UNIT - 1

An Overview of Business Intelligence, Analytics, and Decision Support

1. Analytics to Manage a Vaccine Supply Chain Effectively and Safely

2. Changing Business Environments and Computerized Decision Support

3. Information Systems Support for Decision-Making

4. The Concept of Decision Support Systems(DSS)

5. Business Analytics Overview

6. Brief Introduction to Big Data Analytics

The business environment (climate) is constantly changing, and it is becoming more and more complex. Private and public organizations are under pressures that force them to respond quickly to changing conditions and be innovative in how they operate. Such activities require organizations to be agile and to make frequent and quick strategic, tactical, and operational decisions, some of which are very complex. Making such decisions may require considerable amounts of relevant data, information, and knowledge. Processing these, in the framework of the needed decisions, must be done quickly, frequently in real-time, and usually requires some computerized support.

1.1 Analytics to Manage a Vaccine Supply Chain Effectively and Safely

Magpie Sensing Employs Analytics to Manage a Vaccine Supply Chain Effectively and Safely

A Cold chain in healthcare is defined as a temperature-controlled supply chain involving a system of transporting and storing vaccines and pharmaceutical drugs.

It consists of three major components—

- Transport and storage equipment
- Trained personnel
- Efficient management procedures

The majority of the vaccines in the cold chain are typically maintained at a temperature of 35–46 degrees Fahrenheit [2–8 degrees Centigrade]. Maintaining cold chain integrity is extremely important for healthcare product manufacturers.

Especially for the vaccines, improper storage and handling practices that compromise vaccine viability prove a costly, time-consuming affair.

Vaccines must be stored properly from manufacture until they are available for use.

Any extreme temperatures of heat or cold will reduce vaccine potency; such vaccines, if administered, might not yield effective results or could cause adverse effects. Effectively maintaining the temperatures of storage units throughout the healthcare supply chain in real-time—i.e., beginning from the gathering of the resources, manufacturing, distribution, and dispensing of the products—is the most effective solution desired in the cold chain. Also, the location-tagged real-time environmental data about the storage units helps in monitoring the cold chain for spoiled products. The chain of custody can be easily identified to assign product liability.

A study conducted by the Centers for Disease Control and Prevention (CDC) looked at the handling of cold chain vaccines by 45 healthcare providers around the United States and reported that three-quarters of the providers experienced serious cold chain violations.

A Way Toward a Possible Solution Magpie Sensing, a start-up project under Ebers Smith and Douglas Associated LLC, provides a suite of cold chain monitoring and analysis technologies for the healthcare industry. It is a shippable, wireless temperature and humidity monitor that provides real-time, location-aware tracking of cold chain products during shipment. Magpie Sensing's solutions rely on rich analytics algorithms that leverage the data gathered from the monitoring devices to improve the efficiency of cold chain processes and predict cold storage problems before they occur.

Magpie sensing applies all three analytical techniques—

- Descriptive analytics

- Predictive analytics
- Prescriptive analytics

—to turn the raw data returned from the monitoring devices into actionable recommendations and warnings. The properties of the cold storage system, which include the set point of the storage system's thermostat, the typical range of temperature values in the storage system, and the duty cycle of the system's compressor, are monitored and reported in real-time.

This information helps trained personnel to ensure that the storage unit is properly configured to store a particular product. All the temperature information is displayed on a Web dashboard that shows a graph of the temperature inside the specific storage unit. Based on information derived from the monitoring devices, Magpie's predictive analytic algorithms can determine the set point of the storage unit's thermostat and alert the system's users if the system is incorrectly configured, depending upon the various types of products stored. This offers a solution to the users of consumer refrigerators where the thermostat is not temperature-graded.

Magpie's system also sends alerts about possible temperature violations based on the storage unit's average temperature and subsequent compressor cycle runs, which may drop the temperature below the freezing point. Magpie's further report possible human errors, such as failure to shut the storage unit doors or the presence of an incomplete seal, by analyzing the temperature trend and alerting users via the Web interface, text message, or audible alert before the temperature bounds are actually violated.--In a similar way, a compressor or a power failure can be detected; the estimated time before the storage unit reaches an unsafe temperature also is reported, which prepares the users to look for backup solutions such as using dry ice to restore power. In addition to predictive analytics, Magpie Sensing's analytics systems can provide prescriptive recommendations for improving the cold storage

processes and business decision making. Prescriptive analytics help users dial in the optimal temperature setting, which helps to achieve the right balance between freezing and spoilage risk; this, in turn, provides a cushion-time to react to the situation before the products spoil. Its prescriptive analytics also gather useful meta-information on cold storage units, including the times of day that are busiest and periods when the system's doors are opened, which can be used to provide additional design plans and institutional policies that ensure that the system is being properly maintained and not overused. Furthermore, prescriptive analytics can be used to guide equipment purchase decisions by constantly analyzing the performance of current storage units.

Based on the storage system's efficiency, decisions on distributing the products across available storage units can be made based on the product's sensitivity. Using Magpie Sensing's cold chain analytics, additional manufacturing time and expenditure can be eliminated by ensuring that product safety can be secured throughout the supply chain and effective products can be administered to the patients. Compliance with state and federal safety regulations can be better achieved through automatic data gathering and reporting about the products involved in the cold chain.

Finally, this opening vignette also suggests that innovative applications of analytics can create new business ventures. Identifying opportunities for applications of analytics and assisting with decision making in specific domains is an emerging entrepreneurial opportunity.

Changing Business Environments and Computerized Decision Support

It illustrates how a company can employ technologies to make sense of data and make better decisions. Companies are moving aggressively to computerized support of their operations. To understand why companies are embracing computerized support, including business intelligence, we developed a model called the Business Pressures–Responses–Support Model, which is shown in Figure 1.1.

The Business Pressures–Responses–Support Model :

The Business Pressures–Responses–Support Model, as its name indicates, has three components:

1. Business pressures that result from today's business climate,
2. Responses (actions taken) by companies to counter the pressures (or to take advantage of the opportunities available in the environment)
3. computerized support that facilitates the monitoring of the environment and enhances the response actions taken by organizations.

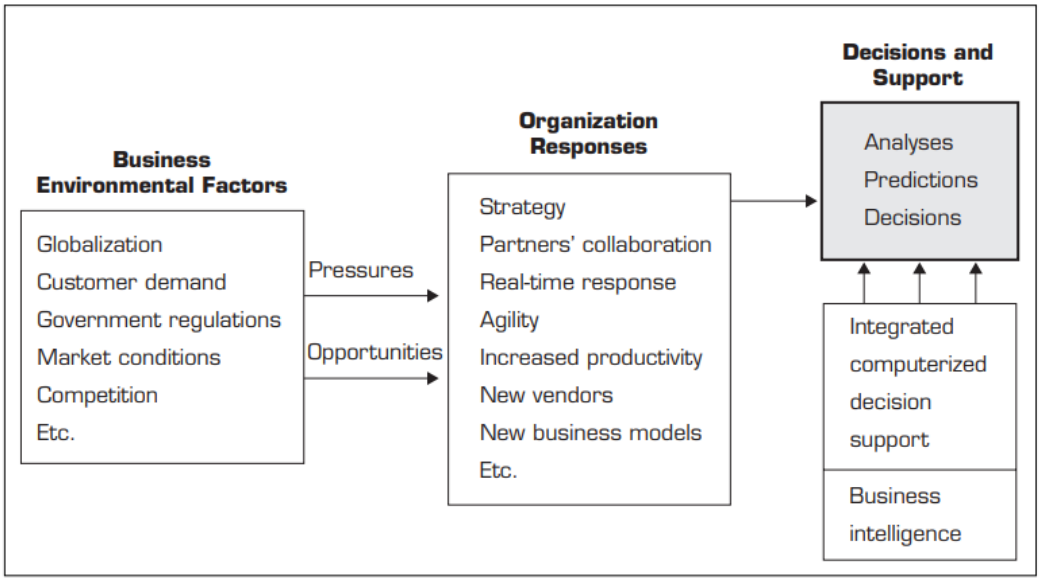


Figure 1.1 The Business Pressures–Responses–Support Model.

The Business Environment:

The environment in which organizations operate today is becoming more and more complex. This complexity creates opportunities on the one hand and problems on the other.

Take globalization as an example. Today, you can easily find suppliers and customers in many countries, which means you can buy cheaper materials and sell more of your products and services; great opportunities exist. However, globalization also means more and stronger competitors. Business environment factors can be divided into four major categories: markets, consumer demands, technology, and societal. These categories are summarized in Table 1.1.

Note that the intensity of most of these factors increases with time, leading to more pressures, more competition, and

so on. In addition, organizations and departments within organizations face decreased budgets and amplified pressures from top managers to increase performance and profit. In this kind of environment, managers must respond quickly, innovate, and be agile. Let's see how they do it.

Organizational Responses:

Be Reactive, Anticipative, Adaptive, and Proactive

Both private and public organizations are aware of today's business environment and pressures. They use different actions to counter the pressures. Vodafone New Zealand Ltd (Krivda, 2008), for example, turned to BI to improve communication and to support executives in its effort to retain existing customers and increase revenue from these customers. Managers may take other actions, including the following:

- Employ strategic planning.
- Use new and innovative business models.
- Restructure business processes.
- Participate in business alliances.
- Improve corporate information systems.
- Improve partnership relationships .

TABLE 1.1 Business Environment Factors That Create Pressures on Organizations

Factor	Description
Markets	Strong competition
	Expanding global markets
	Booming electronic markets on the Internet
	Innovative marketing methods
	Opportunities for outsourcing with IT support
	Need for real-time, on-demand transactions
Consumer demands	Desire for customization
	Desire for quality, diversity of products, and speed of delivery
	Customers getting powerful and less loyal
Technology	More innovations, new products, and new services
	Increasing obsolescence rate
	Increasing information overload
	Social networking, Web 2.0 and beyond
Societal	Growing government regulations and deregulation
	Workforce more diversified, older, and composed of more women
	Prime concerns of homeland security and terrorist attacks
	Necessity of Sarbanes-Oxley Act and other reporting-related legislation
	Increasing social responsibility of companies
	Greater emphasis on sustainability

- Encourage innovation and creativity.
- Improve customer service and relationships.
- Employ social media and mobile platforms for e- commerce and beyond.
- Move to make-to-order production and on-demand manufacturing and services.
- Use new IT to improve communication, data access (discovery of information), and collaboration.
- Respond quickly to competitors' actions (e.g., in pricing, promotions, new products and services).
- Automate many tasks of white-collar employees.

