CHAPTER-9

APPLICATIONS OF DATA MINING
CHAPTER-9

APPLICATIONS OF DATA MINING

9.1 APPLICATIONS OF DATA MINING

This section will be used to discuss some of the typical application of data mining organizations. The uses or applications of data mining will be grouped into main sections.

- **General Data Mining Applications**

  These applications are typically of value to any type of organization that perform the functions discussed here, and is not limited to only certain organizations operating in certain industries. This section highlights some of the applications of data mining in organizations in general typical business functions that these data mining applications fall under are:

  (i) **E-commerce and E-business Functions**

  Data mining can help organizations to identify the information that would be most suitable to put on the Web by conducting an analysis of competitors and potential customers, and determining the expertise of the organization (Thuraisingham, 2003:276) [23] In order for any organization to get onto the World Wide Web, it needs to decide what information to put on its web site. Data mining can also be used to assist organizations waning to conduct e-business with other organizations by providing information with regard to selecting the best partners, identifying competitors and determining the best pricing policies for its products of services. Summaries the e-commerce and e-business applications of data mining as being the following:
• Selecting partner organization.
• Analyzing customer profiles.
• Determining which products to market online.
• Assessing the similarity of user browsing patterns.
• Identifying web pages that are viewed together.
• Assessment of the similarity of web page contents.
• Categorization of web pages based on content.

The web site will assist online users in their decision making by listing for instance other books that have been bought by people who also bought the particular book that the online user is currently looking at. One of the most useful applications of data mining in an e-commerce environment can be found on the Amazon.com web site. The web site also makes use of data mining to bundle books that are often bought in pairs, together and to sell or market these at reduced price if they are bought together. Amazon.com also uses data mining to profile its customers, and online users that return to the web site after having bought a book for instance, will be provided with lists of other book titles that they might be interested in based on the category of book that they just bought.

(ii) Business Intelligence Functions

In order to be able to run a business effectively, management needs intelligence relating to competitors, customer, partners and employees, as well as intelligence relating to market conditions, future trends,
government policies and more lists the applications of data mining in business intelligence as follows:

- Analyzing competitor strategies.
- Analyzing customer profiles.
- Determining your won business strategies.

(iii) Customer Relationship Management Functions

This information is then used to serve the customers in the best possible way. The CRM is a key application of business intelligence and deals with mining information about customers from public as well as private databases in order to build customer profiles. This mined information is then used to provide customers with easy access to other similar products that they might be interested in. Amazon.com once again provides a good example of how this mined information can be put to good use by gathering information such as, customers who bought this book also bought those books. Typical application of data mining in customer relationship management is therefore:

- Building and analyzing customer profiles.
- Developing customer-specific products.

(iv) Marketing and Sales Functions

Marketing and sales were two of the early business functions that drove the development of the data mining. The following are some of the data mining application relevant to these functions:

- Performing targeted marketing.
• Determining the marketing strategies of competitors.
• Prediction of sales trends.
• Market segmentation.
• Lifestyle behavior analysis.
• Online sales support.
• Analyzing or predicting customer reaction to promotions.
• Marketing basket analysis.

(v) Enterprise Resource Management Function

The resources of an organization might include human, inventory, expertise or anything that is of value to the organization, and data mining has many applications in this area. Enterprise resource management involves the management of all the resources of an organization. Some of this application are:

• Analyzing human resources for employee retention benefits and payroll.
• Analyzing enterprise resources or supply chain management.

One of the primary applications of data mining in enterprise resource management is to ensure that the organization’s resources and capabilities are aligned with what is required in the business environment not only today, but also in the future.
(vi) **Manufacturing and Planning Functions**

Data mining has several applications in the manufacturing and planning functions of organizations. Some of these applications include the following:

- Analyzing and developing manufacturing schedules.
- Determining the best possible assembly routines.
- Monitoring plants to detect anomalies and unusual patterns.
- Assisting large manufacturing firms to predict production breakdowns and analyze product defects.

The main advantage data mining offers to manufacturing firms is its ability to assist in identifying the conditions that lead to critical situations. Engineers can then use this information to correct problems and prevent future failures.

(vii) **Education and Training Functions**

These applications related mainly to the identification of both current and future training needs. Data mining has several applications that are relevant to education and training. The following two main applications:

- Developing courses and schedules based on the needs identified by data mining.
- Predicting future trends in education.

- **Industry Specified Data Mining Application**

Data mining application and uses, and it related to specific industries will be discussed in this section. Due to the huge number of industries
affected by data mining, a brief outline of the data mining application in only some industries will be provided.

(i) **Telecommunication Industry**

The telecommunication industry offers an attractive domain for data mining applications due to the data intensive nature of applications employed in this industry. The following as some of the major applications of data mining in the telecommunications industry:

- Determining travel routes and giving advice to wireless users during peak hour travel times.
- Data mining can be used to assist network operators with network intrusion detection and fault monitoring.
- Identifying and understanding the critical issues that determine client loyalty.
- Narrowing the focus and increasing the effectiveness of marketing campaigns.
- Forecasting network behavior.
- Call tracking.
- Churn management
- Fraud detection.

(ii) **Physical Sciences, Social Sciences and Engineering Industries.**

Several disciplines are covered any these three industries and each of those disciplines would have to be analysed in order to identify all the potential and current applications of data mining relevant to each of them. These industries however falls beyond the scope of this study, and
the data mining applications relevant to them will be limited to the following as:

In astronomy, data mining can be used to sift through the vast amounts of data gathered in order to help with the identification of new planets and starts.

- Chemical engineering can benefit from data mining by making use of it to assist in identifying new chemical elements.
- Social scientists could make use of data mining to analyze social behavior.

(iii) Medical and Biotechnical Industries.

Data mining has been applied extensively in these two industries over the years. Typical applications of data mining in these industries include the following:

- Using data mining to help physician with the diagnosis and prevention of illnesses, for instance identifying specific groups of people that are more susceptible to developing cancer, and focusing preventative efforts on those people.
- Analysing genome data in an effort to identify disease causing genes such as those responsible for Alzheimer’s.
- Pharmaceutical companies are also making use of data mining to discover new drugs, and to predict future trends and needs in order to optimize their manufacturing processes.
(iv) Law Enforcement Industry

Defines this probably one of the most critical areas for the application of data mining today. Data mining tools can be used to detect unusual patterns, suspicious behavior and unauthorized intrusions, all of which can be extremely helpful not only in aiding companies and government in their law enforcement efforts, but also in their counter-terrorism efforts. The following includes some of the specific applications of data mining in counter-terrorism:

- Intrusion detection.
- Insider threat analysis.
- Identifying terrorists.

(v) Financial Services Industry

The major benefit offered by data mining to the financial services industry related to the industry’s dominant problem category of trying to predict the future.

The financial services industry can benefit a great deal from data mining, especially with regard to the major role that Business Intelligence and Customer Relationship Management play in this industry. Typical data mining applications in the financial services industry include the following:

- Analyzing stock market quotes.
- Providing clients with customized investment advice.
- Forecasting bankruptcies.
- Forecasting defaulting loans.
- Credit assessment.
- Portfolio management.
- Loan approval.
- Risk classification
- Identifying suspicious transactions.
- Risk management

**(vi) Insurance Industry**

During the past few years, increased competition has forced insurance companies to move away from being traditionally product-centric organizations to becoming customer-centric organizations. Customers needs and adapt their business process in the value chain to effectively meet those needs. The insurance industry is totally dependent on the ability to convert raw data into intelligence about customers, markets, competitors and the business environment.

**(vii) Customer Relationship Management**

They have to utilize the vast amounts of data that they collect and store in order to assists them in gaining a better understanding of the needs of their customers. Typical insurance companies have huge customer bases, varied product lines, May distribution channels and market spreads across geographies. The following applications of data mining that can assist them in doing this:
**Customer profitability**

It is important for organizations to retain and maximize the profitability of their existing customers, rather than simply acquiring new ones. The insight gained from customer profitability analysis can also lead to product customization and new product development.

**Customer Lifetime Value.**

Customer should not only be measured by their current profitability, but the potential future income that could be generated from them should also be assessed and considered. Less profitable customers might also serve as excellent references the at result in the acquisition of more profitable customers.

Data mining tools can assist in modeling this by taking a large number of factors relating to the customer into consideration. Some of the more general marketing and sales based data mining applications, such as customers segmentation, attrition analysis, market basket analysis, target marketing, campaign analysis and cross selling can also be used with great effect to improve an insurance company’s customer relationship management efforts.

**(viii) Channel Management**

The internet threatened to change this, but not many insurance companies have had success at selling their products online. Insurance companies have traditionally made use of independent agents, brokers and direct sales to distribute their products effectively. The focus of data mining in channel management has been more towards gaining insights into the various distribution channels, and helping organizations to
address issues relating to channel management more effectively. Some of these issues that can be addressed by data mining applications are:

- Agent and Sales Force Deployment.
- Agent Development and Relationship Management.
- Channel Analysis.

(ix) Underwriting and Policy Management

Underwriting involves deciding whether the risk taken to insure a client is acceptable or not, and data mining can be used in the following applications to optimize this function of the insurance value chain:

- Premium Analysis
- Loss Analysis

Premium income is the primary source of income for an insurance company, and its optimization will therefore have a significant impact on the overall performance of the company. Similarly, limiting the losses incurred can also help improve the overall performance of insurance companies.

(x) Claims Management

The efficient handling of claims can contribute significantly towards an organization’s overall customer relationship management success, but at the same time, care should be taken not to neglect this function and to allow the increase of fraudulent claim submission and payouts to take place. Claims management is probably one of the most important functions in the insurance value chain. Data mining applications relevant to the claims management function include the following:
• Claims Analysis.
• Fraud Detection.
• Claims Estimation.

The key to gaining a competitive advantage in the insurance industry is found in recognizing that customer databases, if properly managed, analysed and exploited are unique, valuable corporate assets. Insurance companies can use data mining to unlock the intelligence contained in their customer databases.

9.2 Electronic Commerce

Business transactions conducted via electronic means; most often referring to Internet-based relationships between customers and vendors, but also including CD-ROM catalogs; also called internet marketing. The Internet, direct marketing's newest channel, has provided an unprecedented opportunity on a global basis for businesses to interact with, reach out to, and be accessible by their customers without limitations with respect to physical location or time zone. E-commerce practitioners utilize the Internet to disseminate company or product information, generate leads, take orders, and build customer databases. CD-ROM catalogs permit the distribution of vast amounts of personalized product information via a cost efficient medium. Business-to-business sales are expected to dominate ecommerce by 2003 reaching $1.3 trillion, whereas consumer sales are projected to be $108 billion. Electronic commerce may be adopted more readily by buyers of services and products such as software, information, and photos, because these items can be both sold and distributed over the Internet.
E-Commerce is short for:

- Electronic Commerce
- E Commerce Nerd Making Sense of Making Money Online
- Electronic commerce

Electronic commerce, commonly known as e-commerce or ecommerce, consists of the buying and selling of products or services over electronic systems such as the Internet and other computer networks. The amount of trade conducted electronically has grown dramatically since the wide introduction of the Internet. A wide variety of commerce is conducted in this way; including things such as electronic funds transfer, supply chain management, internet marketing, online transaction processing, and electronic data interchange (EDI), automated inventory management systems, and automated data collection systems. Modern electronic commerce typically uses the World Wide Web at least at some point in the transaction's lifecycle, although it can encompass a wider range of technologies such as e-mail as well.

