outcome or target variable based on one or more input variables. Predictive modeling algorithms require a known prior data set to “learn” the model. Descriptive modeling techniques have no target variable to predict; hence there is no test data set. However, techniques in both categories have evaluation approaches.

This phase is not straightforward; usually, the implementation involves repetitive experiments with different parameters to generate several models, and selecting the best one is an additional task.

We will learn the techniques of predictive and descriptive modeling in Week 4 and Week 5 respectively.

1.2.5 Model Evaluation

In most cases, selected data-mining models should help in decision-making. Hence, such models need to be interpretable for actionable deployment because humans are unlikely to establish their decisions on complex "black-box" models. However, depending on the problem formulated, some data-mining projects do not require interpretable models but emphasize precision. For example, a model to predict the

classification of an image may not need to be interpretable according to the image properties.

Note that the model's precision and interpretation goals are somewhat contradictory. Usually, simple models are more interpretable, but they may be less accurate. Therefore, identifying accurate and useful models to select the best model for deployment is crucial. Current data-mining approaches are expected to yield highly accurate results using high-dimensional models. The problem of interpreting these models is considered a separate and critical task, with specific methods to validate the results. Interpreting the models is essential for non-data-mining practitioners because they are unlikely to understand and interpret hundreds of pages of numerical results and use them for successful decision-making.

We will learn the methods to validate data-ming models and select the best model in Week 6.

1.3 Applications of Data Mining

The following data-mining applications examples in sections 1.3.1 - 1.3.6 highlight the use of learning in performance situations. The stress is on performing well via supervised or unsupervised learning techniques to gain knowledge inferred from the data and make high-performance predictions. However, inference and predictions are not usually an application system that is actionable or workable as an application in its own right. Regardless, in the examples, the comprehensible decision structure is a critical element in the successful adoption of the application.

1.3.1 Web Mining

Mining information on the internet is a vast application area. First, search engine companies analyze the webpage hyperlinks to develop a measure to distinguish and rank each website and page. For example, Googe uses the PageRank metric to measure the characteristics of a web page. Also, search engines use PageRank to sort web pages into order before displaying a user's search results.

Another way search engines tackle the problem of ranking web pages is to use machine learning based on a set of example queries containing the terms in the query and human decisions (e.g., click-streams) about how relevant the web pages are to that query. Then a learning algorithm analyzes these examples as training data and develops a model to predict the relevance estimation for any web page and query. Finally, a set of feature values is counted for each web page according to the query term. For example, if it appears in the title tag or the web page's URL, how frequent it occurs on the web page, and how frequent it occurs in the anchor text of hyperlinks that direct and specify it. In addition, search engines mine the click-streams to improve the search next time.

Online product or service merchants mine the purchasing databases to develop recommendations. For example, the commonly found statement "users who bought this product also bought these" has a solid stimulus for giving users personalized choices. Similarly, movie sites recommend movies based on users' previous and others' choices. The merchants will benefit if the recommendations keep customers returning to their website.

Furthermore, there are social networks and other personal data. For example, today, people commonly share their thoughts and photographs on blogs and social life, music, and movie preference in tweets. These user data in social networks provide a massive volume of data for data-mining applications, for instance, applications for analyzing users' opinions, information sharing patterns, or even depression indications according to the words used in their social media platforms.

1.3.2 Loan Default

When a customer applies for a loan, he/she will need to fill in a questionnaire form for relevant financial and personal information. The loan company uses this information to decide whether to give a loan to the applicant. Such decisions usually involve two steps: first, applying learning methods to decide 'reject' and 'accept' cases. The methods estimate a numeric criterion based on the applicants'

information provided. Applicants are accepted if this criterion surpasses a preset threshold or rejected if it is lower than the threshold. The remaining borderline cases are more challenging and require human assessment for decision-making.

On analyzing historical data on whether applicants did repay their loans, it was possible that more or less half of the borderline applicants who were granted loans defaulted. An approach to identify the characteristics of those applicants who are likely to repay their loans is using Data Mining. A machine learning technique can be used in the data-mining process to generate a set of classification rules that predict the borderline cases if they are likely to default. These rules can improve the success rate of the loan decisions and are used to explain the characteristics of applicants who are likely to default or not.

1.3.3 Screening Images

Environmental scientists have been attempting to identify oil slicks from satellite images to provide early warning of ecological disasters and prevent illegal dumping. Also, analysing satellite images enables monitoring coastal waters anytime, regardless of weather conditions.

Oil slicks are dark regions in an image whose dimensions evolve according to the sea and weather conditions. However, other look-alike dark regions may be giving false alarms due to high wind. Detecting oil slicks is a costly manual process demanding highly trained personnel to evaluate each region in the image.

A hazard detection system can be deployed to screen images using a data-mining approach for subsequent manual processing by classifying images into positive or negative results (e.g., oil slicks or not) depending on application domains. For screen images, high dimensional data involves as many attributes are extracted from each geographical region and their vicinity background, distinguishing its area and other features such as shape, size, sharpness, intensity, and proximity to other regions.

The deployment can benefit many users, for instance, government agencies and organisations with different purposes and applications in different geographical areas.

1.3.4 Load Forecasting

It is crucial to decide and estimate future power demand as far in advance as possible in the electricity supply industry. With accurate estimations of the maximum and minimum load for each period (in terms of hour, day, month, year, and season), utility companies can produce notable economies in setting the operating reservation, maintenance schedule, and energy inventory management.