- Automate certain decision processes, especially those dealing with customers
. • Improve decision making by employing analytics.

Many, if not all, of these actions require some computerized support. These and other response actions are frequently facilitated by a computerized decision support System(DSS).

Closing the Strategy Gap:

One of the major objectives of computerized decision support is to facilitate closing the gap between the current performance of an organization and its desired performance, as expressed in its mission, objectives, and goals, and the strategy to achieve them. In order to understand why computerized support is needed and how it is provided, especially for decision-making support

Information Systems Support for Decision Making

From traditional uses in payroll and bookkeeping functions, computerized systems have penetrated complex managerial areas ranging from the design and management of automated factories to the application of analytical methods for the evaluation of proposed mergers and acquisitions. Nearly all executives know that information technology is vital to their business and extensively use information technologies.

Computer applications have moved from transaction processing and monitoring activities to problem analysis and solution applications, and much of the activity is done with Web-based technologies, in many cases

accessed through mobile devices. Analytics and BI tools such as data warehousing, data mining, online analytical processing (OLAP), dashboards, and the use of the Web for decision support are the cornerstones of today's modern management. Managers must have high-speed, networked information systems (wireline or wireless) to assist them with their most important task: making decisions. Besides the obvious growth in hardware, software, and network capacities, some developments have clearly contributed to facilitating the growth of decision support and analytics in a number of ways, including the following:

- **Group communication and collaboration:**

Many decisions are made today by groups whose members may be in different locations. Groups can collaborate and communicate readily by using Web-based tools as well as the ubiquitous smartphones. Collaboration is especially important along the supply chain, where partners—all the way from vendors to customers—must share information. Assembling a group of decision makers, especially experts, in one place can be costly. Information systems can improve the collaboration process of a group and enable its members to be at different locations (saving travel costs).

- I**mproved data management:**

Many decisions involve complex computations. Data for these can be stored in different databases anywhere in the organization and even possibly at Web sites outside the organization. The data may include text, sound, graphics, and video, and they can be in different languages. It may be necessary to transmit data quickly from distant locations. Systems today can search, store, and transmit needed data quickly, economically, securely, and transparently.

- **Managing giant data warehouses and Big Data:**

Large data warehouses, like the ones operated by Walmart, contain terabytes and even petabytes of data. Special methods, including parallel computing, are available to organize, search, and mine the data. The costs related to data warehousing are declining. Technologies that fall under the broad category of Big Data have enabled massive data coming from a variety of sources and in many different forms, which allows a very different view into organizational performance that was not possible in the past.

- **Analytical support:**

With more data and analysis technologies, more alternatives can be evaluated, forecasts can be improved, risk analysis can be performed quickly, and the views of experts (some of whom may be in remote locations) can be collected quickly and at a reduced cost. Expertise can even be derived directly from analytical systems. With such tools, decision-makers can perform complex simulations, check many possible scenarios, and assess diverse impacts quickly and economically. This, of course, is the focus of several chapters in the book.

- **Overcoming cognitive limits in processing and storing information:**

According to Simon (1977), the human mind has only a limited ability to process and store information. People sometimes find it difficult to recall and use information in an error-free fashion due to their cognitive limits. The term cognitive limits indicates that an individual's problem-solving capability is limited when a wide range of diverse information and knowledge is required. Computerized systems enable people to overcome their cognitive limits by quickly accessing and processing vast amounts of stored information.

- **Knowledge management.**

Organizations have gathered vast stores of information about their own operations, customers, internal procedures, employee interactions, and so forth through the unstructured and structured communications taking place among the various stakeholders. Knowledge management systems have become sources of formal and informal support for decision making to managers, although sometimes they may not even be called KMS.

- **Anywhere, any time support.**

Using wireless technology, managers can access information anytime and from any place, analyze and interpret it, and communicate with those involved. This perhaps is the biggest change that has occurred in the last few years. The speed at which information needs to be processed and converted into decisions has truly changed expectations for both consumers and businesses.

These and other capabilities have been driving the use of computerized decision support since the late 1960s, but especially since the mid-1990s. The growth of mobile technologies, social media platforms, and analytical tools has enabled a much higher level of information systems support for managers. In the next sections we study a historical classification of decision support tasks. This leads us to be introduced to decision support systems. We will then study an overview of technologies that have been broadly referred to as business intelligence. From there we will broaden our horizons to introduce various types of analytics.

The Concept of Decision Support Systems (DSS)

	Type of Control		
Type of Decision	**Operational Control**	**Managerial Control**	**Strategic Planning**
Structured	1 Accounts receivable Accounts payable Order entry	2 Budget analysis Short-term forecasting Personnel reports Make-or-buy	3 Financial management Investment portfolio Warehouse location Distribution systems
Semistructured	4 Production scheduling Inventory control	5 Credit evaluation Budget preparation Plant layout Project scheduling Reward system design Inventory categorization	6 Building a new plant Mergers & acquisitions New product planning Compensation planning Quality assurance HR policies Inventory planning
Unstructured	7 Buying software Approving loans Operating a help desk Selecting a cover for a magazine	8 Negotiating Recruiting an executive Buying hardware Lobbying	9 R & D planning New tech development Social responsibility planning

In the early 1970s, Scott-Morton first articulated the major concepts of DSS. He defined decision support systems (DSS) as "**interactive computer-based systems, which help decision makers utilize data**

and models to solve unstructured problems" (Gorry and Scott-Morton, 1971).

Components of DSS:

- DSS Database
- DSS Software System
- DSS User Interface

Types of DSS:

- Data driven:
- Model driven
- Knowledge driven
- Document driven
- Communication driven

The following is another classic DSS definition, provided by Keen and Scott-Morton (1978):

> Decision support systems couple the intellectual resources of individuals with the capabilities of the computer to improve the quality of decisions. It is a computer-based support system for management decision-makers who deal with semistructured problems.

Note that the term decision support system, like management information system (MIS) and other terms in the field of IT, is a content-free expression (i.e., it means different things to different people). Therefore, there is no universally accepted definition of DSS. Actually, DSS can be viewed as a conceptual methodology—that is, a broad, umbrella term. However, some view DSS as a narrower, specific decision support application.

DSS as an Umbrella Term:

The term DSS can be used as an umbrella term to describe any computerized system that supports decision making in an organization.

--An organization may have a knowledge management system to guide all its personnel in their problem solving.

–Another organization may have separate support systems for marketing, finance, and accounting;

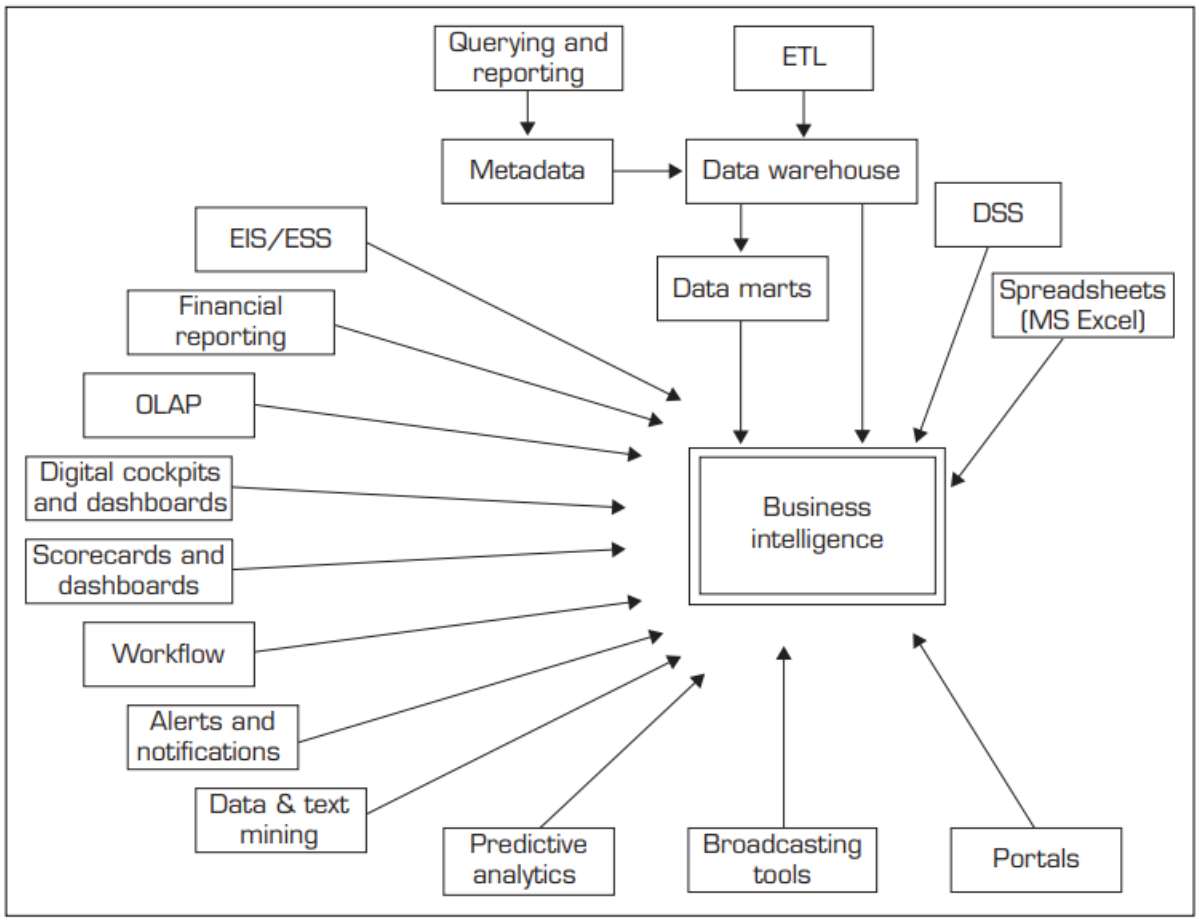
–A supply chain management (SCM) system for production;

–Several rule-based systems for product repair diagnostics and help desks. DSS encompasses them all.