A small percentage of electronic commerce is conducted entirely electronically for "virtual" items such as access to premium content on a website, but most electronic commerce eventually involves physical items and their transportation in at least some way.

E-commerce or electronic commerce is generally considered to be the sales aspect of e-business.

E-commerce, it has three subcategories: business-to-business (B2B), business-to-consumer (B2C), and consumer-to-consumer (C2C).
9.3 E-Commerce Trends

"No single force embodies our electronic transformation more than the evolving medium known as the Internet. Internet technology is having a profound effect on the global trade in services," according to a White House paper in 1997. Electronic commerce is estimated to have been in the range of $63 billion in 1999 and is expected to soar to $1,444 trillion by 2003 (For rester Research, 1999). Electronic commerce is a broad term describing business activities with associated technical data that are conducted electronically. It is an entire set of different, digitally enabled activities that are progressively replacing the more traditional brick-and-mortar commercial functions. While the wider phenomenon of "electronization of economic activities" encompasses the digitalization of all processes of economic wealth generation-including economic analysis, production, storage, information provisioning, marketing, and so on-it is the area of sales and related processes facilitated by electronic media that have been popularly termed "electronic commerce." Consequently within the more general phenomenon of digitalization of modern life, we find its most important component electronic commerce.

9.4 New Ways of Doing Business

Corporate, not-for-profit, and governmental systems are incorporating many increasingly digitalized processes that are leading to astounding productivity gains in the world economy because these processes are becoming less expensive, less time consuming, and more useful ("Why the Productivity Revolution," 2000). For example, a directory-assistance call formerly required an operator, look-up in paper-based directories, and a localized search. Now it involves a national (or international)
computer database, voice synthesis, and automatic connection. Furthermore, the process has been expanded; one can do reverse searches through the Internet that will point to the owner of a telephone number, link this to one's telephone bill, and not involve any individual as the service provider. Thousands of "system processes" are undergoing this type of mutation, leading to cheaper, less time-consuming, and expanded types of services. Figure 1 shows several components of the business process (e.g., marketing) and electronic commerce tools (e.g., Web banners) that are structurally changing ways of doing business.

The marketing, advertising, and care triad are the core of the phenomenon. One-to-one marketing (whereby large customer databases link much information about clients, thus creating very efficient leads) is linked to very tailored advertising:

The firm knows the client when he or she is connected to the Internet, and it fires off a series of individually targeted banners catering very closely to the client’s needs. These advertising banners can explore the geography of the client at that moment (for example, if in a car, the closest gas station, drug store, or sports bar), linkages among products or recent purchase (bought a computer, needs parts and software), personal factors (getting married, needing a dress, birthday, death in the family), and other factors. The e-care part of the triad is the emerging process of the new organization. Technologically rich products need superior, technologically based support. E-care—a mix of e-mail, Web based support, and, when essential, phone support—is cheaper and more powerful if properly done than the traditional means. Organizations are finding that the same stringent standards of traditional care must be applied in the e-organization.
The electronic commerce revolution is in its initial phases and will progressively take over all processes either directly or indirectly. The distinction between "snail" commerce and e-commerce will disappear, with all processes becoming either digital or aided by digital supporting processes. The pace of this transformation is what differentiates winning from losing competitors, industries, and investors. The intrinsic nature of the product and processes, as well as the dynamics and resistance to change of corporations and industries, will determine the pace of change and the gains in productivity. Together with the telephone, railroads, and electricity, the Internet is one of the major agents of change of modern life.

Two major factors affect the speed of change in terms of product: (1) bitability and (2) e-commoditization. A product is bitable or not. If it can be transmitted over the Internet, the product (or service) is bitable. Software, information, remote support services, banking, brokerage, and insurance, for example, fall into this category. If a product is bitable it does not ultimately have to be physically delivered, although it may take some time to acquire sufficient bandwidth or get consumers used to the idea. The e-commoditization factor is more complex. An commodity is an item that one does not need to touch, see, Tryon, tryout, taste, or squeeze before buying. Clearly high-fashion clothes, cars, foods, meats, vegetables, and girlfriends tend not to fall into this category. On the other hand, this is very much a question of attitude and need. Busy executives will forgo the examination of food items for the convenience of having them at home when they arrive there. Once a teenager has tried on an item of brand-name clothing for size, it becomes a commodity, since sizes tend to be quality controlled. A buyer who lives in a very remote location may consider an item to be a commodity because the cost of its
examination does not warrant extensive travel-and, in the case of clothes, they can be altered. Ultimately, bitable goods and bitable commodity goods present the highest potentials for e-commerce.

Research predicts that there will be a wide range of business expansion on the Internet, which incorporates a series of predictions from four different organizations.

Travel and apparel are expected to be the largest B2C (business to consumer) areas. At the same time, the volumes of B2B (business to business) trade are expected to be six to ten times larger than B2C, but with much narrower margins.

**9.5 Emerging Principles of Electronic Commerce**

An entirely new set of principles of commerce is emerging. First is the realization that a Malthusian physical world is giving way to a place where information is abundant and eyeballs limited. Second is the realization that paradoxes exist because of technology, and that giving things away for free, not protecting software against privacy, and paying for visitors, may be the paradigms of the e-world. Third, the meaning of the words competitor and industry are changing. In the faceless world of the Internet, one's current and future customers and suppliers are both one's competitors and one's allies. Fourth, industries are blending and changing, and affiliation agreements allow for the creation of entire product cycles without the ownership of inventory or production facilities. Finally, pricing models are changing; hybrids of fixed pricing, auctions, variable pricing, contingent pricing, and name-your-price pricing are emerging and creating new business models. While technology gets most of the credit, actually successes are usually based
on the triad of: (1) technology, (2) business model innovation, and (3) a family of facilitating (profitable) services.

The B2B sector of e-commerce will present both vertical and horizontal models. In the vertical model, the firm will focus on an industry and develop great industry expertise in order to develop its markets. In the horizontal model the firm will focus on one type of product or service and offer it across industries (e.g., Internet payroll services). The B2B sector is intrinsically different from B2C. Buyers are well informed, possess many resources and can negotiate based on volume. Brand name is much less of a consideration than price, quality, delivery time, and reliability. Three different models have emerged for B2B transactions: (1) the e-catalog model for situations in which there are many different items at distributed locations and the price is fixed (e.g., auto parts); (2) the auction model, in which products are not standardized and there are great differences in the perceptions of value (e.g., auctions of used capital plant products); and (3) the commodity auction model, in which there are not too many variations on the type of product and there are large buyers and sellers (e.g., natural gas, pork bellies, coffee, etc.).

Electronic commerce is progressively and irreversibly changing the face of many businesses because of three dominant phenomena: (1) disintermediation, whereby one party to a transaction is eliminated (e.g., brokers in on-line trading); (2) reintermediation, whereby a new electronic intermediary comes between the seller and the buyer (e.g., electronic booksellers that take orders and farm them out to providers that have the book in stock); and (3) cannibalization, whereby businesses progressively give up their traditional brick-and-mortar ventures for the
superior electronic model (e.g., traditional pharmacies opening on-line drug stores).

9.6 Problems in Search of Solutions

The e-commerce juggernaut is not without its dangers and shortcomings. It is drastically affecting traditional firms that cannot continue to do business according to the traditional economic model. For example, a stock brokerage firm that on average charged $90 per trade cannot compete with $10 on-line trades; if any adaptations meet with strong resistance, the organization could be dooming itself to extinction. The security weaknesses of the e-commerce infrastructure have been well publicized: Viruses, security intrusions, and inability to provide services because of volume attacks are not phenomena that will disappear. A continuous struggle is evolving among facilitating technologies, intrinsic technological dangers, and the management of these factors.

Privacy issues present a different set of challenges. The same technology that facilitates business activities and provides wonderful services is also a major threat to individual freedom. Large databases linking information about economic activity from different sources (purchases, banking activity, medical records) provide great economic advantages by making marketing more efficient, loans more targeted, and medical information ubiquitous. They also create great dangers for privacy potentials for abuse. Doubleclick.com, a marketing analysis technology firm, tracks customer activities in Web sites. Its click-path analysis allows firms to understand customer behavior and improve their offerings. However they have 11,000 clients, and linking together the buyers' profiles from these sites is considered to be far too intrusive by many people. Complaints
have been filed with the Federal Trade Commission and boycotts proposed. As a response, Double click lured bigger profile "primary" officers for its board." Solutions, however, are not as straightforward as they might seem. Making certain actions illegal can actually create arbitrage opportunities and extraordinary advantages for Internet players. Because gambling rules are strict in the United States, electronic casinos are created in cooperative havens in the Bahamas: A legal obstacle in the United States is a business opportunity for another country. When restrictions are placed on the use and content of databases in Germany, offshore database havens appear immediately. When Web-site censorship appears in China, free-Chinese Web sites are constructed. Telephone systems are monitored and taxed by PTTs, supranational satellite telephone networks are being created.

Consequently, easy fixes are not possible and new methods of establishing order, efficiency, and decency will have to be created. Because the Internet is truly a supranational entity, nations need to band together to maintain order and efficiency and reasonableness in cyberspace. The same economic factors that allow for arbitrage can also be used for self-policing and monitoring of the ecommerce environment. To benefit fully from this medium, companies/entities and nations must have payment-clearing solutions, customs solutions, and access to the large markets of the economy. Rogue countries can be excluded from the payment-clearing chains; rogue companies behaving in unacceptable ways can be boycotted and excluded from any affiliation and linking deals. Self-policing seals such as the American Institute of Certified Public Accountants' (AI CPA) Web Trust and SysTrust products, inspections, and certificates can be used for monitoring and supervision. International information structures, involving many cooperating
organizations, can be on the alert for rogue behavior and spearhead a drive to create reasonable and unbiased rules. Technology can be used to monitor and detect money laundering, illegal product flows, and information trafficking. But such monitoring must be carefully conceived and supervised, because it could turn into a "Big Brother" type of behavior. Most important of all is not to succumb to the easy temptation of creating restrictive and ill-thought-out laws of the sort that legislators tend to create when some local scandal occurs.

9.7 The Impact of E-Commerce

There can be no denial that e-commerce has made a definite and significant impact upon the global economy. Its integration into society has affected the ways people manage and conduct business. The spread of e-commerce will continue as organizations use it to increase productivity as well as another avenue for sales and service. In fact, Clarke (1993) predicts that business-to-business (B2B) and business-to-consumer (B2C) e-commerce will become so popular that most businesses will be forced to enter the digital economy in order to retain competitive advantage.

