What is your Strategy to Extract Intelligence from your Data?

Data Mining: a powerful tool for decision making

Hassan Ghaziri, Ph.D. Operation & Information Systems, American University of Beirut, Olayan School of Business

Introduction

Today, as we are confronted with increasingly large volumes of data, Data Mining (DM) is more than ever a critical component in the decision making process. Data mining is the process of extracting useful patterns and regularities from large bodies of data (i.e. databases which are too large to be analyzed by hand).

Data Mining is an umbrella term that covers the tools and techniques used in the analysis of large databases. Every corporation has data. Data is the raw material feeding the corporate body. Datamining is the process that is transforming raw food into blood that is nourishing the various functions. In this paper, the basic concepts of Data Mining will be introduced. The objective is to provide the reader with a practical understanding of this new tool. The purpose of a DM solution will be understood from a business point of view. Thus, DM will be put into the context of the corporate environment, and then the way DM fits into the overall corporate business model will be discussed.

What is Data Mining?

Gaining insights to improve business functions remains a major concern of any manager or decision-maker. The past two decades have seen a wide diffusion of information technologies in all business activities. The intensive use of new electronic devices such as point of sales, remote-sensing devices, ATMs, and the Internet has contributed to the explosion of available data. It is similar to the gold rush: a new continent was available to all kinds of miners. Suddenly managers and decision-makers became conscious of this continent of data. It was natural to think about the best way to benefit from it. This is where Data Mining comes in.

Data Mining is becoming a fundamental component of the global business infrastructure that assists firms in the decision-making process and helps them capture the complexity of the new economy. It is estimated that the amount of information in the world doubles every 20 months. What are we supposed to do with this flood of raw data? Clearly little of it will be seen by human eyes."
With Data Mining, questions that can make a difference could be answered. A report generated by RDBMS can answer such questions as the following:

- Which region sold the most last month?
- Which salesman sold the most during the last year?
- How much each channel was able to sell?

These questions can be answered by a transactional oriented system. However, this system is not designed to answer questions like, 'Why are we getting these results?' Until now, answering such question has been the exclusivity of human experts. So what has changed? And why do we now need to change our strategy to reach the right answer to the question 'Why'? In fact, two major factors have converged and led to the emergence of new technologies centered around Data Mining. The first factor is the amount of data collected from various sources including the functional databases (e.g. the accounting system, the sales information system), the web, and the intranets. This huge amount of data, collected and stored in various data bases, constitutes a challenge for decision-makers to extract from this flood of raw data useful information that could be used successfully in the decision making process. It is also a challenge for them to take advantage of this goldmine of data that could be used to gain competitive advantage and improve the efficiency of the firm. The second factor is the new developments in the area of analytical methods that are able to provide inference and to learn efficiently from data at a reasonable computational cost. Answering the question 'Why?' means that one needs to identify the factors causing the considered result. The manager can use his know-how skills and expertise in making decision and drawing sound conclusions. The objective of Data Mining is to optimize the use of available data and reduce the risk of making wrong decisions. Analytical methods allow managers to draw hypotheses and test them, predict events and adapt decisions accordingly, find unexpected patterns and exploit them, discover hidden associations and undertake appropriate actions. Without analytical methods, there is no effective analysis. Without analysis, there is no business intelligence. Without business intelligence, there is no hope to assimilate gigabytes of data and consistently make decisions that keep one's business ahead of competition.

Data Mining in the Corporate Environment

DM is a component of the business intelligence architecture. Business intelligence makes it possible to take action that will ultimately result in measurable benefits. Traditionally, it includes query and reporting, online analytical processing (OLAP), data visualization, and data analysis. DM complements these tools and provides a global approach that integrates the conventional tools in a whole process that leads to new insights and actionable knowledge for managers and decision-makers. It allows users to access information from different sources through client server or web-based query systems to visualize models and interact with the results of those queries to gain new business knowledge.

Figure 1 shows the framework for Decision Support Systems in the context of the new corporate business model.