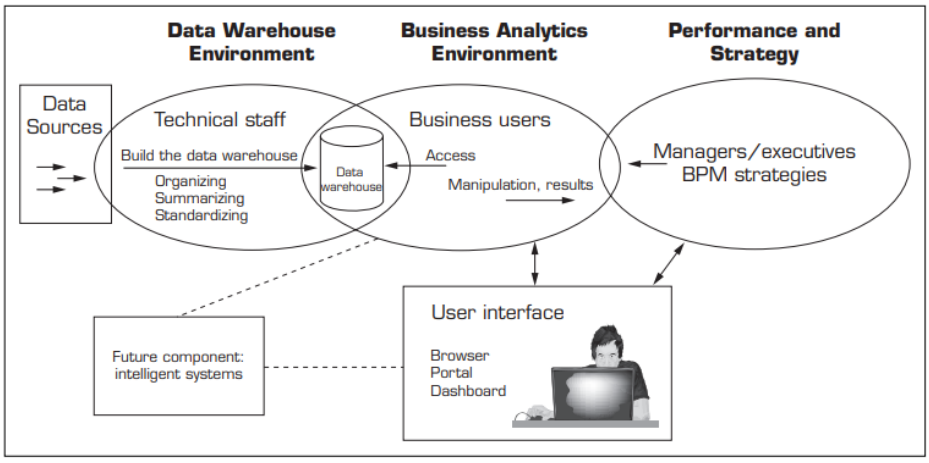
Evolution of DSS into Business Intelligence

In the early days of DSS, managers let their staff do some supportive analysis by using DSS tools. As PC technology advanced, a new generation of managers evolved—one that was comfortable with computing and knew that technology can directly help make intelligent business decisions faster. New tools such as OLAP, data warehousing, data mining, and intelligent systems, delivered via Web technology, added promised capabilities and easy access to tools, models, and data for computer-aided decision making. These tools started to appear under the names BI and business analytics in the mid-1990s.

Business intelligence:



Evolution of business intelligence



A High-Level Architecture of BI.

Business Analytics Overview

The word “analytics” has replaced the previous individual components of computerized decision support technologies that have been available under various labels in the past.

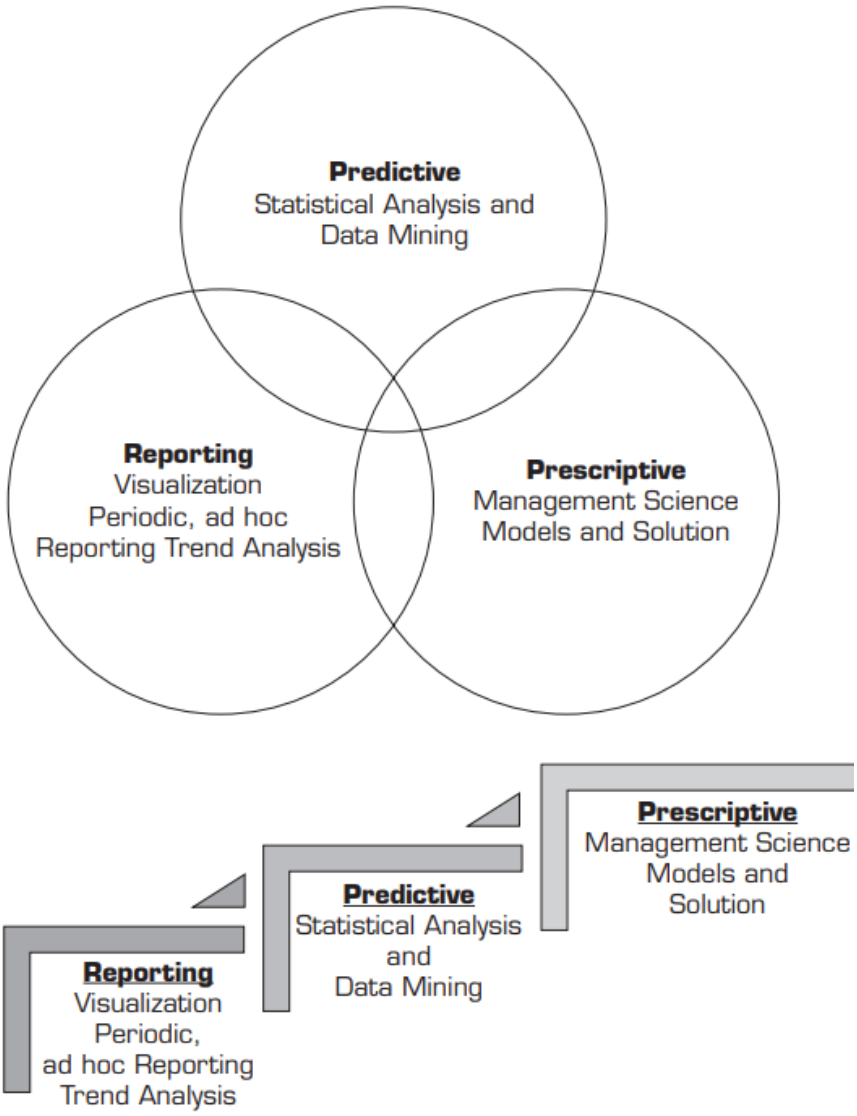
Indeed, many practitioners and academics now use the word analytics in place of BI. Although many authors and consultants have defined it slightly differently, one can view analytics as the process of developing actionable decisions or recommendations for actions based upon insights generated from historical data.

The Institute for Operations Research and Management Science (INFORMS) has created a major initiative to organize and promote analytics. According to INFORMS, analytics represents the combination of computer technology, management science techniques, and statistics to solve real problems. Of course, many other organizations have proposed their own interpretations and motivation for analytics. For example, SAS Institute Inc. proposed eight levels of analytics that begin with standardized reports from a computer system. These reports essentially provide a sense of what is happening with an organization. Additional technologies have enabled us to create more customized reports that can be generated on an ad hoc basis.

The next extension of reporting takes us to online analytical processing (OLAP)–type queries that allow a user to dig deeper and determine the specific source of concern or opportunities. Technologies available today can also automatically issue alerts for a decision maker when performance issues warrant such alerts. At a consumer level we see such alerts for weather or other issues. But similar alerts can also be generated in specific settings when sales fall above or below a certain level within a certain time period or when the inventory for a specific product is running low. All of these applications are made possible

through analysis and queries on data being collected by an organization. The next level of analysis might entail statistical analysis to better understand patterns.

These can then be taken a step further to develop forecasts or models for predicting how customers might respond toa specific marketing campaign or ongoing service/product offerings. When an organization has a good view of what is happening and what is likely to happen, it can also employ other techniques to make the best decisions under the circumstances.



Three Types of Analytics.

This idea of looking at all the data to understand what is happening, what will happen, and how to make the best of it has also been encapsulated by INFORMS in proposing three levels of analytics. These three levels are identified (informs.org/ Community/Analytics) as descriptive, predictive, and prescriptive. The above figure presents two graphical views of these three levels of analytics. One view suggests that these three are somewhat independent steps (of a ladder) and one type of analytics application leads to another. The interconnected circles view suggests that there is actually some overlap across these three types of analytics. In either case, the interconnected nature of different types of analytics applications is evident. We next introduce these three levels of analytics.

Descriptive Analytics

Descriptive or reporting analytics refers to knowing what is happening in the organization and understanding some underlying trends and causes of such occurrences. This involves, first of all, consolidation of data sources and availability of all relevant data in a form that enables appropriate reporting and analysis. Usually development of this data infrastructure is part of data warehouses, which we study in Chapter 3. From this data infrastructure we can develop appropriate reports, queries, alerts, and trends using various reporting tools and techniques.

A significant technology that has become a key player in this area is visualization. Using the latest visualization tools in the marketplace, we can now develop powerful insights into the operations of our organization. Color renderings of such applications are available on the companion Web site and also on Tableau's Web site.

Predictive Analytics

Predictive analytics aims to determine what is likely to happen in the future. This analysis is based on statistical techniques as well as other more recently developed techniques that fall under the general category of data mining. The goal of these techniques is to be able to predict if the customer is likely to switch to a competitor (“churn”), what the customer is likely to buy next and how much, what promotion a customer would respond to, or whether this customer is a creditworthy risk. A number of techniques are used in developing predictive analytical applications, including various classification algorithms.

For example, we can use classification techniques such as decision tree models and neural networks to predict how well a motion picture will do at the box office. We can also use clustering algorithms for segmenting customers into different clusters to be able to target specific promotions to them. Finally, we can use association mining techniques to estimate relationships between different purchasing behaviors. That is, if a customer buys one product, what else is the customer likely to purchase? Such analysis can assist a retailer in recommending or promoting related products.

Prescriptive Analytics

The third category of analytics is termed prescriptive analytics. The goal of prescriptive analytics is to recognize what is going on as well as the likely forecast and make decisions to achieve the best performance possible. This group of techniques has historically been studied under the umbrella of operations research or management sciences and has generally been aimed at optimizing the performance of a system. The goal here is to provide a decision or a recommendation for a specific action. These recommendations can be in the forms of a specific yes/no decision for a problem, a specific amount (say, price for a specific item or airfare to charge), or a complete set of production plans. The decisions

may be presented to a decision maker in a report or may directly be used in an automated decision rules system (e.g., in airline pricing systems). Thus, these types of analytics can also be termed decision or normative analytics.

Brief Introduction To Big Data Analytics

What Is Big Data?

Our brains work extremely quickly and are efficient and versatile in processing large amounts of all kinds of data: images, text, sounds, smells, and video. We process all different forms of data relatively easily. Computers, on the other hand, are still finding it hard to keep up with the pace at which data is generated—let alone analyze it quickly. We have the problem of Big Data.

Big data is data that cannot be stored in a single storage unit. Big Data typically refers to data that is arriving in many different forms, be they structured, unstructured, or in a stream.