In a study encompassing the first half of 2000, the Internet Economy was, in the United States, found to support more than 3 million workers (CREC 2001). Online businesses numbered 550,000 by mid-2000 (Cerpa and Jamieson 2002), up 30 percent from the previous year. The United States Department of Commerce estimated that retail e-commerce sales for the fourth quarter of 2001 totaled $10 billion (Pastore 2002), up from $5.3 billion in the same period in 1999 (Armstrong 2000). In contrast,
total retail sales were $821.2 billion and $860.8 billion in the fourth quarter of 1999 and 2001 respectively. Although ecommerce only accounts for a miniscule portion of all retail sales, e-commerce sales have doubled proportionate to total retail sales in the two year period, reflecting an increasing amount of e-commerce usage. That ecommerce sales only compose about one percent of total retail sales demonstrates there is plenty of room for e-commerce to continue expanding into.

From a worldwide perspective, IDC found that e-commerce spending grew 68 percent between 2000 and 2001 to reach $600 billion. IDC estimates that this will continue increasing to a massive $1 trillion in 2002 (Pastore 2002). The numbers above are primarily in reference to B2C transactions. It is postulated that B2B transactions outstrip B2C ones with the Gartner Group predicting 2004 worldwide B2B revenues at $7.3 trillion. It is this profit potential that has lured venture capitalists into investing into 'dot com' companies which are trying to 'ride the wave' and establish themselves as profitable businesses.

There are many other statistics that may be cited. However, one thing is clear that ecommerce's prominence in business is increasing. In the next few years, this growth is forecasted to continue unabated.

From an organizational and management perspective, the changes e-commerce has wrought have been just as dramatic. Most notably, the restructuring of the 'Big Five' multinational accounting firms to separate their e-commerce consulting arms from their auditing arms. The impetus for this is to ensure that their audit work is not compromised as a conflict of interest exists if a firm both consults and audits the same client (Kane 2002). Accenture's separation from Andersen, as a result, also gave it independence such that when Andersen was shaken by the
collapse of Enron and consequential legal proceedings, Accenture was relatively untouched. PricewaterhouseCoopers has spun off its consulting arm which was acquired by IBM, Deloitte Touche Tohmatsu spun off its consulting arm into Braxton, with KPMG likewise turning theirs into Bearing Point.

Apart from sales and marketing, e-commerce systems are also employed for operational and supply purposes, including finance, logistics and procurement. Incumbent firms especially have managed to take advantage of these types of systems, enabling cost reduction and greater process efficiencies (Turban 2000).

9.8 Advantages of E-Commerce

E-commerce possesses a variety of attributes that have made it attractive to businesses and to their customers. Implementing e-commerce systems has enabled organizations and their customers to conduct business in ways previously not possible.

A company conducting e-commerce via the Internet has, potentially, a global reach, due to the interconnected, ubiquitous nature of it. This has greatly lowered the barriers of entry for many industries, especially those conducive to electronically-based ventures, as opposed to more traditional and well established markets (Dertouzos 2000). Previously, penetrating into a market, especially a geographically based one, required organizations to have a physical presence and physical assets. E-commerce allows businesses to have market presence without physical presence, and thus less capital is required. Systems are also available 24 hours, 7 days a week, therefore are not limited by time zone constraints (Vacca 1995). As a result, many ‘dot com’ companies such as Dell have
been able to successfully enter markets dominated by traditional incumbent giants like Hewlett-Packard and IBM. Additionally, intangible assets such as technological innovation and strong customer relationships are increasingly important in a business’ strategic planning (Anandarajah and Lek 2000).

The ability to leverage new technology and quickly establish strong brand equity has allowed Amazon.com to prosper against incumbent Barnes and Noble. Even though Barnes and Noble entered the e-commerce marketplace shortly afterwards, Amazon.com had created enough brand loyalty such that it eventually returned a profit.

In a Forrester Research survey (Martin 2002), over 50 percent of B2C respondents indicated that penetrating new markets is the reason they are involved in e-commerce, followed by the ability to deliver a better quality of service. For B2B firms, over 30 percent cited lower operating costs as their motivation, followed by over 25 percent citing better information delivery. Therefore, e-commerce offers more than just new market opportunities for business. Performing transactions electronically decreases the cost of creating, process, distributing, storing and retrieving paper-based information. For example, the implementation of an electronic procurement system may reduce a company’s administration costs by up to 85 percent (Turban 2000). E-commerce may incite business process re-engineering, as e-commerce systems change the way business processes interact. Supply chain management and procurement systems necessarily remodel and optimize supply chain processes (Turban 2000).

The open nature of e-commerce allows businesses to select from a greater variety of business partners - the opportunity for businesses to
build network-based ad hoc partnerships (Wang et al. 2000). The formation of strategic alliances with other firms, and achieving knowledge sharing through the integration of their systems, allows firms to diversify, leveraging on each others' expertise. The result is businesses being able to adapt and to respond to shifting market forces. For example, Amazon.com's alliance with Drugstore.com allowed Amazon to diversify into the pharmaceuticals market, while giving Drugstore access to Amazon's large customer base.

The advantages are not restricted to businesses alone, though. For e-commerce to work, incentives for customers must exist as well. Greater information accessibility is enabled via e-commerce. Information that may have taken days to obtain is now available to customers immediately from a single computer terminal, enhancing the level of customer service. Furthermore, because the information is exchanged in electronic format, customers are able to integrate transaction processing into their own e-commerce systems, thereby automating the procurement process and reducing costs.

EDI systems have allowed this in the past, although XML-based information is increasingly being used as a standard format for information exchange over the Internet (Yeomans 2001). The advantages of paperless transactions are therefore applicable to customers.

Customers also get more choice in terms of vendors. Compared to traditional retailing, each electronic vendor is just as accessible as another, providing customers with greater convenience. Quick price comparisons between a large number of competing vendors is also available. Online marketplaces allow customers to place requests for
tenders, and businesses to bid for them. Finally, quicker delivery of products is possible, especially with digitilized products (Turban 2000).

9.9 E-Commerce Risks and Weaknesses

Despite the prodigious growth of e-commerce, the new environment is not without its weaknesses. From these weaknesses arise risks that must be noted and accounted for, lest they are exploited by the unscrupulous or cause other unintended disruption. Ecommerce is relatively new, and that, combined with the swiftness of its uptake, has meant that there are few formalized methodologies and guidelines in place for developing those systems (Ng and Wong 1999). Systems, as a result, have weaknesses that may be exploited. As businesses and customers become increasingly reliant on their e-commerce systems, they should be aware of the risks they have increased exposure to. Weaver et al. (2000) discuss four different possible trends in the future growth of ecommerce. Two of the possible trends include a decline in e-commerce sales, both due to flaws in ecommerce (one which the Internet recovers from, and one which is irrecoverable from such as the breaking of commonly used encryption standards).

Consequently, it is important to identify these flaws lest e-commerce suffers from one of them. E-commerce must be regarded as a safe and practical way to do business, by both businesses and customers, to succeed. The feasibility of ecommerce is well proven, although it is clear that many are not convinced that ecommerce is secure. A majority of the insecurities associated with e-commerce fall under the branch of 'electronic crime' (ACPR 2001), which is crime perpetrated via information systems. This encompasses both traditional 'real world'
crimes that have migrated online (such as credit card fraud), as well as new crimes that have arisen alongside the development of new information systems (such as viruses).

Insecure systems are a recurring theme when e-commerce flaws are discussed in literature. Weaver et al. (2000, p. 30) say that "at least two major research areas will affect the growth-or non growth-of Internet businesses over the next three to five years: wireless technology and security." Udo's (2001) survey concludes that security forms a major barrier for the spread e-commerce with regards to consumers: "Security concern is one of the main reasons Web users give for not purchasing over the Web."

(Udo 2001, p. 166) An empirical study by Elliot and Fowell (2000) of consumer experiences with e-commerce retailing discovered that 50 percent of transactions rated as unsatisfactory stemmed from security concerns. Other factors forming barriers are privacy issues, censorship concerns, e-mail safety concerns and impersonation/forged identity concerns (which are security related).

For businesses, security is a major issue. The case of the 'love bug' virus, which crippled many corporate e-mail systems and coasted billions of dollars in lost productivity is widely cited in research (Wang, Hidvegi and Whinstone 2000; Udo 2001; Weaver et al. 2000). A distributed denial of service attack occurred in February 2000 which disabled numerous large sites of businesses including eBay, Microsoft and Yahoo. Following the attack, which caused a 22 hour outage on eBay, eBay experienced an 18 percent drop in share price and an immediate 43 percent decline in business volume (O'Brien and Mercer 2002). Other security breaches, such as the Code Red worm, and various hacking attacks which have
disclosed databases of customer credit card details (egg: Leyden 2000) all lower customer and business confidence in e-commerce. More traditional crimes, have found their way online.

Most prominent of these IS fraud, which, in an online environment, can be perpetrated in ways unique to e-commerce (for example, the faceless nature of the Internet means that properly identifying customers is difficult, if not impossible). Not surprisingly then, e-commerce security and protecting against electronic crime is a significant and fertile field of research (Cerpa and Jamieson 2001). Security controls are necessary to mitigate these risks and remove some of the barriers inhibiting the growth of e-commerce. Accordingly, a myriad of control frameworks and approaches have been devised in the research literature. These include research on introducing security into e-commerce systems at design-time, through secure mechanism design (Wang, Hidvegi and Whinstone 2000), the development of architectures for real-time intrusion monitoring (Furnell and Dowland 2000) and the use of continuous auditing (Cerpa and Jamieson 2001).

This thesis addresses one facet of e-commerce security, albeit a crucial one. A central unit of e-commerce is the transaction. Transactions form the lifeblood of e-commerce in which information and money are traded amongst businesses and consumers. Thus, ensuring that these transactions are legitimate in nature (i.e. not fraudulent), and that participating parties receive what they expect from them, is a fundamental issue of e-commerce.

If businesses and consumers are unconvinced that performing transactions on the Internet is secure, a lack of trust will develop. This lack of trust acts as an inhibiting factor (Elliot and Fowell 2000). This
thesis focuses on transaction security or more specifically, the fraud auditing of transactions.

E-commerce has changed the face of most business functions in competitive enterprises. Internet technologies have seamlessly automated interface processes between customers and retailers, retailers and distributors, distributors and factories, and factories and their myriad suppliers. In general, e-commerce and ebusiness (henceforth referred to as e-commerce) have enabled on-line transactions. Also, generating large-scale real-time data has never been easier. With data pertaining to various views of business transactions being readily available, it is only apposite to seek the services of data mining to make (business) sense out of these data sets.

Data mining (DM) has as its dominant goal, the generation of non-obvious yet useful information for decision makers from very large databases. The various mechanisms of this generation include abstractions, aggregations, summarizations, and characterizations of data. These forms, in turn, are the result of applying sophisticated modeling techniques from the diverse fields of statistics, artificial intelligence, database management and computer graphics.