Automated load-forecasting aids have been functioning at a significant utility supplier for many years to develop hourly forecasts a few days in advance. However, the initial step of this automation approach is that data is usually collected manually over the previous few years to create a load model to realize the automation. Accordingly, the load model is static, constructed manually from historical data, and implicitly presumes typical conditions over the year. The last step was to consider weather conditions by finding the previous day most similar to the current cases and using the historical data from that day as a predictor.

The predictive modeling technique used in the data-mining approach is considered a supplementary correction to the static load model. The data collection includes temperature, humidity, wind speed, and cloud surface at local weather centers of the past year's historical records, enabling the construction of data sets for modeling and the difference between the actual and predicted loads by the static model.

The resulting system adopting a data-mining approach yielded a similar performance as trained human forecasters. However, it was faster by far, which only takes seconds instead of hours to produce a daily forecast.

1.3.5 Sales and Marketing

In sales and marketing applications, predictions are the principal interest. Banking industries were the early adopters of the data mining approach because of their successes in using machine learning techniques for credit assessment. Data mining reduces customer attrition by identifying changes in customer banking patterns that may hint at a change of bank or even preference changes. These changes may reveal, for example, a cluster of customers with an above-average attrition rate who perform most of their online banking when the service response is slow. In addition, data mining may determine new appropriate services for specific clusters. For instance, customers rarely get cash advances from their credit cards except in December, when they need to pay excessive interest rates for the holiday season.

Other domains, such as mobile phone companies, tackle churn by detecting behavior patterns that could benefit from new services and then advertise such services to retain their customer base. However, incentives to retain existing customers can be costly. A successful data mining project enables them to target customers who tend to obtain maximum benefit accurately.

Market basket analysis uses association techniques in a data mining approach to discover items that are likely to occur concurrently in transactions, such as supermarket transaction data containing items purchased information. For example, the data may include date/time of purchase, item quantity, and unit price. The data-mining approach uses the data that may discover that beer customers also buy chips. This discovery could be meaningful from the supermarket operator's perspective. Alternatively, it may come up with customers who usually purchase beer and diapers together on Fridays. This surprising result makes sense as young parents stock up for weekends at home. Supermarket operators could use such information for many purposes, such as planning store layouts and special discounts on specific items that tend to be purchased together.

Direct marketing is another favorite application domain for data mining. Bulk-mail promotional offers are pricey but have a highly profitable response rate. Anything that enables focus promotions, fulfilling a similar response from a smaller sample, is beneficial. Commercially available databases containing demographic information that distinguish individuals according to ZIP codes can be associated with information on existing customers to predict what kind of individuals tend to buy which items. This association can also predict likely future customers. Unlike shopping-mall retailers, direct mail companies have more complete customer purchasing histories and can use data mining to differentiate those who respond to special offers; besides saving money and reducing hassle, direct offers only to those likely to desire the product.

1.3.6 Other Applications

There are numerous other real-world, including scientific applications of Data Mining across different industry domains.

Advanced manufacturing processes often involve adjusting control parameters. British Petroleum used data mining to create rules for establishing the parameters. For example, it is challenging to separate crude oil from natural gas, which is vital to refining oil and controlling the process. However, establishing parameters using the data-mining generated rules only requires a few minutes, whereas human experts may take more than a day. Westinghouse faced difficulties manufacturing nuclear fuel pellets and used a data-mining approach to develop rules to control the process. The Tennessee printing company adopted a similar idea to control rotogravure printing presses to lower artifacts caused by improper parameter settings, decreasing 94% of the artifacts each year.

Customer support and service examples emerge when a customer reports a problem, and the company must determine and assign an appropriate technician to the job. Again, Bell Atlantic used a data-mining approach to make this decision according to the generated rules, saving more than $10 million per year by having fewer incorrect decisions.

In biomedicine, data mining approaches predict drug activity by analyzing drugs' chemical properties, including three-dimensional structures. This prediction accelerates drug discovery and decreases its cost. Data-mining approaches also help predict specific organic compounds structures from magnetic resonance spectra in chemistry.

Finally, cybersecurity is the primary concern in today's vulnerable networked computer systems; data-mining approaches allow the detection of intrusion by identifying unusual operation patterns.

1.4 Discussion Forum Activity

Discussion forum activity

Activity 1 (Research activity)

There are several standard methodologies for Data Mining as a process proposed by various stakeholders. Research the following methodologies, and examine the phases involved in each of the methodologies.

- CRISP-DM (Cross Industry Standard Process for Data Mining)
- KDD (Knowledge Discovery in Databases)
- SEMMA (Sample, Explore, Modify, Model, Assess)

Activity 2 (Discussion activity)

Discuss the similarities and differences between the methodologies researched in Activity 1 from the perspectives of their concept and phases involved in the data-mining process.

Time: 30 minutes

Purpose: The purpose of this task is to introduce students to several standard methodologies for Data Mining as a process proposed by various stakeholders and discuss the similarities and differences between the methodologies researched.

Practice Questions:

Question 1:

What are the phases involved in a data-mining process?

- Data Transformation
- Data understanding
- Data modeling
- Application development

Question 2:

Generating results from a database query using a large volume of data set is considered Data Mining.

- True
- False

Question 3

Which of the following statement(s) are precise?

- Data Mining enables an automated process of deriving data patterns.
- Data Mining is a process to discover interrelationships between data.
- Data Mining approaches can be used in predicting rules for sophisticated manufacturing processes controlling and setting parameters.
- Data-mining approaches always produce positive and useful results.