Major sources of such data are clickstreams from Web sites, postings on social media sites such as Facebook, or data from traffic, sensors, or weather. A Web search engine like Google needs to search and index billions of Web pages in order to give you relevant search results in a fraction of a second. Although this is not done in real time, generating an index of all the Web pages on the Internet is not an easy task. Luckily for Google, it was able to solve this problem. Among other tools, it has employed Big Data analytical techniques.

There are two aspects to managing data on this scale: **storing and processing**. If we could purchase an extremely expensive storage solution to store all the data at one place on one unit, making this unit fault tolerant would involve major expense. An ingenious solution was

proposed that involved storing this data in chunks on different machines connected by a network, putting a copy or two of this chunk in different locations on the network, both logically and physically. It was originally used at Google (then called Google File System) and later developed and released as an Apache project as the **Hadoop Distributed File System** (HDFS). However, storing this data is only half the problem. Data is worthless if it does not provide business value, and for it to provide business value, it has to be analyzed. How are such vast amounts of data analyzed? Passing all computation to one powerful computer does not work; this scale would create a huge overhead on such a powerful computer. Another ingenious solution was proposed: Push computation to the data, instead of pushing data to a computing node. This was a new paradigm, and it gave rise to a whole new way of processing data. This is what we know today as the **MapReduce programming paradigm**, which made processing Big Data a reality. MapReduce was originally developed at Google, and a subsequent version was released by the Apache project called Hadoop MapReduce.

Today, when we talk about storing, processing, or analyzing Big Data, HDFS and MapReduce are involved at some level. Other relevant standards and software solutions have been proposed. Although the major toolkit is available as open source, several companies have been launched to provide training or specialized analytical hardware or software services in this space. Some examples are HortonWorks, Cloudera, and Teradata Aster. Over the past few years, what was called Big Data changed more and more as Big Data applications appeared. The need to process data coming in at a rapid rate added velocity to the equation.

One example of fast data processing is algorithmic trading. It is the use of electronic platforms based on algorithms for trading shares on the financial market, which operates in the order of microseconds. The need to process different kinds of data added variety to the equation. Another

example of the wide variety of data is sentiment analysis, which uses various forms of data from social media platforms and customer responses to gauge sentiments. Today Big Data is associated with almost any kind of large data that has the characteristics of volume, velocity, and variety.

Unit - 2

Text Analytics and Text Mining

- **Machine versus Men on Jeopardy!: The Story of Watson**
- **Text Analytics and Text Mining Concepts and Definitions**
- **Natural Language Processing**
- **Text Mining Applications**
- **Text Mining Process**
- **Text Mining Tools**

Machine versus Men on Jeopardy!: The Story of Watson

Can machine beat the best of man in what man is supposed to be the best at? Evidently, yes, and the machine's name is Watson. Watson is an extraordinary computer system (a novel combination of advanced hardware and software) designed to answer questions posed in natural human language. It was developed in 2010 by an IBM Research team as part of a DeepQA project and was named after IBM's first president, Thomas J. Watson.

Background

Roughly 3 years ago, IBM Research was looking for a major research challenge to rival the scientific and popular interest of Deep Blue, the computer chess-playing champion, which would also have clear relevance to IBM business interests. The goal was to advance computer science by exploring new ways for computer technology to affect science, business, and society. Accordingly, IBM Research undertook a challenge to build a computer system that could compete at the human champion level in real time on the American TV quiz show, Jeopardy! The extent of the challenge included fielding a real-time automatic contestant on the show, capable of listening, understanding, and responding—not merely a laboratory exercise.

Competing Against the Best

In 2011, as a test of its abilities, Watson competed on the quiz show Jeopardy!, which was the first ever human-versus-machine matchup for the show. In a two-game, combined-point match (broadcast in three Jeopardy! episodes during February 14–16), Watson beat Brad Rutter, the biggest all-time money winner on Jeopardy!, and Ken Jennings, the record holder for the longest championship streak (75 days). In these episodes, Watson consistently outperformed its human opponents on the game's signaling device, but had trouble responding to a few categories, notably those having short clues containing only a few words. Watson had access to 200 million pages of structured and unstructured content consuming four terabytes of disk storage. During the game Watson was not connected to the Internet.

Meeting the Jeopardy! Challenge required advancing and incorporating a variety of QA technologies (text mining and natural language processing) including parsing, question classification, question decomposition, automatic source acquisition and evaluation, entity and relation detection, logical form generation, and knowledge representation and reasoning. Winning at Jeopardy! required accurately computing confidence in your answers. The questions and content are ambiguous and noisy and none of the individual algorithms are perfect.

Therefore, each component must produce a confidence in its output, and individual component confidences must be combined to compute the overall confidence of the final answer. The final confidence is used to determine whether the computer system should risk choosing to answer at all. In Jeopardy! parlance, this confidence is used to determine whether the computer will "ring in" or "buzz in" for a question. The confidence must be computed during the time the question is read and before the opportunity to buzz in. This is roughly between 1 and 6 seconds with an average around 3 seconds.

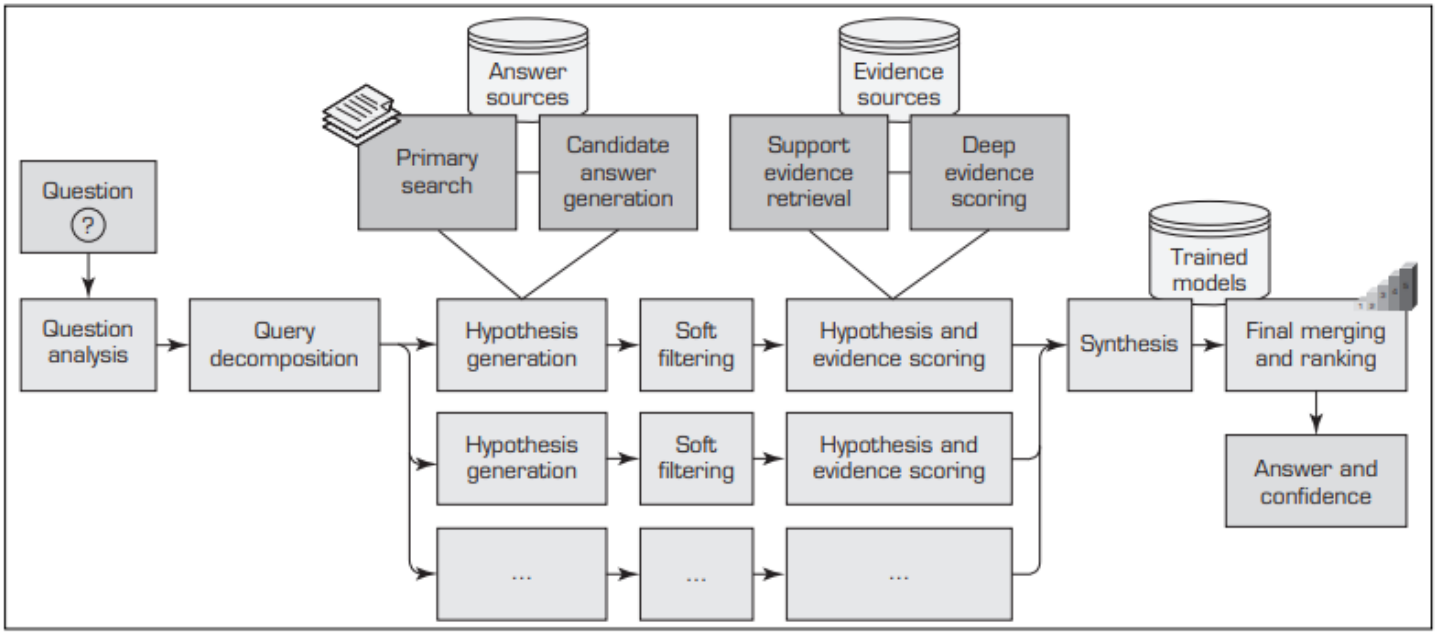
How Does Watson Do It?

The system behind Watson, which is called DeepQA, is a massively parallel, text mining– focused, probabilistic evidence-based computational architecture. For the Jeopardy! challenge, Watson used more than 100 different techniques for analyzing natural language, identifying sources, finding and generating hypotheses, finding and scoring evidence, and merging and ranking hypotheses. What is far more important than any particular technique that they used was how they combine them in DeepQA such that overlapping approaches can bring their strengths to bear and contribute to improvements in accuracy, confidence, and speed.

DeepQA is an architecture with an accompanying methodology, which is not specific to the Jeopardy! challenge. The overarching principles in DeepQA are massive parallelism, many experts, pervasive confidence estimation, and integration of the-latest-and-greatest in text analytics.

- **Massive parallelism:** Exploit massive parallelism in the consideration of multiple interpretations and hypotheses.
- **Many experts:** Facilitate the integration, application, and contextual evaluation of a wide range of loosely coupled probabilistic question and content analytics.

- **Pervasive confidence estimation:** No component commits to an answer; all components produce features and associated confidences, scoring different question and content interpretations. An underlying confidence-processing substrate learns how to stack and combine the scores.
- **Integrate shallow and deep knowledge:** Balance the use of strict semantics and shallow semantics, leveraging many loosely formed ontologies.



Conclusion:

The Jeopardy! challenge helped IBM address requirements that led to the design of the DeepQA architecture and the implementation of Watson. After 3 years of intense research and development by a core team of about 20 researchers, Watson is performing at human expert levels in terms of precision, confidence, and speed at the Jeopardy! quiz show.

IBM claims to have developed many computational and linguistic algorithms to address different kinds of issues and requirements in QA. Even though the internals of these algorithms are not known, it is imperative that they made the most out of text analytics and text mining. Now IBM is working on a version of Watson to take on surmountable problems in healthcare and medicine (Feldman et al., 2012).

Text Analytics and Text Mining Concepts and Definitions

Text analytics, is the process of extracting valuable insights, patterns, and information from unstructured textual data. It involves using computational techniques and algorithms to analyze and understand the content of text documents, enabling organizations to derive actionable intelligence from large volumes of text.

The primary objectives of text analytics include:

1. **Information Extraction:** Identifying and extracting specific pieces of information from text, such as names, dates, locations, and other relevant data.