9.10 History

Early signification

The meaning of the term "electronic commerce" has changed over the last 30 years. Originally, "electronic commerce" meant the facilitation of commercial transactions electronically, usually using technology like Electronic Data Interchange (EDI) and Electronic Funds Transfer (EFT),
where both were introduced in the late 1970s, for example, to send commercial documents like purchase orders or invoices electronically.

The 'electronic' or 'e' in e-commerce refers to the technology/systems; the 'commerce' refers to traditional business models. E-commerce is the complete set of processes that support commercial business activities on a network. In the 1970s and 1980s, this would also have involved information analysis. The growth and acceptance of credit cards, automated teller machines (ATM) and telephone banking in the 1980s were also forms of e-commerce. However, from the 1990s onwards, this would include enterprise resource planning systems (ERP), data mining and data warehousing. Perhaps the earliest example of many-to-many electronic commerce in physical goods was the Boston Computer Exchange, a marketplace for used computers, launched in 1982. The first online information marketplace, including online consulting, was likely the American Information Exchange, another pre-Internet online system, introduced in 1991.

9.11 Activities

In the dot com era, it came to include activities more precisely termed "Web commerce" -- the purchase of goods and services over the World Wide Web, usually with secure connections (HTTPS, a special server protocol that encrypts confidential ordering data for customer protection) with e-shopping carts and with electronic payment services, like credit card payment authorizations.
Today, it encompasses a very wide range of business activities and processes, from e-banking to offshore manufacturing to e-Logistic. The ever growing dependence of modern industries on electronically enabled business processes gave impetus to the growth and development of supporting systems, including backend systems, applications and middleware. Examples are broadband and fiber-optic networks, supply-chain management software, customer relationship management software, inventory control systems and financial accounting software.

9.12 Web development

When the Web first became well-known among the general public in 1994, many journalists and pundits forecast that e-commerce would soon become a major economic sector. However, it took about four years for security protocols (like HTTPS) to become sufficiently developed and widely deployed. Subsequently, between 1998 and 2000, a substantial number of businesses in the United States and Western Europe developed rudimentary web sites.

Although a large number of "pure e-commerce" companies disappeared during the dot-com collapse in 2000 and 2001, many "brick-and-mortar" retailers recognized that such companies had identified valuable niche markets and began to add e-commerce capabilities to their Web sites. For example, after the collapse of online grocer Web van, two traditional supermarket chains, Albertsons and Safeway, both started e-commerce subsidiaries through which consumers could order groceries online.
The emergence of e-commerce also significantly lowered barriers to entry in the selling of many types of goods; accordingly many small home-based proprietors are able to use the internet to sell goods. Often, small sellers use online auction sites such as eBay, or sell via large corporate websites like Amazon.com, in order to take advantage of the exposure and setup convenience of such sites.

$259 billion of online sales including travel are expected in 2007 in USA, a 18 percent increase from the previous year, as forecasted by the "State of Retailing Online 2007" report from the National Retail Federation (NRF) and Shop.org.[I] Success factors

In many cases, an e-commerce company will survive not only based on its product, but by having a competent management team, good post-sales services, well-organized business structure, network infrastructure and a secured, well designed website. A company that wants to succeed will have to perform two things: Technical and organizational aspects and customer-oriented. Following factors will make business of companies succeed in e-commerce:

**9.13 Technical and organizational aspects**

Sufficient work done in market research and analysis. E-commerce is not exempt from good business planning and the fundamental laws of supply and demand. Business failure is as much a reality in e-commerce as in any other form of business.

A good management team armed with information technology strategy. A company's IT strategy should be a part of the business re-design process. Providing an easy and secured way for customers to effect transactions.
Credit cards are the most popular means of sending payments on the internet, accounting for 90% of online purchases. In the past, card numbers were transferred securely between the customer and merchant through independent payment gateways. Such independent payment gateways are still used by most small and home businesses. Most merchants today process credit card transactions on site through arrangements made with commercial banks or credit cards companies.

Providing reliability and security. Parallel servers, hardware redundancy, fail-safe technology, information encryption, and firewalls can enhance this requirement. Providing a 360-degree view of the customer relationship, defined as ensuring that all employees, suppliers, and partners have a complete view, and the same view, of the customer. However, customers may not appreciate the big brother experience. Constructing a commercially sound business model.

Engineering an electronic value chain in which one focuses on a "limited" number of core competencies -- the opposite of a one-stop shop. (Electronic stores can appear as either specialist or generalist if properly programmed.)

Operating on or near the cutting edge of technology and staying there as technology changes (but remembering that the fundamentals of commerce remain indifferent to technology).

Setting up an organization of sufficient alertness and agility to respond quickly to any changes in the economic, social and physical environment.
Providing an attractive website. The tasteful use of color, graphics, animation, photographs, fonts, and white-space percentage may aid success in this respect. Streamlining business processes, possibly through re-engineering and information technologies.

Providing complete understanding of the products or services offered, which not only includes complete product information, but also sound advisors and selectors.

Naturally, the e-commerce vendor must also perform such mundane tasks as being truthful about its product and its availability, shipping reliably, and handling complaints promptly and effectively. A unique property of the Internet environment is that individual customers have access to far more information about the seller than they would find in a brick-and-mortar situation. (Of course, customers can, and occasionally do, research a brick-and-mortar store online before visiting it, so this distinction does not hold water in every case.)

9.14 Customer experience

A successful e-commerce organization must also provide an enjoyable and rewarding experience to its customers. Many factors go into making this possible. Such factors include:

**Providing value to customers:** Vendors can achieve this by offering a product or product-line that attracts potential customers at a competitive price, as in nonElectronic commerce.

**Providing service and performance:** Offering a responsive, user-friendly purchasing experience, just like a flesh-and-blood retailer, may go some way to achieving these goals.
Providing an incentive for customers to buy and to return. Sales promotions to this end can involve coupons, special offers, and discounts. Cross-linked websites and advertising affiliate programs can also help.

Providing personal attention: Personalized web sites, purchase suggestions, and personalized special offers may go some of the way to substituting for the face-to-face human interaction found at a traditional point of sale.

Providing a sense of community: Chat rooms, discussion boards, soliciting customer input and loyalty programs (sometimes called affinity programs) can help in this respect.

Owning the customer’s total experience: E-tailers foster this by treating any contacts with a customer as part of a total experience, an experience that becomes synonymous with the brand.

Letting customers help themselves: Provision of a self-serve site, easy to use without assistance, can help in this respect. This implies that all product information is available, cross-sell information, advise for product alternatives, and supplies & accessory selectors.

Helping customers do their job of consuming: E-tailers and online shopping directories can provide such help through ample comparative information and good search facilities. Provision of component information and safety-and-health comments may assist e-tailers to define the customers’ job.
9.15 Problems

Product suitability

Certain products or services appear more suitable for online sales; others remain more suitable for offline sales. While credit cards are currently the most popular means of paying for online goods and services, alternative online payments will account for 26% of e-commerce volume by 2009 according to Client.[2]

Many successful purely virtual companies deal with digital products, (including information storage, retrieval, and modification), music, movies, office supplies, education, communication, software, photography, and financial transactions. Examples of this type of company include: Google, eBay and PayPal. Other successful marketers such as use Drop shipping or Affiliate marketing techniques to facilitate transactions of tangible goods without maintaining real inventory. Examples include numerous sellers on eBay.

Virtual marketers can sell some non-digital products and services successfully. Such products generally have a high value-to-weight ratio, they may involve embarrassing purchases, they may typically go to people in remote locations, and they may have shut-ins as their typical purchasers. Items which can fit through a standard letterbox - such as music CDs, DVDs and books - are particularly suitable for a virtual marketer, and indeed Amazon.com, one of the few enduring dot-com companies, has historically concentrated on this field.

Products such as spare parts, both for consumer items like washing machines and for industrial equipment like centrifugal pumps, also seem
good candidates for selling online. Retailers often need to order spare parts specially, since they typically do not stock them at consumer outlets -- in such cases, e-commerce solutions in spares do not compete with retail stores, only with other ordering systems. A factor for success in this niche can consist of providing customers with exact, reliable information about which part number their particular version of a product needs, for example by providing parts lists keyed by serial number. Purchases of pornography and of other sex-related products and services fulfill the requirements of both virtuality (or if non-virtual, generally high-value) and potential embarrassment; unsurprisingly, provision of such services has become the most profitable segment of e-commerce. [citation needed]

There are also many disadvantages of e-commerce; one of the main ones is fraud. This is where your details (name, bank Card. number, age, and national insurance number) are entered into what look to be a safe site but really it is not. These details can then be used to steal money from you and can be used to buy things on line that you are completely unaware of until it is too late. If this information is leaked into the wrong hands, people are able to steal your identity, and commit more fraud crimes under your name.

Products less suitable for e-commerce include products that have a low value-to-weight ratio, products that have a smell, taste, or touch component, products that need trial fittings - most notably clothing - and products where color integrity appears important. Nonetheless, Tesco.com has had success delivering groceries in the UK, albeit that many of its goods are of a generic quality, and clothing sold through the internet is big business in the U.S. Also, the recycling program Cheap
cycle sells goods over the internet, but avoids the low value-to-weight ratio problem by creating different groups for various regions, so that shipping costs remain low.

**Acceptance**

Consumers have accepted the e-commerce business model less readily than its proponents originally expected. Even in product categories suitable for ecommerce, electronic shopping has developed only slowly. Several reasons might account for the slow uptake, including:

**Concerns about security.** Many people will not use credit cards over the Internet due to concerns about theft and credit card fraud.

**Lack of instant gratification with most e-purchases (non-digital purchases).** Much of a consumer’s reward for purchasing a product lies in the instant gratification of using and displaying that product. This reward does not exist when one's purchase does not arrive for days or weeks.

The problem of access to web commerce, mainly for poor households and for developing countries. Low penetration rates of Internet access in some sectors greatly reduces the potential for e-commerce.

Electronic commerce relates to the usage of electronic communication networks to conduct business transactions. The emergence of e-commerce in society has profoundly impacted upon how people manage and conduct business. It has changed how companies operate internally, whilst also giving them the opportunity to expand into new, previously unstoppable, markets. The ubiquitous nature of e-commerce has also accelerated globalization as instantaneous information exchange is
possible anywhere on the planet. The smallest of firms employing ecommerce potentially have access to a global market. The largest of firms have redefined or remodeled themselves in response to the advent of ecommerce. Indeed, e-commerce not only affects the way business is conducted, but its nascent influence reverberates through to changing the world economy. None the less, this new dimension of business has problems, barriers and disadvantages that inhibit its expansion. It is a phenomenon undergoing continual, rapid change and maturity. Increasing levels of integration of e-commerce systems into business has led to an increasing level of reliance on these systems. Inter organizational systems and globally distributed data means that ensuring the availability, integrity and confidentiality of the information these systems process is of paramount importance.