For years companies recognized that that the volume of data generated by operations (On-line Transaction Processing System (OLTP), Point of Service System (POS), ATM system, the Internet etc) could be a volatile asset. The challenge is to make data available anytime, anywhere for decision-making process. Data is now used in a proactive way that provides value for the firm. This approach has led to the development of the Decision Support Systems (DSS) and the Executive Information Systems (EIS) which go beyond the static reporting concepts. The approach uses computing power and graphical interfaces to manipulate, "slice and dice" data easily and quickly at the convenience of the corporate.

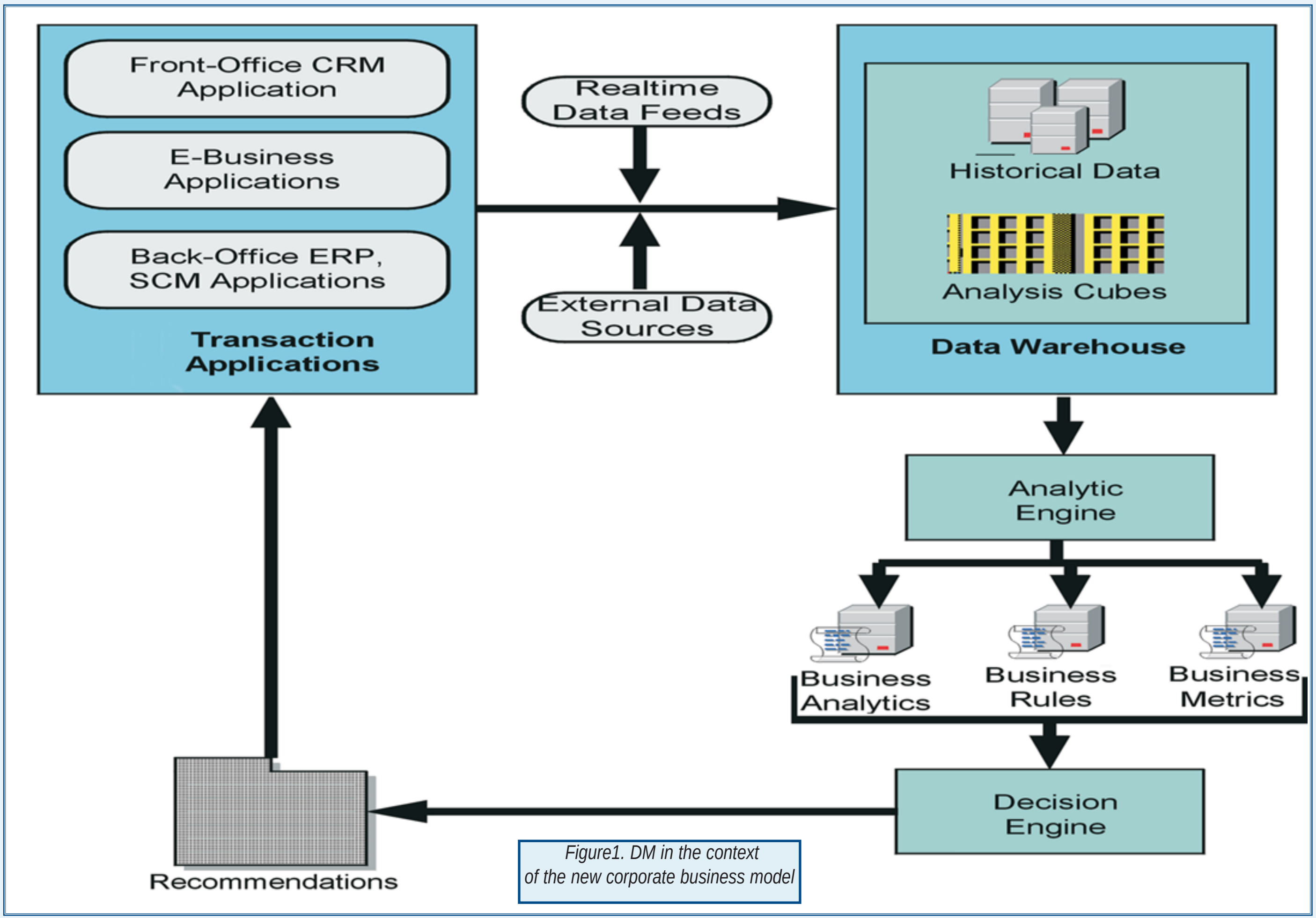

Figure1. DM in the context of the new corporate business model

What are the roots of Data Mining?

In this approach, the stress is on the ability to show data along several dimensions. The manager should be able to drill down into the ultimate detail of a transaction and zoom up for a general view. This ability is not possible without Data Warehousing, which collects data from various operational systems, organizes them, transforms them and makes them available for further investigation, and On-line Analytical Processing tools (OLAP), that require a Data Warehouse to facilitate the preparation of data multi-dimensional analysis.

Evolutionary Step	Business Question	Enabling Technologies	Characteristics
Operational Transaction Systems (1960s)	"What was my total turnover last year?"	Computers, tapes, disks	Retrospective, static data delivery
Client Server Applications (1980s)	"How much the New York branch was able to sell?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What was the amount of sales in New England last year? Drill down to Boston."	On-line analytic processing (OLAP), data warehouses	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next year? Why?"	Advanced algorithms, large size databases	Prospective, proactive information delivery

Table 1. Steps in the Evolution of Data Mining

Most DM applications focus on discovering the unknown from databases. Therefore, practitioners tend to adopt a bottom-up approach in which no hypothesis is previously formulated.