2. **Sentiment Analysis:** Determining the emotional tone or sentiment expressed in text, which is particularly useful for gauging public opinion, customer sentiment, and brand perception.

3. **Text Classification:** Categorizing text documents into predefined categories or classes, such as spam detection, news categorization, and content tagging.

4. **Topic Modeling:** Discovering underlying topics or themes within a collection of documents, aiding in document organization and summarization.

5. **Text Summarization:** Generating concise and coherent summaries of longer texts, helping users quickly grasp the main ideas and key points.

6. **Pattern Recognition:** Identifying recurring patterns or trends in text data, which can be valuable for market research, competitive analysis, and trend forecasting.

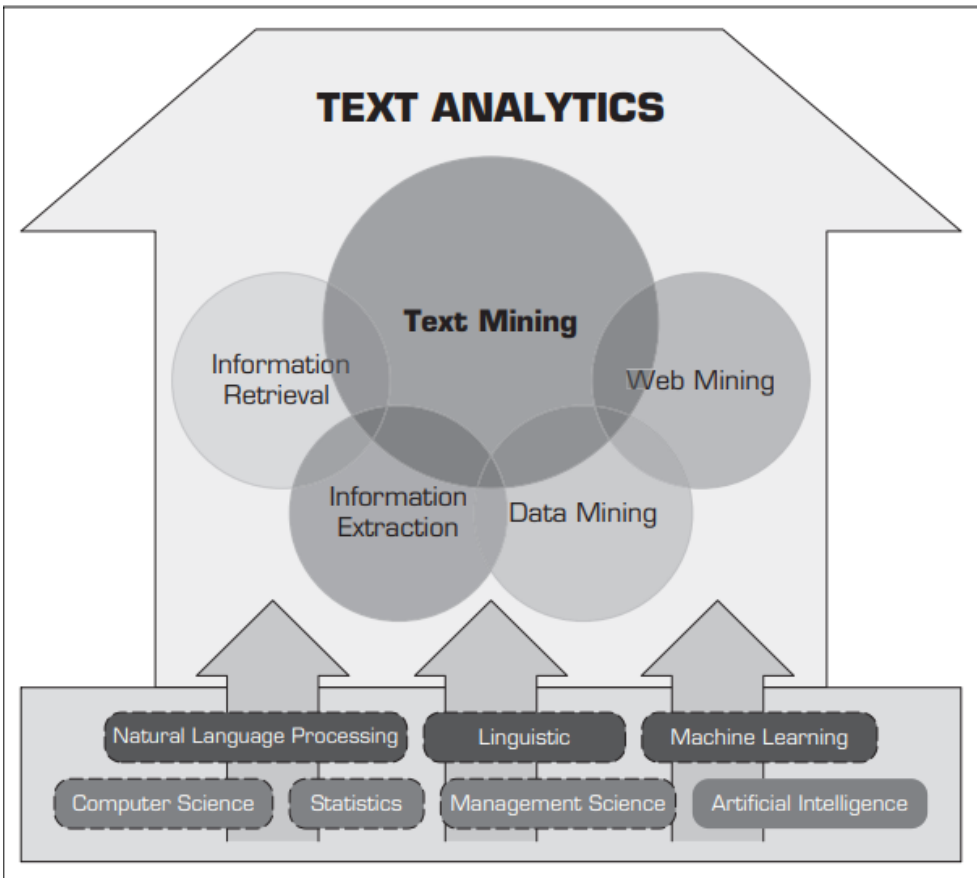
7. **Search and Retrieval:** Improving the accuracy and relevance of text-based search results, making it easier for users to find information within large datasets.

Text analytics leverages techniques such as natural language processing, machine learning, statistical analysis, and linguistic rules to process and analyze text data. It has a wide range of applications across industries, including marketing, customer service, finance, healthcare, and social media monitoring, among others. The insights gained from text analytics can inform decision-making, automate manual tasks, and enhance the understanding of textual content in various contexts.

Text Analytics = Information Retrieval + Information Extraction +Data Mining + Web Mining,

or simply

Text Analytics = Information Retrieval + Text Mining



Text Analytics, Related Application Areas, and Enabling Disciplines

Text analytics involves various concepts and techniques for analyzing and extracting valuable information from unstructured text data. Here are some key text analytics concepts:

1. **Text Preprocessing:** Before analysis, text data often undergoes preprocessing, which includes tasks like tokenization (breaking text into words or phrases), removing stop words, punctuation, and special characters, and stemming or lemmatization to reduce words to their base forms.

2. **Tokenization:** Tokenization is the process of splitting text into individual words or tokens. It is a fundamental step in text analysis.

3. **Stop Words:**Stop words are common words (e.g., "the," "and," "in") that are often removed from text because they carry little semantic meaning and can introduce noise in analysis.

4. **Stemming and Lemmatization:** These techniques reduce words to their base or root forms. Stemming removes prefixes and suffixes, while lemmatization maps words to their dictionary form.

5. **Bag of Words (BoW):** BoW is a simple text representation technique where a document is represented as a collection of its words, often used in text classification and clustering.

6. **Term Frequency-Inverse Document Frequency (TF-IDF):** TF-IDF is a numerical statistic that reflects the importance of a word in a document relative to a corpus. It's commonly used for text mining and information retrieval.

7. **N-grams**: N-grams are contiguous sequences of n words in a text, such as bigrams (two-word sequences) or trigrams (three-word sequences).

8. **Sentiment Analysis**:Sentiment analysis determines the emotional tone expressed in text, categorizing it as positive, negative, or neutral. It's used for gauging opinions and sentiment trends.

9. **Named Entity Recognition (NER):** NER is the process of identifying and classifying entities such as names of people, places, organizations, and dates in text.

10. **Topic Modeling:** Topic modeling techniques like Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) identify underlying topics or themes in a collection of documents.

11. **Text Classification:** Text classification involves assigning predefined categories or labels to documents based on their content, often used in spam detection, content tagging, and news categorization.

12. **Text Clustering:** Text clustering groups similar documents together based on their content, aiding in document organization and summarization.

13. **Text Summarization:** Text summarization techniques aim to generate concise and coherent summaries of longer texts, either extractively (selecting and combining sentences) or abstractively (generating new sentences).

14. **Word Embeddings:** Word embeddings are vector representations of words in a continuous space, capturing semantic relationships between words. Techniques like Word2Vec and GloVe are commonly used.

15. **Language Models:** Advanced language models like GPT-3 and BERT are capable of understanding and generating human-like text and are used in various NLP tasks, including chatbots and content generation.

16. **Text Analytics Tools:** Various software libraries and tools, such as NLTK, spaCy, scikit-learn, and deep learning frameworks, provide functionality for text analysis.

17. **Ethical Considerations:** Ethical considerations are crucial in text analytics, as it often involves handling sensitive data and addressing potential biases, privacy concerns, and fairness issues.

These concepts form the foundation of text analytics and are applied to analyze and extract insights from textual data in various domains and applications. The choice of techniques and tools depends on the specific goals and challenges of a text analytics project.

Text mining (also known as text data mining or knowledge discovery in textual databases) is the semi-automated process of extracting patterns (useful information and knowledge) from large amounts of unstructured data sources. Remember that data mining is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases, where the data are organized in records structured by categorical, ordinal, or continuous variables.

Text mining is the same as data mining in that it has the same purpose and uses the same processes, but with text mining the input to the process is a collection of unstructured (or less structured) data files such as Word documents, PDF files, text excerpts, XML files, and so on. In essence, text mining can be thought of as a process (with two main steps) that starts with imposing structure on the text-based data sources, followed by extracting relevant information and knowledge from this structured text-based data using data mining techniques and tools.

The benefits of text mining are obvious in the areas where very large amounts of textual data are being generated, such as law (court

orders), academic research (research articles), finance (quarterly reports), medicine (discharge summaries), biology (molecular interactions), technology (patent files), and marketing (customer comments). For example, the free-form text-based interactions with customers in the form of complaints (or praises) and warranty claims can be used to objectively identify product and service characteristics that are deemed to be less than perfect and can be used as input to better product development and service allocations. Likewise, market outreach programs and focus groups generate large amounts of data. By not restricting product or service feedback to a codified form, customers can present, in their own words, what they think about a company's products and services. Another area where the automated processing of unstructured text has had a lot of impact is in electronic communications and e-mail. Text mining not only can be used to classify and filter junk e-mail, but it can also be used to automatically prioritize e-mail based on importance level as well as generate automatic responses (Weng and Liu, 2004). The following are among the most popular application areas of text mining:

- **Information extraction**:Identification of key phrases and relationships within text by looking for predefined objects and sequences in text by way of pattern matching. Perhaps the most commonly used form of information extraction is named entity extraction. Named entity extraction includes named entity recognition (recognition of known entity names—for people and organizations, place names, temporal expressions, and certain types of numerical expressions, using existing knowledge of the domain), co-reference resolution (detection of co-reference and anaphoric links between text

entities), and relationship extraction (identification of relations between entities).

- **Topic tracking:** Based on a user profile and documents that a user views, text mining can predict other documents of interest to the user.
- **Summarization:** Summarizing a document to save time on the part of the reader.
- **Categorization:** Identifying the main themes of a document and then placing the document into a predefined set of categories based on those themes.
- **Clustering:** Grouping similar documents without having a predefined set of categories.
- **Concept linking:** Connects related documents by identifying their shared concepts and, by doing so, helps users find information that they perhaps would not have found using traditional search methods.
- **Question answering:** Finding the best answer to a given question through knowledge-driven pattern matching

The following list describes some commonly used text mining terms:

- **Unstructured data (versus structured data)**: Structured data has a predetermined format. It is usually organized into records with simple data values (categorical, ordinal, and continuous variables) and stored in databases. In contrast, **unstructured data** does not have a predetermined format and is stored in the form of textual documents. In essence, the structured data is for the computers to process while the unstructured data is for humans to process and understand.