Unfortunately, it is the pace of e-commerce system development that amplifies the huge challenge of ensuring those same systems are secure. This thesis examines specifically the threat of fraud which is the largest security risk that has direct implications upon the revenue flows and costs of a business.

It is for this reason that e-commerce security should receive collaborative attention from research institutions and commercial organizations, such that security may be able to keep in step with the latest advancements in e-commerce. Current approaches tend to be fragmented in nature, due to the wide variety of systems in the marketplace, and the trend of inter organizational systems integration means that unless a more unified approach to Suring up security is taken, the rise in number of points a large e-commerce system has that are exploitable will be increasingly detrimental. A system vulnerable to different types of fraud stands to be
a large liability over more traditional means of business and undermines the attractiveness of e-commerce. Moreover, customers that perceive that their ecommerce transactions are susceptible to fraud are not encouraged to engage in such business. Only when security systems are developed that can, with a reasonable degree of effectiveness, detect fraud, will this barrier to e-commerce uptake be assuaged.

### 9.16 Factors for implementation DM for success

The success of a DM exercise is driven to a very large extent by the following factors

**Availability of data with rich descriptions:** This means that unless the relations captured in the database are of high degree, extracting hidden patterns and relationships among the various attributes will not make any practical sense. Availability of a large volume of data: This is mostly mandated for statistical significance of the rules to hold. Absence of say, at least a hundred thousand transactions will most likely reduce the usefulness of the rules generated from the transactional database.

**Reliability of the data available:** Although a given terabyte database may have hundreds of attributes per relation, the DM algorithms run on this dataset may be rendered defunct if the data itself was generated by manual and error prone means and wrong default values were set. Also, the lesser the integration with legacy applications, the better the accuracy of the dataset.
Ease of quantification of the return on investment (ROI) in DM: Although the earlier three factors may be favorable, unless a strong business case can be easily made, investments in the next level DM efforts may not be possible. In other words, the utility of the DM exercise needs to be quantified vis-a-vis the domain of application.

Ease of interfacing with legacy systems: It is commonplace to find large organizations run on several legacy systems that generate huge volumes of data. A DM exercise which is usually preceded by other exercises like extract, transformation and loading (ETL), data filtering etc, should not add more overheads to system integration.

It must now be noted that e-commerce data, being the result of on-line transactions, do satisfy all the above proper criteria for data mining. We observe that once the back-end databases are properly designed to capture customer buying behavior, and provided that default data take care of missing and nonexistent data, the first issue of availability of data with rich descriptions is taken care of. Similarly, the reliability of data collected is also ensured because it is possible to increase the so called no-touch-throughput in e-commerce transactions. Technologies like ebXML, BizTalk and Rosetta Net enhance the quality of data that is generated.

The ROI in DM exercises related to e-commerce can be easily quantified. For instance, mining the web logs certainly enhances web server architecture-related decisions. Improved web server availability results in faster transactions, thus increasing the revenue. Observe that increasing the number of transactions directly results in improved profits. Lastly, e-commerce systems usually follow the MVC (Model-View-Controller) pattern with the business execution systems conforming to the model
tier, the browser being the view tier and interfacing mechanisms like Java Servlets or Microsoft ASP forming the controller tier. Data mining in e-commerce mostly relies on the controller for generating the data to mine on. Thus integration issues also do not surface in this case. In summary, it is little surprise that e-commerce is the killer application for data mining.

### 9.17 A review of data-mining methods

Given a truly massive amount of data, the challenge in data mining is to unearth hidden relationships among various attributes of data and between several snapshots of data over a period of time. These hidden patterns have enormous potential in predictions and personalization's in e-commerce. Data mining has been pursued as a research topic by at least three communities: the statisticians, the artificial intelligence researchers, and the database engineers. We now present a brief overview of some of the features of each of these approaches.

### 9.18 Role of statistics in data mining

Extracting causal information from data is often one of the principal goals of data mining and more generally of statistical inference. Statisticians have done aggregate data analyses on data for decades; thus DM has actually existed from the time large scale statistical modeling has been made possible.

Statisticians consider the causal relationship between the dependent variables and independent variables as proposed by the user (usually the
domain expert), and endeavor to capture the degree and nature of dependence between the variables. Modeling methods include simple linear regression, multiple regressions, and nonlinear regression. Such models are often parameter driven and are arrived at after solving attendant optimization models. For a more detailed overview of regression methods.

The regression methods may be considered analogous to the association rules in data mining. In the latter case, rule-mining algorithms propose the correlation of item sets in a database, across various attributes of the transactions. For instance, rules could be of the form if a customer visits Page A.html, 90% of the times she will also visit Page B.html. We assume here that the database (here, the web logs) has transactions recorded on a per-customer basis. Each record in the database indicates whether the customer visited a page during her entire session. Such rules can and need to be validated using the well-known statistical regression methods. Also, in some cases, the number of association rules may be very large. To draw meaningful rules that have real business value, it may be worthwhile to select the statistically most significant set of rules from the large pool of rules generated by a rule-mining algorithm. We note that methods such as principal components analysis and factor analysis could be used to unearth hidden classes or clusters. Time series modeling on the other hand, is more relevant in sequential mining. This is used to unearth correlations and patterns in temporally ordered data. For a more detailed overview of time series methods.

Model validations based on hypothesis testing start with an initial hypothesis of (usually) a linear relation between variable sets X and Y, and after conducting tests, the data are either shown to prove or disprove
the hypothesis. Data mining involves designing a search architecture requiring evaluation of hypotheses at the stages of the search, evaluation of the search output, and appropriate use of the results. Although statistics may have little to offer in understanding search architectures, it has indeed a great deal to offer in evaluation of hypotheses in the above stages. While the statistical literature has a wealth of technical procedures and results to offer data mining, one has to take note of the following while using statistics to validate the rules generated using data mining.

• Prove that the estimation and search procedures used in data mining are consistent under conditions reasonably assumed to apply in applications

• Use and reveal uncertainty and not hide it; some data-mining approaches ignore the causal relations due to lack of sufficient data. Such caveats can be unearthed using statistical methods.

• Calibrate the errors of search to take advantages of model averaging. This is relevant where predicting the future is important, as in data mining applied to forecasting a time series. Model averaging is beneficial where several models may be relevant to build a forecast.

9.19 The role of AI in data mining

Artificial intelligence, on the other hand, has provided a number of useful methods for DM. Machine learning is a set of methods that enable a computer to learn relations from the given data sets. With minimal or no hypothesis from the user, learning algorithms do crone up with meaningful relations and also explain them well. Some of the most
popular learning systems include the neural networks and support vector machines. We briefly present the relevant issues below.

Neural networks are predominantly used to learn linear and nonlinear relationships between variables of interest. The architecture, in general, consists of a preceptor with input and output nodes with weighted edges connecting the two nodes. A neural network with two layers is thus a bipartite acyclic graph. The preceptor, which is the learning machine, is 'trained' in order to arrive at an optimal 'weight vector'. The output is then expressed as a (weighted) linear combination of the inputs. Learning consists of solving an underlying optimization model which is solved using gradient descent based methods.

It is worth noting here that the corresponding statistical methods available for estimating nonlinear relationships are based on the Maximum Likelihood Estimate problem. This problem is rather unwieldy since it requires the solution of highly nonlinear optimization problems; this result in tedious computations involved in solving algebraic equations. It is here that neural networks outperform their statistical counterparts, by resorting to the supervised learning methods based on gradient descent to solve such estimation problems. In other words, instead of explicitly solving equations to arrive at the maximum likelihood weights, neural networks 'learn' these weights via gradient descent based search methods.

To learn more complex relationships including multi-variate nonlinear ones, it is not uncommon to have more layers than two. Such layers are called the hidden layers. The empiricism associated with neural networks is due to the no availability of methods that would help fix the rate of convergence and the optimal number of layers. In the above learning
process, if the outputs are Boolean, the problem is essentially a supervised learning mechanism to classify data sets. In such cases, often-times, a sigmoidal function (a nonlinear transformation) is applied to obtain the relevant output.

Apart from learning relationships as above, neural networks are also useful in clustering data sets. The most popular method available to cluster data sets is the K-means algorithm. Given an M-dimensional data set, the idea is to try and locate the minimal number of cancroids around which the data set clusters itself. Thus the onus is to define an appropriate distance vector that helps partition the data sets into as minimally overlapping sub-sets as possible. In general, Euclidean distance metrics (including the L2, L1, and the Ll norms) are proposed for 'optimally' partitioning a given data set. The optimization is again based on minimizing the sum of squares of an appropriate error term, using classical gradient based methods.

The advantages of neural networks over the conventional statistical analysis methods are as follows. First, neural networks are good at modeling nonlinear relationships and interaction while conventional statistical analysis in most cases assumes linear relationship between independent variables and dependent variables. Neural networks build their own models with the help of learning process whether the relationships among variables are linear or not. Secondly, neural networks perform well with missing or incomplete data. A single missing value in regression analysis leads to removal of the entire observation or removal of the associated variable from all observations in the data set being analyzed. However, neural networks update weights between input, output, and intermediate nodes, so that even incomplete data can
contribute to learning and produce desired output results. Finally, neural networks do not require scale adjustment or statistical assumptions, such as normality or independent error terms.

Artificial intelligence based methods using neural networks are used in clustering and classification methods of data mining. They can also be used in sequential mining. For instance, market basket analysis which concerns itself with identifying hidden customer segments could be solved using neural networks with unsupervised learning. An online web store may want to provide different grades of service to its users depending on the frequency of customers' visits to their websites. Identifying the basket of such customer segments could be done using clustering methods. Finally, if the web store wants to identify the factors contributing to repeat customers, they could use the nonlinear regression expressions obtained using neural networks.

9.20 The role of database research in data mining

Keeping in mind that data mining approaches rely heavily on the availability of high quality data sets, the database community has invented an array of relevant methods and mechanisms that need to be used prior to any DM exercise. Extract, transform and load (ETL) applications are worthy of mention in this context. Given an enterprise system like an enterprise resource planning system (ERP), it is likely that the number of transactions that happen by the minute could run into hundreds, if not thousands. Data mining can certainly not be run on the transaction databases in their native state. It requires to be extracted at periodic intervals, transformed into a form usable for analysis, and loaded on to the servers and applications that work on the transformed
data. Today, software systems exist in the form of data warehousing solutions that are often bundled with the ERP system, to perform this complex and important task.

It is to be observed that data warehouses are essentially snapshots of transactional data aggregated along various dimensions (including time, geographies, demographics, products etc.) In order to run data mining algorithms, it is common practice to use the data available in the data warehouse rather than by running real time scripts to fetch transactional data. This is for the simple reason that for practical purposes, it is sufficient to include snapshots of data taken at say, weekly or monthly basis, for analysis. Real-time data is not relevant for tactical decision making, which is where data mining is used. Data warehousing is nevertheless fraught with technological challenges.