This approach could be directed or undirected. In the directed approach, also called the supervised approach, the data miner has some idea about the data he is exploring. In the undirected approach, also called the unsupervised approach, the data miner has no idea what he is looking for.

Let us consider a sales management. A hypothesis could be, "Customer churning earns less than $10,000K." A predictive problem could be, "Which customers might leave." A descriptive problem could be, "If a customer is buying service A, what are the chances that he buys service B".

Let us categorize the different Data Mining tasks according to the business problems. The following table shows which techniques to be used for which task in which business situation. In this table we are linking the Data Mining tasks with Data Mining techniques. Those techniques are out of the scope of this paper. A more detailed description can be found in the reference [1].

Business Applications	Customer Management	Market Management	Risk Management
	Customer relation	Market basket analysis	Fraud detection
	Target marketing	Market segmentation	Bankruptcy prediction
	Customized catalog	Cross selling	

DM Approaches	Prediction	Description	Verification
	Find future values, trends and behaviors	Discover new facts and relationships	Test hypothesis

DM Task	Classification	Link Analysis	Clustering	Regression

DM Task	Neural Networks	Classification Trees	Association Rules

Table 2. Data Mining management applications with the corresponding tasks and techniques

DM Process

Many methodologies for conducting a DM project have been proposed. Among the most used is the CRISP-DM methodology.

The methodology proposed is a seven-step process: <u>business understanding, data understanding, data preparation, data modeling, analysis of the results, knowledge assimilation and deployment, and evaluation</u>. The different tasks involved in each step have to be described in detail; the appropriate involvement in each step of various technical and business specialists also need to be identified.

Why a methodology?

It enables enterprises to produce consistent and meaningful results. It helps to define and enforce a process. It provides a means to apply a proven approach that has been refined from previous experience, which in turn helps reduce the risk and complexity surrounding these projects .