- **Corpus.** In linguistics, a corpus (plural corpora) is a large and structured set of texts (now usually stored and processed electronically) prepared for the purpose of conducting knowledge discovery.
- **Terms**. A term is a single word or multiword phrase extracted directly from the corpus of a specific domain by means of natural language processing (NLP) methods.
- **Concepts.** Concepts are features generated from a collection of documents by means of manual, statistical, rule-based, or hybrid categorization methodology. Compared to terms, concepts are the result of higher level abstraction.
- **Stemming**. Stemming is the process of reducing inflected words to their stem (or base or root) form. For instance, stemmer, stemming, and stemmed are all based on the root stem.
- **Stop words.** Stop words (or noise words) are words that are filtered out prior to or after processing of natural language data (i.e., text). Even though there is no universally accepted list of stop words, most natural language processing tools use a list that includes articles (a, am, the, of, etc.), auxiliary verbs (is, are, was, were, etc.), and context-specific words that are deemed not to have differentiating value.
- **Synonyms and polysemes.** Synonyms are syntactically different words (i.e., spelled differently) with identical or at least similar meanings (e.g., movie, film, and motion picture). In contrast, polysemes, which are also called homonyms, are syntactically identical words (i.e., spelled exactly the same) with different meanings (e.g., bow can mean "to bend forward," "the front of the ship," "the weapon that shoots arrows," or "a kind of tied ribbon").

- **Tokenizing.** A token is a categorized block of text in a sentence. The block of text corresponding to the token is categorized according to the function it performs. This assignment of meaning to blocks of text is known as tokenizing. A token can look like anything; it just needs to be a useful part of the structured text.
- **Term dictionary**. A collection of terms specific to a narrow field that can be used to restrict the extracted terms within a corpus.
- **Word frequency.** The number of times a word is found in a specific document.
- **Part-of-speech tagging.** The process of marking up the words in a text as corresponding to a particular part of speech (such as nouns, verbs, adjectives, adverbs, etc.) based on a word's definition and the context in which it is used.
- **Morphology.** A branch of the field of linguistics and a part of natural language processing that studies the internal structure of words (patterns of word-formation within a language or across languages).
- **Term-by-document matrix** (occurrence matrix). A common representation schema of the frequency-based relationship between the terms and documents in tabular format where terms are listed in rows, documents are listed in columns, and the frequency between the terms and documents is listed in cells as integer values.
- **Singular-value decomposition (latent semantic indexing).** A dimensionality reduction method used to transform the term-by-document matrix to a manageable size by generating an intermediate representation of the frequencies using a matrix manipulation method similar to principal component analysis.

Natural Language Processing:

Natural language processing (NLP) is an important component of text mining and is a subfield of artificial intelligence and computational linguistics. It studies the problem of "understanding" the natural human language, with the view of converting depictions of human language (such as textual documents) into more formal representations (in the form of numeric and symbolic data) that are easier for computer programs to manipulate. The goal of NLP is to move beyond syntax-driven text manipulation (which is often called "word counting") to a true understanding and processing of natural language that considers grammatical and semantic constraints as well as the context.

The definition and scope of the word "understanding" is one of the major discussion topics in NLP. Considering that the natural human language is vague and that a true understanding of meaning requires extensive knowledge of a topic (beyond what is in the words, sentences, and paragraphs), will computers ever be able to understand natural language the same way and with the same accuracy that humans do? Probably not! NLP has come a long way from the days of simple word counting, but it has an even longer way to go to really understanding natural human language.

The following are just a few of the challenges commonly associated with the implementation of NLP:

•**Part-of-speech tagging.** It is difficult to mark up terms in a text as corresponding to a particular part of speech (such as nouns, verbs, adjectives, adverbs, etc.) because the part of speech depends not only on the definition of the term but also on the context within which it is used.

•**Text segmentation.** Some written languages, such as Chinese, Japanese, and Thai, do not have single-word boundaries. In these instances, the text-parsing task requires the identification of word boundaries, which is often a difficult task. Similar challenges in

speech segmentation emerge when analyzing spoken language, because sounds representing successive letters and words blend into each other.

•**Word sense disambiguation.** Many words have more than one meaning. Selecting the meaning that makes the most sense can only be accomplished by taking into account the context within which the word is used.

•**Syntactic ambiguity**. The grammar for natural languages is ambiguous; that is, multiple possible sentence structures often need to be considered. Choosing the most appropriate structure usually requires a fusion of semantic and contextual information.

•**Imperfect or irregular input**. Foreign or regional accents and vocal impediments in speech and typographical or grammatical errors in texts make the processing of the language an even more difficult task.

•**Speech acts.** A sentence can often be considered an action by the speaker. The sentence structure alone may not contain enough information to define this action. For example, "Can you pass the class?" requests a simple yes/no answer, whereas "Can you pass the salt?" is a request for a physical action to be performed

It is a longstanding dream of the artificial intelligence community to have algorithms that are capable of automatically reading and obtaining knowledge from text. By applying a learning algorithm to parsed text, researchers from Stanford University's NLP lab have developed methods that can automatically identify the concepts and relationships between those concepts in the text.

NLP has successfully been applied to a variety of domains for a variety of tasks via computer programs to automatically process natural human language that previously could only be done by humans. Following are among the most popular of these tasks:

•**Question answering.** The task of automatically answering a question posed in natural language; that is, producing a human-language answer when given a human-language question. To find the answer to a question, the computer program may use either a prestructured database or a collection of natural language documents (a text corpus such as the World Wide Web).

•**Automatic summarization.** The creation of a shortened version of a textual document by a computer program that contains the most important points of the original document.

•**Natural language generation.** Systems convert information from computer databases into readable human language.

•**Natural language understanding.** Systems convert samples of human language into more formal representations that are easier for computer programs to manipulate.

•**Machine translation.** The automatic translation of one human language to another.

•**Foreign language reading.** A computer program that assists a nonnative language speaker to read a foreign language with correct pronunciation and accents on different parts of the words.

•**Foreign language writing.** A computer program that assists a nonnative language user in writing in a foreign language.

•**Speech recognition.** Converts spoken words to machine-readable input. Given a sound clip of a person speaking, the system produces a text dictation.

•**Text-to-speech.** Also called speech synthesis, a computer program automatically converts normal language text into human speech.

•**Text proofing.** A computer program reads a proof copy of a text in order to detect and correct any errors.

•**Optical character recognition.** The automatic translation of images of handwritten, typewritten, or printed text (usually captured by a scanner) into machineeditable textual documents.

The success and popularity of text mining depend greatly on advancements in NLP in both generation as well as understanding of human languages. NLP enables the extraction of features from unstructured text so that a wide variety of data mining techniques can be used to extract knowledge (novel and useful patterns and relationships) from it. In that sense, simply put, text mining is a combination of NLP and data mining.

Text Mining Applications

As the amount of unstructured data collected by organizations increases, so does the value proposition and popularity of text mining tools. Many organizations are now realizing the importance of extracting knowledge from their document-based data

repositories through the use of text mining tools. Following are only a small subset of the exemplary application categories of text mining.

Marketing Applications:

Text mining can be used to increase cross-selling and up-selling by analyzing the unstructured data generated by call centers. Text generated by call center notes as well as transcriptions of voice conversations with customers can be analyzed by text mining algorithms to extract novel, actionable information about customers' perceptions toward a company's products and services. Additionally, blogs, user reviews of products at independent Web sites, and discussion board postings are a gold mine of customer sentiments. This rich collection of information, once properly analyzed, can be used to increase satisfaction and the overall lifetime value of the customer (Coussement and Van den Poel, 2008).

Text mining has become invaluable for customer relationship management. Companies can use text mining to analyze rich sets of unstructured text data, combined with the relevant structured data extracted from organizational databases, to predict customer perceptions and subsequent purchasing behavior. Coussement and Van den Poel (2009) successfully applied text mining to significantly improve the ability of a model to predict customer churn (i.e., customer attrition) so that those customers identified as most likely to leave a company are accurately identified for retention tactics.

Ghani et al. (2006) used text mining to develop a system capable of inferring implicit and explicit attributes of products to enhance retailers' ability to analyze product databases. Treating products as sets of attribute–value pairs rather than as atomic entities can potentially boost the effectiveness of many business applications, including demand forecasting, assortment optimization, product recommendations, assortment comparison across retailers and manufacturers, and product supplier selection. The proposed system allows a business to represent its products in terms of attributes and attribute values without much manual effort. The system learns these attributes by applying supervised and semi-supervised learning techniques to product descriptions found on retailers' Web sites.

Security Applications:

One of the largest and most prominent text mining applications in the security domain is probably the highly classified ECHELON surveillance system. As rumor has it, ECHELON is assumed to be capable of identifying the content of telephone calls, faxes, e-mails, and other types of data and intercepting information sent via satellites, public switched telephone networks, and microwave links.

In 2007, EUROPOL developed an integrated system capable of accessing, storing, and analyzing vast amounts of structured and unstructured data sources in order to track transnational organized crime. Called the Overall Analysis System for Intelligence Support (OASIS), this system aims to integrate the most advanced data and text mining technologies available in today's market. The system has enabled EUROPOL to make significant progress in supporting its law enforcement objectives at the international level (EUROPOL, 2007).

The U.S. Federal Bureau of Investigation (FBI) and the Central Intelligence Agency (CIA), under the direction of the Department for Homeland Security, are jointly developing a supercomputer data and text mining system. The system is expected to create a gigantic data warehouse along with a variety of data and text mining modules to meet the knowledge-discovery needs of federal, state, and local law enforcement agencies. Prior to this project, the FBI and CIA each had its own separate databases, with little or no interconnection.

Another security-related application of text mining is in the area of **deception detection**. Applying text mining to a large set of real-world criminal (person-of-interest) statements, Fuller et al. (2008) developed prediction models to differentiate deceptive statements from truthful ones. Using a rich set of cues extracted from the textual statements, the model predicted the holdout samples with 70 percent accuracy, which is believed to be a significant success considering that the cues are extracted only from textual statements (no verbal or visual cues are present). Furthermore, compared to other deception-detection techniques, such as polygraph, this method is nonintrusive and widely applicable to not only textual data, but also (potentially) to transcriptions of voice recordings.