We note that database researchers have predominantly investigated association rule-mining within the field of DM. When one has terabytes of data available, the goal of database engineers in DM is to create structures and mechanisms to efficiently read in data into memory and run algorithms like A priori. Such algorithms assume the so-called item sets. Consider a database where the transactions pertain to a retail store. Customers buy various products and each transaction records the products bought by the customer. Observe that such databases can grow enormously in size, especially for large retailers who have web storefronts, like Amazon. Com. Each item set is a record in the database, with attributes mentioning if a particular product was purchased or not. The algorithms compute, given a certain support and confidence, the rules that apply on the given item sets.
A rule is of the form $X \rightarrow Y$, where $X$ and $Y$ are attribute subsets of the original set of attributes. For instance, a rule could be that, of the 80% transactions where customers bought product subset $X$, they also bought product subset $Y$, and this holds good for 50% of the total transactions. Here, support is defined as the percentage of transactions where both the attributes had a value 1 (indicating that both were bought). Confidence is defined as the percentage of records with attribute value 1 for subset $Y$ whenever the value was 1 for subset $X$. In the above case, the support is 50% while the confidence is 80%.

It is to be noted that A priori-like algorithms work on a given item set only. If the underlying transactional database is dynamic, then there are methods known as incremental mining proposed by the database community. Such methods resort to minimizing the number of passes over a given database for computing the support and confidence values in rule-mining.

Concerning e-commerce itself, it may be noted that data warehousing may not be required depending on the application. For instance, if one is keen on analyzing the click stream patterns from users, enhancements of the web-logging schemas could be just sufficient to provide the right kind of data for mining. On the other hand, if the online store has an ERP system running at the back end, then ETL applications may well be required.

With respect to rule-mining itself, an application that begs attention is that of analyzing the customer's pattern of website visits so that the websites could be designed and presented better. For instance, if the customer visits two or three levels deep into a website in order to conduct a transaction (like booking a ticket online, searching for availability of
trains etc.), then rule-mining could be used to unearth such patterns and identify the desirability of keeping the most frequently visited subsets of websites/homepages together. This will surely enable the customer to locate information easily while conducting a transaction at ease.

9.21 E-Commerce and data mining

In this section, we survey articles that are very specific to DM implementations in e-commerce. The salient applications of DM techniques are presented first. Later in this section, architecture and data collection issues are discussed.

9.22 DM in customer profiling

It may be observed that customers drive the revenues of any organization. Acquiring new customers, delighting and retaining existing customers, and predicting buyer behavior will improve the availability of products and services and hence the profits. Thus the end goal of any DM exercise in e-commerce is to improve processes that contribute to delivering value to the end customer. Consider an on-line store like http://www.dell.com where the customer can configure a PC of his/her choice, place an order for the same, track its movement, as well as pay for the product and services. With the technology behind such a web site, Dell has the opportunity to make the retail experience exceptional. At the most basic level, the information available in web log files can illuminate what prospective customers are seeking from a site. Are they purposefully shopping or just browsing? Buying something they're familiar with or something they know little about? Are they shopping from home, from work, or from a hotel dial-up? The information available
in log files is often used to determine what profiling can be dynamically processed in the background and indexed into the dynamic generation of HTML, and what performance can be expected from the servers and network to support customer service and make e-business interaction productive.

Companies like Dell provide their customers access to details about all of the systems and configurations they have purchased so they can incorporate the information into their capacity planning and infrastructure integration. Back-end technology systems for the website include sophisticated DM tools that take care of knowledge representation of customer profiles and predictive modeling of scenarios of customer interactions. For example, once a customer has purchased a certain number of servers, they are likely to need additional routers, switches, load balancers, backup devices etc. Rule-mining based systems could be used to propose such alternatives to the customers.

9.23 DM in recommendation systems

Systems have also been developed to keep the customers automatically informed of important events of interest to them. The article discusses an intelligent framework called PENS that has the ability to not only notify customers of events, but also to predict events and event classes that are likely to be triggered by customers. The event notification system in PENS has the following components: Event manager, event channel manager, registries, and proxy manager. The event prediction system is based on association rule-mining and clustering algorithms. The PENS system is used to actively help an e-commerce service provider to forecast the demand of product categories better. Data mining has also
been applied in detecting how customers may respond to promotional offers made by a credit card e-commerce company. Techniques including fuzzy computing and interval computing are used to generate if-then-else rules.

A method to build customer profiles in e-commerce settings, based on product hierarchy for more effective personalization. They divide each customer profile into three parts: basic profile learned from customer demographic data; preference profile learned from behavioral data, and rule profile mainly referring to association rules. Based on customer profiles, the authors generate two kinds of recommendations, which are interest recommendation and association recommendation. They also propose a special data structure called profile tree for effective searching and matching.

**9.24 DM in web personalization**

A comprehensive overview of the personalization process based on web usage mining. In this context, the author discusses a host of web usage mining activities required for this process, including the preprocessing and integration of data from multiple sources, and common pattern discovery techniques that are applied to the integrated usage data. The goal of this paper is to show how pattern discovery techniques such as dusting, association rule-mining, and sequential pattern discovery, performed on web usage data, can be leveraged effectively as an integrated part of a web personalization system. The author observes that the log data collected automatically by the Web and application servers represent the fine-grained navigational behavior of visitors.
9.24 (a) Data to be captured by web logs

Depending on the goals of the analysis, e-commerce data need to be transformed and aggregated at different levels of abstraction. E-Commerce data are also further classified as usage data, content data, structure data, and user data. Usage data contain details of user sessions and page views. The content data in a site are the collection of objects and relationships that are conveyed to the user. For the most part, the data comprise combinations of textual material and images. The data sources used to deliver or generate data include static HTML/XML pages, images, video clips, sound files, dynamically generated page segments from scripts or other applications, and collections of records from the operational database(s).

Site content data also include semantic or structural metadata embedded within the site or individual pages, such as descriptive keywords, document attributes, semantic tags, or HTTP variables. Structure data represent the designer's view of the content organization within the site. This organization is captured via the inter-page linkage structure among pages, as reflected through hyperlinks. Structure data also include the intra-page structure of the content represented in the arrangement of HTML or XML tags within a page. Structure data for a site are normally captured by an automatically generated site map which represents the hyperlink structure of the site. The operational database(s) for the site may include additional user profile information. Such data may include demographic or other identifying information on registered users, user ratings on various objects such as pages, products, or movies, past purchase or visit histories of users, as well as other explicit or implicit representations of users' interests.
Once the data types are clear, data preparation is easily achieved by processes such as data cleansing, page view identification, user identification, session identification, the inference of missing references due to caching, and transaction (episode) identification. For example, in collaborative filtering applications, such weights may be determined based on user ratings of items. In most web usage mmmg tasks, the focus is generally on anonymous user navigational activity where the primary sources of data are server logs. This allows one to choose two types of weights for page views: weights can be binary, representing the existence or non-existence of a page view in the transaction; or they can be a function of the duration of the page view in the user's session.

9.24(b) an illustration

For example, consider a site with 6 page views A, B, C, D, E, and F. Assuming that the page view weights associated with a user transaction are determined by the number of seconds spent on them, a typical transaction vector may look like: (11, 0, 22, 5, 127, 0). In this case, the vector indicates that the user spent 11 seconds on page A, 22 seconds on page C, 5 seconds on page D, and 127 seconds on page E. The vector also indicates that the user did not visit pages B and F during this transaction. Given this representation, the set of all m user transactions can be conceptually viewed as an m x n transaction-page view matrix which we shall denote by TP. This transaction-page view matrix can then be used to perform various data mining tasks. For example, similarity computations can be performed among the transaction vectors (rows) for clustering and k-NN neighborhood formation tasks, or an association rule discovery algorithm, such as Apriority, can be applied (with page views as items) to find frequent item sets of page views.
9.25 DM and multimedia e-commerce

Applications in virtual multimedia catalogs are highly interactive, as in e-malls selling multimedia content based products. It is difficult in such situations to estimate resource demands required for presentation of catalog contents. A method to predict presentation resource demands in interactive multimedia catalogs. The prediction is based on the results of mining the virtual mall action log file that contains information about previous user interests and browsing and buying behavior.

DM and buyer behavior in e-commerce

For a successful e-commerce site, reducing user-perceived latency is the second most important quality after good site-navigation quality. The most successful approach towards reducing user-perceived latency has been the extraction of path traversal patterns from past user's access history to predict future user traversal behavior and to pre fetches the required resources. However, this approach is suited for only non-e-commerce sites where there IS no purchase behaviour. Describe an approach to predict user behavior in e-commerce sites. The core of their approach involves extracting knowledge from integrated data of purchase and path traversal patterns of past users (obtainable from web server logs) to predict the purchase and traversal behavior of future users.

Web sites are often used to establish a company's image, to promote and sell goods and to provide customer support. The success of a web site affects and reflects directly the success of the company in the electronic market. A methodology to improve the success of web sites, based on the exploitation of navigation-pattern discovery. In particular, the authors present a theory, in which success is modeled on the basis of the
navigation behavior of the site's users. They then exploit web usage miner (WUM), a navigation pattern discovery miner, to study how the success of a site is reflected in the users' behavior. With WUM the authors measure the success of a site's components and obtain concrete indications of how the site should be improved.

In the context of web mining, clustering could be used to cluster similar click streams to determine learning behaviors in the case of e-learning or general site access behaviors in e-commerce. Most of the algorithms presented in the literature to deal with clustering web sessions treat sessions as sets of visited pages within a time period and do not consider the sequence of the click-stream visitation. This has a significant consequence when comparing similarities between web sessions. Propose an algorithm based on sequence alignment to measure similarities between web sessions where sessions are chronologically ordered sequences of page accesses.

**Enabling data collection in e-commerce**

It may be observed that there are various ways of procuring data relevant to ecommerce DM. Web server log files, web server plug-ins (instrumentation), TCP/IP packet sniffing, application server instrumentation are the primary means of collecting data. Other sources include transactions that the user performs, marketing programs (banner advertisements, emails etc), demographic (obtainable from site registrations and subscriptions), call centers and ERP systems.

It is quite common to expend about 80% of any DM effort in e-commerce in data filtering. This is largely in part to the heavy reliance on the web logs that are generated by the HTTP protocol. This protocol being
stateless, it becomes very difficult to cull out customer buying behavior-related information along with the product details. Architecture for supporting the integration of DM and ecommerce. The architecture is found to dramatically reduce the preprocessing, cleaning, and data understanding effort. They emphasize the need for data collection at the application server layer and not the web server, in order to support tagging of data and metadata that is essential to the discovery process. They also describe the data transformation bridges required from the transaction processing systems and customer event streams (e.g. click streams) to the data warehouse.

9.26 Analyzing web transactions

Once the data are collected via any of the above mentioned mechanisms, data analysis could follow suit. This could be done along session level attributes, customer attributes, product attributes and abstract attributes. Session level analysis could highlight the number of page views per session, unique pages per session, time spent per session, average time per page, fast vs. slow connection etc. Additionally, this could throw light on whether users went through registration, if so, when, did the users look at the privacy statement; did they use search facilities, etc. The user level analysis could reveal whether the user is an initial or repeat or recent visitor/purchaser; whether the users are readers, browsers, heavy spenders, original referrers etc.

The view of web transactions as sequences of page views allows one to employ a number of useful and well-studied models which can be used to discover or analyze user navigation patterns. One such approach is to model the navigational activity in the website as a Markov chain. In the
context of web transactions, Markov chains can be used to model transition probabilities between page views. In web-usage analysis, they have been proposed as the underlying modeling machinery for web pre-fetching applications or to minimize system latencies.

A new approach called on-line analytical mining for web data.

Their approach consists of data capture, web house construction, pattern discovery and pattern evaluation.

The authors describe the challenges in each of these phases and present their approach for web usage mining. Their approach is useful in determining the most profitable customers, the difference between buyers and non-buyers, identification of website parts that attract most visits, parts of website that are session killers, parts of the site that lead to the most purchases, identifying the typical path of customers that leads to a purchase or otherwise etc. The web house is akin to the data warehouse.

9.27 An architecture for DM

In a B2B e-commerce setting, it is very likely that vendors, customers and application service providers (ASP) (usually the middlemen) have varying DM requirements. Vendors would be interested in DM tailored for market basket analysis to know customer segments. On the other hand, end customers are keen to know updates on seasonal offerings and discounts all the while. The role of the ASP is then to be the common meeting ground for vendors and customers. Propose a distributed DM architecture that enables a DM to be conducted in such a naturally distributed environment. The proposed distributed data mining system is
intended for the ASP to provide generic data mining services to its subscribers. In order to support the robust functioning of the system it possesses certain characteristics such as heterogeneity, costing infrastructure availability, presence of a generic optimization engine, security and extensibility. Heterogeneity implies that the system can mine data from heterogeneous and distributed locations. The proposed systems designed to support user requirements with respect to different distributed computing paradigms (including the client-server and mobile agent based models). The costing infrastructure refers to the system having a framework for estimating the costs of different tasks. This implies that a task that requires higher computational resources and/or faster response time should cost the users more on a relative scale of costs. Further, the system should be able to optimize the distributed data mining process to provide the users with the best response time possible (given the constraints of the mining environment and the expenses the user is willing to incur). The authors have indeed designed and implemented such a framework.

Maintaining security implies that in some instances, the user might be mining highly sensitive data that should not leave the owner's site. In such cases, the authors provide the option to use the mobile-agent model where the mining algorithm and the relevant parameters are shipped to the data site and at the end of the process the mobile agent is destroyed on the site itself. The system is extensible to provide for a wide range of mining algorithms. The authors provide a facility wherein the user can register their algorithms with the ASP for use in their specific distributed DM jobs.
9.28 Cases in e-commerce data mining

In this section, we first present an interesting application of DM in e-commerce. We then present some important lessons learnt by some authors while implementing DM in e-commerce.

(A) Distributed spatial data mining

In various e-commerce domains involving spatial data (real estate, environmental planning, and precision agriculture), participating businesses may increase their economic returns using knowledge extracted from spatial databases. However, in practice, spatial data is often inherently distributed at multiple sites. Due to security, competition and a lack of appropriate knowledge discovery algorithms, spatial information from such physically dispersed sites is often not properly exploited. To develop a distributed spatial knowledge discovery system for precision agriculture. In the proposed system, a centralized server collects proprietary site-specific spatial data from subscribed businesses as well as relevant data from public and commercial sources and integrates knowledge in order to provide valuable management information to subscribed customers. Spatial data mining software interfaces this database to extract interesting and novel knowledge from data. Specific objectives include a better understanding of spatial data, discovering relationships between spatial and non spatial data, construction of spatial knowledge-bases, query optimization and data reorganization in spatial databases. Knowledge extracted from spatial data can consist of characteristic and discriminate rules, prominent structures or clusters, spatial associations and other forms.
Challenges involved in spatial data mining include multiple layers of data, missing attributes and high noise due to a low sensibility of instruments and to spatial interpolation on sparsely collected attributes. To address some of these problems, data are cleaned by removing duplicates, removing outliers and by filtering through a median filter with a specified window size.

The goal of precision agriculture management is to estimate and perform site specific crop treatment in order to maximize profit and minimize environmental damage. Through a knowledge discovery (KDD) process, to propose learning algorithms that perform data modeling using data sets from different fields in possibly different regions and years. Each dataset may contain attributes whose values are not manageable (e.g. topographic data), as well as those attributes that are manageable (e.g. nutrient concentrations).

In order to improve prediction ability when dealing with heterogeneous spatial data, an approach employed in the proposed system by is based on identifying spatial regions having similar characteristics using a clustering algorithm. A clustering algorithm is used for partitioning multivariate data into meaningful subgroups (clusters), so that patterns within a cluster are more similar to each other than are patterns belonging to different clusters.

Local regression models are built on each of these spatial regions describing the relationship between the spatial data characteristics and the target attribute.
(B) DM applied to retail e-commerce

Kohavi et al (2004) have attempted a practical implementation of data mining in retail e-commerce data. They share their experience in terms of lessons that they learnt. They classify the important issues in practical studies, into two categories: business-related and technology-related. We now summarize their findings on the technical issues here.

(1) Collecting data at the right level of abstraction is very important. Web server logs were originally meant for debugging the server software. Hence they convey very little useful information on customer-related transactions. Approaches including seasoning the web logs may yield better results. A preferred alternative would be have the application server itself log the user related activities. This is certainly going to be richer in semantics compared to the state-less web logs, and is easier to maintain compared to state-full web logs.

(2) Designing user interface forms needs to consider the DM issues in mind. For instance, disabling default values on various important attributes like Gender, Marital status, Employment status, etc., will result in richer data collected for demographical analysis. The users should be made to enter these values, since it was found by Kohavi et al (2004) that several users left the default values untouched.

(3) Certain important implementation parameters in retail e-commerce sites like the automatic time outs of user sessions due to perceived inactivity at the user end need to be based not purely on DM algorithms, but on the relative importance of the users to the organization. It should not turn out that large clients are made to lose their shopping carts due to the time outs that were fixed based on a DM of the application logs.
Generating logs for several million transactions is a costly exercise. It may be wise to generate appropriate logs by conducting random sampling, as is done in statistical quality control. But such a sampling may not capture rare events, and in some cases like in advertisement referral based compensations, the data capture may be mandatory. Techniques thus need to be in place that can do this sampling in an intelligent fashion.

Auditing of data procured for mining, from data warehouses, is mandatory. This is due to the fact that the data warehouse might have collated data from several disparate systems with a high chance of data being duplicated or lost during the ETL operations.

Mining data at the right level of granularity is essential. Otherwise, the results from the DM exercise may not be correct.

9.29 In Future Securing Web Services

The IT industry has been talking about Web services for almost four years. Web services allows applications (e.g. automated business transactions, stock trading and order- tracking systems) to communicate with each other within organizations, across enterprises, and across the Internet in a loosely-coupled, platform- and programming language- independent manner. Several key standards have formed the foundation for Web services: XML (Extensible Markup Language), WSDL (Web Services Definition Language), SOAP (Simple Object Access Protocol), and UDDI (Universal Description, Discovery, and Integration).

Since, the key benefit of Web services is to deliver integrated & interoperable solutions, ensuring the integrity, confidentiality & security
is the most important key area that needs to be addressed for Web services.

Traditionally, the barriers of integration are due to the tight-coupling, where one application that calls another one is tied strongly by the function and the parameters. There is low flexibility or adaptability to changing environments or needs, due to:

1. Different programming languages
2. Different operating systems or hardware platforms
3. Different software vendors & in-house coding
4. It's difficult to integrate these systems internally
5. It's even harder to integrate with external business partners

9.30 What are Web Services?

According to W3C, a Web service is defined as: "A Web service is a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically WSDL). Other systems interact with the Web service in a manner prescribed by its description using SOAP messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards."

In substance, Web services are technology that allows applications to communicate with each other in a platform-, hardware- and programming language-independent manner. It uses XML based
protocols to describe a collection of operations that can be accessed, executed or data exchanged over the network. A group of Web services interacting together in this manner defines a particular Web service application in a Service-Oriented Architecture (SOA).

9.31 Web services exhibit the following definitive characteristics:

1. Web services communicate using platform-, hardware-independent and language-neutral Web protocols. These Web protocols ensure easy integration over the network & loosely coupling between applications.

2. A Web service provides an interface that can be called from another program. This application-to-application programming interface can be invoked from any type of application client or service.

3. A Web service is registered and can be located through a Web Service Registry. The registry enables service consumers to find services that match their needs.

9.32 Advantages of Web Services

1. Flexibility - Web services allow loose-coupling, which means that interactions between service applications may not break even there is a change. These universal interfaces can cope with inevitable changes in software caused by changing business needs.
2. Agility and Productivity - Rapid application assembly tools allow integration for new business opportunities or trying new business ideas.

3. Cost Savings - It allows automatic transactions, replace manual methods, reduce staffing requirements, replace paper processing & reduce errors.

4. Leverage Existing Investments - Web services provide existing or legacy software applications with service interfaces without changing the original applications, allowing them to fully operate in the service environment. Old software can be used in new ways by building Web services layer for universal access. This adapts existing applications to changing business conditions and customer needs.

5. Leverage Developer Skilllets - The plumbing code is generated automatically and can be integrated & tested with traditional methods.


9.33 Uses of Web Services

What can I do with Web services? While Web services provide all the advantages stated above, Web services allow us to implement as :
1. A credit card service that processes credit card transactions for a given account number.
2. A market data service that provides stock market data associated with a specified stock symbol
3. An airline service that provides flight schedule, availability, and reservation functionalities.

9.34 Web Services Architecture. Conceptually, Web services stack can be defined as the figures below.

1. Transport
2. Service Communication Protocol
3. Service Description
4. Service
5. Business Process
6. Service Registry
7. Policy - Security
8. Transaction
9. Management

![Web Services Architecture Diagram]
Web services are widely adopted standards such as HTTP and extensible Markup Language (XML). Typically, these standards are maintained by independent, nonprofit standards organizations.

A few of the major Web services standards groups are listed below:

**W3C (World Wide Web Consortium)** - The driving force behind the largest number of highly adopted standards in the Web services space including some Web building blocks such as HTML.

**OASIS** - Source of the original specification from which XML evolved, as well as the home of the current XML and Universal Description, Discovery and Integration (UDDI) specification.

**WS- I (Web Services Interoperability Organization)** - Acts as a watchdog group to ensure interoperability between implementations of Web services standards.

A typical Web services model consists of three entities:

1. Service providers who create Web services and publish them to the outside world by registering the services with service brokers.

2. Service brokers who maintain a registry of Published services.
3. Service requesters who find required services by searching the service broker's registry. Requesters than bind their applications to the service provider to use particular services.