Biomedical Applications:

Text mining holds great potential for the medical field in general and biomedicine in particular for several reasons. First, the published literature and publication outlets (especially with the advent of the open source journals) in the field are expanding at an exponential rate. Second, compared to most other fields, the medical literature is more standardized and orderly, making it a more "minable" information source. Finally, the terminology used in this literature is relatively constant, having a fairly standardized ontology. What follows are a few exemplary studies where text mining techniques were successfully used in extracting novel patterns from biomedical literature.

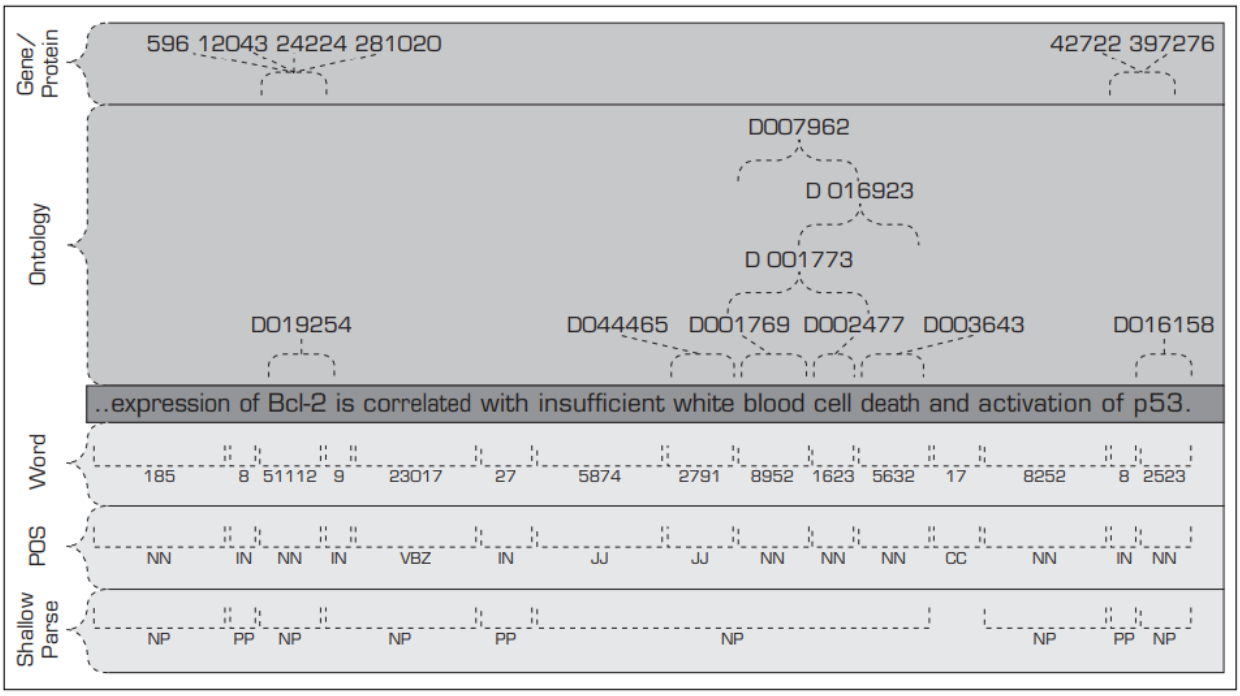
Experimental techniques such as DNA microarray analysis, serial analysis of gene expression (SAGE), and mass spectrometry proteomics, among others, are generating large amounts of data related to genes and proteins. As in any other experimental approach, it is necessary to analyze this vast amount of data in the context of previously known information about the biological entities under study. The literature is a particularly valuable source of information for experiment validation and interpretation. Therefore, the development of automated text mining tools to assist in such interpretation is one of the main challenges in current bioinformatics research.

Knowing the location of a protein within a cell can help to elucidate its role in biological processes and to determine its potential as a drug target. Numerous locationprediction systems are described in the literature; some focus on specific organisms, whereas others attempt to analyze a wide range of organisms. Shatkay et al. (2007) proposed a comprehensive system that uses several types of sequence- and text-based features to predict the location of proteins. The main novelty of their system lies in the way in which it selects its text sources and features and integrates them with sequencebased features. They tested the system on previously used data sets and on new data sets devised specifically to test its predictive power. The results showed that their system consistently beat previously reported results.

Chun et al. (2006) described a system that extracts disease–gene relationships from literature accessed via MedLine. They constructed a dictionary for disease and gene names from six public databases and extracted relation candidates by dictionary matching. Because dictionary matching produces a large number of false positives, they developed a method of machine learning–based named entity recognition (NER) to filter out false recognitions of disease/gene names. They found that the success of

relation extraction is heavily dependent on the performance of NER filtering and that the filtering improved the precision of relation extraction by 26.7 percent, at the cost of a small reduction in recall.

Figure below shows a simplified depiction of a multilevel text analysis process for discovering gene–protein relationships (or protein–protein interactions) in the biomedical literature (Nakov et al., 2005). As can be seen in this simplified example that uses a simple sentence from biomedical text, first (at the bottom three levels) the text is tokenized using part-of-speech tagging and shallow-parsing. The tokenized terms (words) are then matched (and interpreted) against the hierarchical representation of the domain ontology to derive the gene–protein relationship. Application of this method (and/or some variation of it) to the biomedical literature offers great potential to decode the complexities in the Human Genome Project.



Academic Applications

The issue of text mining is of great importance to publishers who hold large databases of information requiring indexing for better retrieval. This is particularly true in scientific disciplines, in which highly specific information is often contained within written text. Initiatives have been launched, such as Nature's proposal for an Open Text

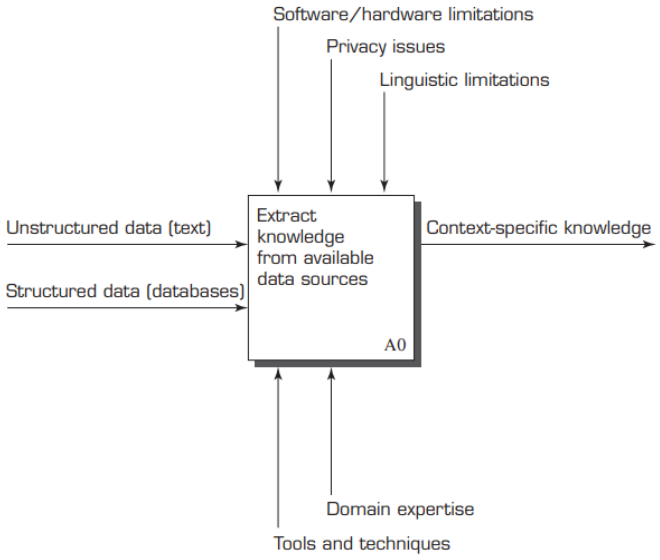
Mining Interface (OTMI) and the National Institutes of Health's common Journal Publishing Document Type Definition (DTD), which would provide semantic cues to machines to answer specific queries contained within text without removing publisher barriers to public access.

Academic institutions have also launched text mining initiatives. For example, the National Centre for Text Mining, a collaborative effort between the Universities of Manchester and Liverpool, provides customized tools, research facilities, and advice on text mining to the academic community. With an initial focus on text mining in the biological and biomedical sciences, research has since expanded into the social sciences. In the United States, the School of Information at the University of California, Berkeley, is developing a program called BioText to assist bioscience researchers in text mining and analysis.

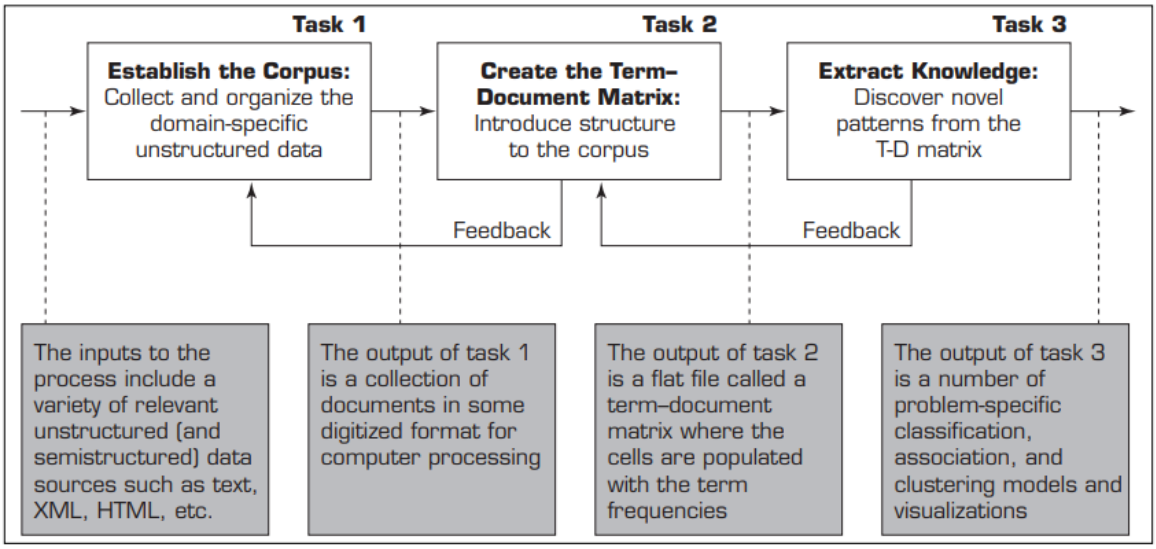
As described in this section, text mining has a wide variety of applications in a number of different disciplines.

Text Mining Process

In order to be successful, text mining studies should follow a sound methodology based on best practices. A standardized process model is needed similar to CRISP-DM, which is the industry standard for data mining projects. Even though most parts of CRISP-DM are also applicable to text mining projects, a specific process model for text mining would include much more elaborate data preprocessing activities. Depicts a high-level context diagram of a typical text mining process (Delen and Crossland, 2008). This context diagram presents the scope of the process, emphasizing its interfaces with the larger environment. In essence, it draws boundaries around the specific process to explicitly identify what is included in (and excluded from) the text mining process.