9.37 Web Services Approach for SOA Architecture:

Web services allow applications interact with one another over the Web, so there is necessary for them to find one another, discover the information and patterns to interconnect. Therefore, Web services involve a family of related protocols to describe, deliver, and interact with services. And Web services require several related XML-based technologies to transport and to transform data into and out of programs and databases.

9.38 Web services are essentially founded upon four major technologies:

1. XML (extensible Markup Language) is the markup language that underlies most of the specifications used for Web services. XML is a
generic language that can be used to describe any kind of content in a structured way, separated from its presentation to a specific device.

2. WSDL (Web Services Description Language) - WSDL is a series of XML statements that constitute the definition for the interfaces of each service.

3. UDDI (Universal Description, Discovery and Integration) - UDDI lets Web services register their characteristics with a registry so that other applications can look them up.

4. SOAP (Simple Object Access Protocol) - SOAP provides the means for communication between Web services and client applications. It handles the issues of messaging, interface description, addressing and delivery.

9.39 The Importance of the Security in Web Services

In February / March 2003 CBDI Forum carried out a survey of its subscribers who had practical experience in implementing Web services to understand how they were applying Web services, their motivation for adoption, and their experience to date as well as their detailed as well as their further plans for 2003.

The responses came from a number of industry sectors, Systems Integrators 27%, Government 1%, Telecoms 13%, Finance 26%, Manufacturing / Process 10%, Travel/Transport 7%, Retail/Logistics 7%.

From the survey result, we can understand the motivation for adopting Web services and the reasons for using Web services.
Motivation for Adopting Web Services

- Support of new business opportunity
- Business process reengineering or optimization
- Providing an information based product
- Device and platform independence
- Cost savings
- Infrastructure initiative
- Technology pilot

Reasons for Using Web Services

- Scenario not really deliverable without them
- More cost effective
- More practical than existing techniques
- Compliance with emerging standard
- Delivers more flexible solution for the business
- Delivers a more flexible solution for IT
- Strategic decision

However, we also find that the security is the highest barrier to wider adopting Web services.

Barriers to wider Adoption

- Security and related concerns
- Immaturity of platform technology
- Immaturity or lack of Web service standards
• Risks of 3rd party dependencies
• No perceived business need
• No perceived technical need
• Financial Constraints
• Skills shortage

Though there is nothing can ever be proven to 100% secure, we should make enough security to make Web services practical.

9.40 Securing Web Services

Though there is nothing can ever be proven to 100% secure, we should make enough security to make Web Services practical.

a) Security and Web Services

Security is important for any distributed computing environment. But, security is even more important for Web services due to the following reasons:

1. The boundary of interaction between communicating partners is expected to expand from intranets to the Internet. Obviously, security problem is much critical in Internet because Internet communication is much less protected than intranet communication.

2. There will have more anonymous to access the web services since communicating partners are more likely to interact with each other without establishing a business or human relationship first. This means that all security requirements such as authentication, access control, non-repudiation, data integrity, and privacy must be addressed by the underlying security technology.
3. More and more interactions are expected to occur from programs to programs rather than from humans to programs. Therefore, the interaction between communicating partners using Web services is anticipated to be more dynamic and instantaneous.

4. As more and more business functions are exposed as Web services, the number of participants in a Web services environment will be larger than what we have seen in other environments.

b) Security Considerations

Security is about protecting assets. In the Web services context data and computational services are assets under consideration. The following security considerations must be considered as part of a comprehensive security framework

1. Identification - The party accessing the resource is able to identify itself to the system.

2. Authentication - the proven identification of users in a computer system.

3. Authorization - There exists a set of transactions the authenticated party is allowed to perform.

4. Integrity - the prevention of unauthorized modification of data.

5. Confidentiality - the prevention of unauthorized disclosure of data.

6. Accountability - the provision of activity logs recording all user activity.
7. Non-repudiation - Both parties are able to provide legal proof to a third party that the sender did send the information, and the receiver received the identical information.

c) Web Services Security Schemes

Web services security language can be defined into two types: computer security and communications security.

1. Computer security is a node-oriented security focus and it is essentially access control within a computer system. A permission rule expresses restrictions on the usage at the server side and a client can execute the operations only if the permission rule is allowed.

2. Communications security is a connection-oriented security focus and it is about providing a secure logical connection between two agents. A requirements rule expresses the necessary security-relevant preparations for the use of a service, or security measures needed after the service execution. The activities authenticates and encrypts are associated with authentication and confidentiality, respectively.

And currently, the most common security scheme available for today’s Web service is SSL (Secure Socket Layer), which is typically used with HTTP. It provides authentication, confidentiality, and message integrity. However, SSL is designed to provide point-to-point security, which falls short for Web services because:

1. We need end-to-end security, where multiple intermediary nodes could exist between the two endpoints.
2. SSL secures communication at transport level rather than at message level. As a result, messages are protected only while in transit on the wire.

3. HTTPS in its current form does not support non-repudiation well. Non-repudiation is critical for business Web services.

4. SSL does not provide element-wise signing and encryption.

In order to complement SSL, the technology industry has been working on various XML-based security schemes to provide comprehensive and unified security schemes for Web services. These schemes include:

**1. XML digital signature** - XML digital signature provides authentication, data integrity and non-repudiation. It is to develop XML syntax for representing digital signatures over any data type. The XML digital signature specification also defines procedures for computing and verifying such signatures. Another important area that XML digital signature addresses is the canonicalization of XML documents. Canonicalization enables the generation of the identical message digest and thus identical digital signatures for XML documents that are syntactically equivalent but different in appearance. XML digital signature provides a flexible means of signing and supports diverse sets of Internet transaction models.

**2. XML encryption** - Its goal is to develop XML syntax for representing encrypted data and to establish procedures for encrypting and decrypting such data. (Unlike SSL, with XML encryption, you can encrypt only the data that needs to be encrypted.)
3. **XKMS (XML Key Management Specification)** - XKMS consists of two parts: XKISS (XML Key Information Service Specification) and XKRSS (XML Key Registration Service Specification). XKISS defines a protocol for resolving or validating public keys contained in signed and encrypted XML documents, while XKRSS defines a protocol for public key registration, revocation, and recovery. The key aspect of XKMS is that it serves as a protocol specification between and XKMS client and an XKMS server in which the XKMS server provides trust services to its clients by performing various PKI operations.

4. **XACML (Extensible Access Control markup Language)** - Its goal is to standardize access control language in XML syntax.

5. **SAML (Secure Assertion Markup Language)** - It’s to outline a standard XML framework for exchanging authentication and authorization information. As a framework, it deals with three things. First, it defines syntax and semantics of XML-encoded assertion messages. Second, it defines request and response protocols between requesting and asserting parties for exchanging security information. Third, it defines rules for using assertions with standard transport and message frameworks.

6. **WS- Security (Web Services Security)** - It defines a set of SOAP header extensions for end-to-end SOAP messaging. Security. It supports message integrity and confidentiality by allowing communicating partners to exchange signed and encrypted messages in a Web services environment.

7. **ebXML Message Service** - The ebXML initiative is a set of next-generation XML-based standards enabling electronic business
transactions via the Internet. One of the ebXML standards is ebXML Message Service, which defines how to securely and reliably send and receive SOAP messages.

The SAML assertions can be digitally signed using XML digital signature. The same assertions can be encrypted using XML Encryption to ensure privacy. The public key used for digital signing and encryption can be validated and registered via XKMS. As for XACML, an SAML asserting party could use it to define an access control policy as a basis for handling SAML-based assertion requests.

Take an example: When a client placing an order, she uses XML digital signature and encryption to digitally sign and encrypt the purchase order XML document. She then sends the document to her supplier using SOAP, whose header structure is defined either in the WS- Security or ebXML Message Service standard. The document’s receiver then could use XKMS to look up and validate the public key. Once the key is determined trustworthy, the receiver then validates and decrypts the purchase order. Finally, the receiver checks a policy server for authorization by sending and receiving SAML requests and responses. The policy server might maintain the access control policy information in XACML.

### 9.41 Identity Management

#### Overview of Identity Management

Identity is a set of attributes that describes a profile of an individual, business organization, or software entity. E-business initiatives - such as enterprise, B2B and B2C applications typically reach throughout and
beyond and enterprise, requiring users to move across networks, applications, and security domains. If there is lack of a well-integrated and interoperable identity management architecture, this makes managing Web properties, applications, identities, and policies non-scalable, and effectively prohibits the interaction of identities across applications or Web services. To be effective, this movement must be transparent to the user. Consider what’s involved in this: a single identity with one registration process and one login procedure.

There are two identity management architectures, centralized model and federated model.

1. Centralized model - In the centralized model, a single operator performs authentication and authorization by owning and controlling all the identity information. It makes the constructing and managing the identity network much easier. However, there is the dangerous potential for the single operator becoming a tollgate for all transactions over the Internet, the single operator could represent a single point of security failure or hacker attack and a single operator can take away the most important business asset (customer identity and profile information).

2. Federated model - In the federated model, both authentication and authorization tasks are distributed among federated communities. It’s to create an open standard for identity, authentication and authorization, which will lower e-commerce
cost and accelerate organizations’ commercial opportunities, while at the same time increasing customer satisfaction. Furthermore, organizations can maintain their own customer data while sharing identity data with partners based on their business objectives and customer preferences.

As centralized model cannot effectively manage or control an e-business initiative from beginning to end, especially when multiple partners are involved. That’s why organizations are turning to federated identity management. The appeal of federation is that they are intended to allow user to seamlessly traverse different sites within a given federation. Federations provide a simple and flexible mechanism to identify and validate users from partner organizations and provide them with seamless access to Web sites within that trusted federation without requiring re-authentication. In addition, Federation standards also deal with the matter of providing trusted attributes about users allowing for privacy and business-specific rules.

Perhaps nothing is more important to the future of Web services than federated identity, the ability to securely establish a person’s or a service’s identity and to share that identity across domains and enterprises.

Establishing a unique identity is the key component in being able to take advantage of services and applications beyond a domain or firewall, which is the ultimate promise of Web services.

**9.42 What is federated identity?**

Federated identity is a way to establish someone’s identity across companies, domains and applications. The idea is that once that identity
is established in one place, it can be carried across to other Web services. So complex transactions and applications can be used, without the person having to log into separate applications or services and information about that person can be carried across as well.

**9.43 Federated Identity Standards**

A number of different standards apply to federated identity, but there are three primary ones:

1. **SAML** - This standard concerns itself with authentication and authorization. The current version is 1.1, but a major new version, 2.0, is due out this summer, and integrates more closely with the Liberty Alliance federated identity standards.

2. **WS- Federation (Web Services Federation Language)** - This is an attempt to build an overriding federated identity standard, to work in concert with SAML and other security standards. Prime movers behind it are BEA, IBM, Microsoft, RSA Security and Veri Sign.

3. **Liberty Alliance** - This is a set of standards for federated identity overseen by a group of companies called the Liberty Alliance, of which Sun was a prime mover and founder.

In this way we understand in E-commerce Securing Web Services is the major part in Data Mining.