As the context diagram indicates, the input (inward connection to the left edge of the box) into the text-based knowledge-discovery process is the unstructured as well as structured data collected, stored, and made available to the process. The output (outward extension from the right edge of the box) of the process is the context-specific knowledge that can be used for decision making. The controls, also called the constraints (inward connection to the top edge of the box), of the process include software and hardware limitations, privacy issues, and the difficulties related to processing the text that is presented in the form of natural language. The mechanisms (inward connection to the bottom edge of the box) of the process include proper techniques, software tools, and domain expertise. The primary purpose of text mining (within the context of knowledge discovery) is to process unstructured (textual) data (along with structured data, if relevant to the problem being addressed and available) to extract meaningful and actionable patterns for better decision making.



Task 1: Establish the Corpus

The main purpose of the first task activity is to collect all of the documents related to the context (domain of interest) being studied. This collection may include textual documents, XML files, e-mails, Web pages, and short notes. In addition to the readily available textual data, voice recordings may also be transcribed using speech-recognition algorithms and made a part of the text collection.

Once collected, the text documents are transformed and organized in a manner such that they are all in the same representational form (e.g., ASCII text files) for computer processing. The organization of the documents can be as simple as a

collection of digitized text excerpts stored in a file folder or it can be a list of links to a collection of Web pages in a specific domain. Many commercially available text mining software tools could accept these as input and convert them into a flat file for processing. Alternatively, the flat file can be prepared outside the text mining software and then presented as the input to the text mining application.

Task 2: Create the Term–Document

Matrix In this task, the digitized and organized documents (the corpus) are used to create the term–document matrix (TDM). In the TDM, rows represent the documents and columns represent the terms. The relationships between the terms and documents are characterized by indices (i.e., a relational measure that can be as simple as the number of occurrences of the term in respective documents).

Terms / Documents	Investment Risk	Project Management	Software Engineering	Development	SAP	. . .
Document 1	1			1		
Document 2		1				
Document 3			3		1	
Document 4		1				
Document 5			2	1		
Document 6	1			1		
. . .						

Figure: A Simple Term–Document Matrix.

The goal is to convert the list of organized documents (the corpus) into a TDM where the cells are filled with the most appropriate indices. The assumption is that the essence

of a document can be represented with a list and frequency of the terms used in that document. However, are all terms important when characterizing documents? Obviously, the answer is "no." Some terms, such as articles, auxiliary verbs, and terms used in almost all of the documents in the corpus, have no differentiating power and therefore should be excluded from the indexing process. This list of terms, commonly called stop terms or stop words, is specific to the domain of study and should be identified by the domain experts. On the other hand, one might choose a set of predetermined terms under which the documents are to be indexed (this list of terms is conveniently called include terms or dictionary). Additionally, synonyms (pairs of terms that are to be treated the same) and specific phrases (e.g., "Eiffel Tower") can also be provided so that the index entries are more accurate.

Another filtration that should take place to accurately create the indices is stemming, which refers to the reduction of words to their roots so that, for example, different grammatical forms or declinations of a verb are identified and indexed as the same word. For example, stemming will ensure that modeling and modeled will be recognized as the word model.

The first generation of the TDM includes all of the unique terms identified in the corpus (as its columns), excluding the ones in the stop term list; all of the documents (as its rows); and the occurrence count of each term for each document (as its cell values). If, as is commonly the case, the corpus includes a rather large number of documents, then there is a very good chance that the TDM will have a very large number of terms. Processing such a large matrix might be time-consuming and, more importantly, might lead to extraction of inaccurate patterns.

Task 3: Extract the Knowledge

Using the well-structured TDM, and potentially augmented with other structured data elements, novel patterns are extracted in the context of the specific problem being addressed. The main categories of knowledge extraction methods are classification, clustering, association, and trend analysis. A short description of these methods follows.

Classification Arguably the most common knowledge-discovery topic in analyzing complex data sources is the classification (or categorization) of certain objects. The task is to classify a given data instance into a predetermined set of

categories (or classes). As it applies to the domain of text mining, the task is known as text categorization, where for a given set of categories (subjects, topics, or concepts) and a collection of text documents the goal is to find the correct topic (subject or concept) for each document using models developed with a training data set that includes both the documents and actual document categories. Today, automated text classification is applied in a variety of contexts, including automatic or semiautomatic (interactive) indexing of text, spam filtering, Web page categorization under hierarchical catalogs, automatic generation of metadata, detection of genre, and many others.

The two main approaches to text classification are knowledge engineering and machine learning (Feldman and Sanger, 2007). With the knowledge-engineering approach, an expert's knowledge about the categories is encoded into the system either declaratively or in the form of procedural classification rules. With the machine-learning approach, a general inductive process builds a classifier by learning from a set of reclassified examples. As the number of documents increases at an exponential rate and as knowledge experts become harder to come by, the popularity trend between the two is shifting toward the machine-learning approach.

Clustering

Clustering is an unsupervised process whereby objects are classified into "natural" groups called clusters. Compared to categorization, where a collection of preclassified training examples is used to develop a model based on the descriptive features of the classes in order to classify a new unlabeled example, in clustering the problem is to group an unlabelled collection of objects (e.g., documents, customer comments, Web pages) into meaningful clusters without any prior knowledge.

Clustering is useful in a wide range of applications, from document retrieval to enabling better Web content searches. In fact, one of the prominent applications of clustering is the analysis and navigation of very large text collections, such as Web pages. The basic underlying assumption is that relevant documents tend to be more similar to each other than to irrelevant ones. If this assumption holds, the clustering of documents based on the similarity of their content improves search effectiveness (Feldman and Sanger, 2007):

•**Improved search recall.** Clustering, because it is based on overall similarity as opposed to the presence of a single term, can improve the recall of a query-based

search in such a way that when a query matches a document its whole cluster is returned.

•**Improved search precision.** Clustering can also improve search precision. As the number of documents in a collection grows, it becomes difficult to browse through the list of matched documents. Clustering can help by grouping the documents into a number of much smaller groups of related documents, ordering them by relevance, and returning only the documents from the most relevant group (or groups).

The two most popular clustering methods are scatter/gather clustering and queryspecific clustering:

•**Scatter/gather.** This document browsing method uses clustering to enhance the efficiency of human browsing of documents when a specific search query cannot be formulated. In a sense, the method dynamically generates a table of contents for the collection and adapts and modifies it in response to the user selection.

•**Query-specific clustering.** This method employs a hierarchical clustering approach where the most relevant documents to the posed query appear in small tight clusters that are nested in larger clusters containing less similar documents, creating a spectrum of relevance levels among the documents. This method performs consistently well for document collections of realistically large sizes.

Association

A formal definition and detailed description of association was provided in the chapter on data mining (Chapter 5). Associations, or association rule learning in data mining, is a popular and well-researched technique for discovering interesting relationships among variables in large databases. The main idea in generating association rules (or solving market-basket problems) is to identify the frequent sets that go together.

In text mining, associations specifically refer to the direct relationships between concepts (terms) or sets of concepts. The concept set association rule A ⇒ B, relating two frequent concept sets A and C, can be quantified by the two basic measures of support and confidence. In this case, confidence is the percentage of documents that include all the concepts in C within the same subset of those documents that include all the concepts in A. Support is the percentage (or number) of documents that include all the concepts in A and C. For instance, in a document collection the concept “Software Implementation Failure” may appear most often in association with “Enterprise Resource Planning” and “Customer Relationship Management” with significant

support (4%) and confidence (55%), meaning that 4 percent of the documents had all three concepts represented together in the same document and of the documents that included "Software Implementation Failure," 55 percent of them also included "Enterprise Resource Planning" and "Customer Relationship Management."

Text mining with association rules was used to analyze published literature (news and academic articles posted on the Web) to chart the outbreak and progress of bird flu (Mahgoub et al., 2008). The idea was to automatically identify the association among the geographic areas, spreading across species, and countermeasures (treatments).

Trend Analysis Recent methods of trend analysis in text mining have been based on the notion that the various types of concept distributions are functions of document collections; that is, different collections lead to different concept distributions for the same set of concepts. It is therefore possible to compare two distributions that are otherwise identical except that they are from different subcollections. One notable direction of this type of analyses is having two collections from the same source (such as from the same set of academic journals) but from different points in time. Delen and Crossland (2008) applied trend analysis to a large number of academic articles (published in the three highest-rated academic journals) to identify the evolution of key concepts in the field of information systems.

Text Mining Tools:

As the value of text mining is being realized by more and more organizations, the number of software tools offered by software companies and nonprofits is also increasing. Following are some of the popular text mining tools, which we classify as commercial software tools and free (and/or open source) software tools.

Commercial Software Tools

The following are some of the most popular software tools used for text mining. Note that many companies offer demonstration versions of their products on their Web sites.

1. ClearForest offers text analysis and visualization tools.
2. IBM offers SPSS Modeler and data and text analytics toolkits.

3. Megaputer Text Analyst offers semantic analysis of free-form text, summarization, clustering, navigation, and natural language retrieval with search dynamic refocusing.

4. SAS Text Miner provides a rich suite of text processing and analysis tools.

5. KXEN Text Coder (KTC) offers a text analytics solution for automatically preparing and transforming unstructured text attributes into a structured representation for use in KXEN Analytic Framework.

6. The Statistica Text Mining engine provides easy-to-use text mining functionality with exceptional visualization capabilities.

7. VantagePoint provides a variety of interactive graphical views and analysis tools with powerful capabilities to discover knowledge from text databases.

8. The WordStat analysis module from Provalis Research analyzes textual information such as responses to open-ended questions, interviews, etc.

9. Clarabridge text mining software provides end-to-end solutions for customer experience professionals wishing to transform customer feedback for marketing, service, and product improvements.

Free Software Tools

Free software tools, some of which are open source, are available from a number of nonprofit organizations:

1. RapidMiner, one of the most popular free, open source software tools for data mining and text mining, is tailored with a graphically appealing, drag-and-drop user interface.

2. Open Calais is an open source toolkit for including semantic functionality within your blog, content management system, Web site, or application.

3. GATE is a leading open source toolkit for text mining. It has a free open source framework (or SDK) and graphical development environment.

4. LingPipe is a suite of Java libraries for the linguistic analysis of human language.

5. S-EM (Spy-EM) is a text classification system that learns from positive and unlabeled examples.

6. Vivisimo/Clusty is a Web search and text-clustering engine.

Often, innovative application of text mining comes from the collective use of several software tools.