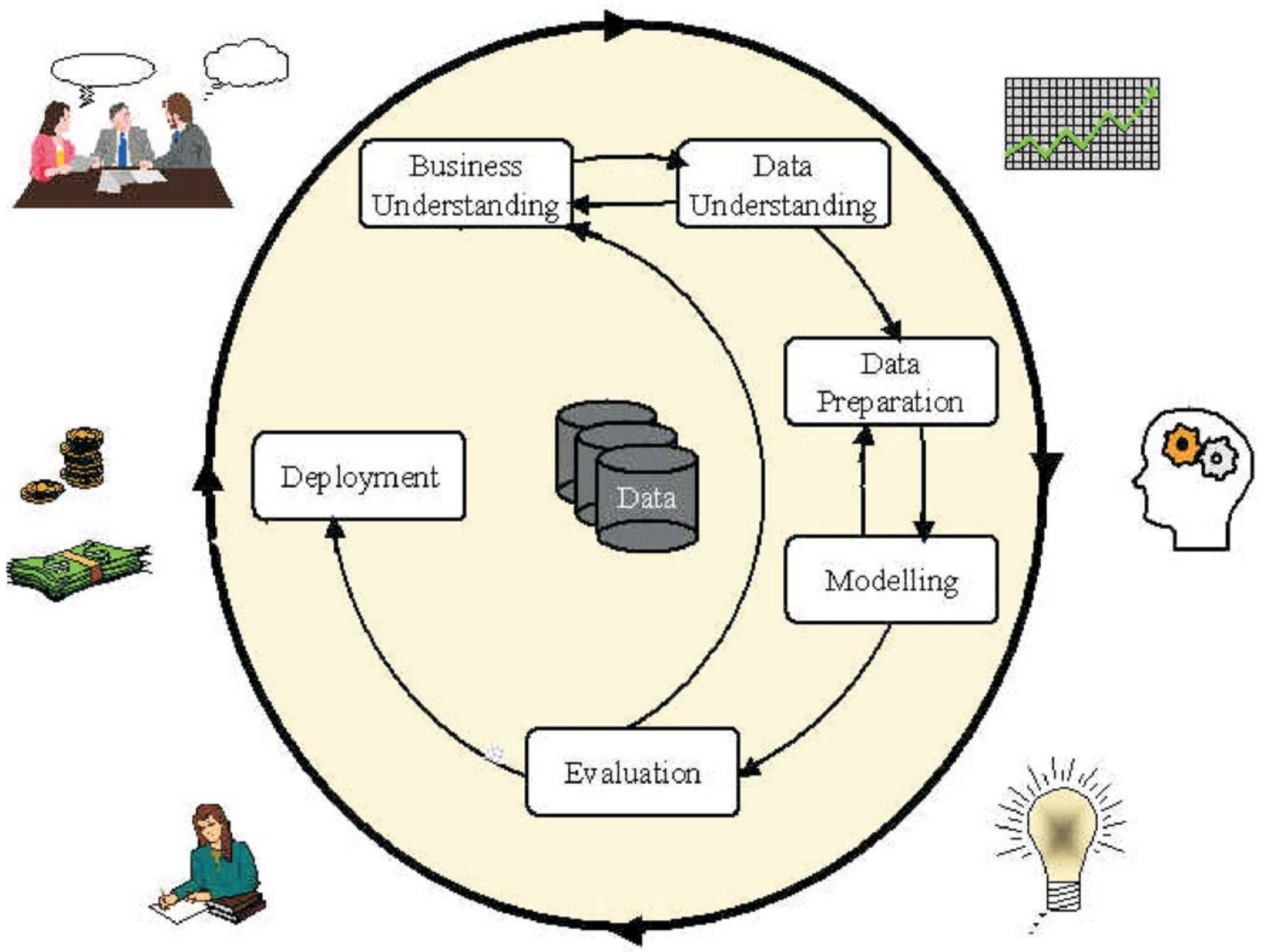

Figure 2: Phases of the Current CRISP-DM Process Model

DM in Practice

DM can be used in two ways:

a) To improve the internal business functions.

b) To address the company's relationship with its customers.

Most efforts are being made to understand customer behavior and adjust a company's products and services accordingly. Over 70 percent of DM projects are currently conducted by marketing departments. It can be explained by the evolution of a company's culture. Companies today are customer centric and not service-for-product centric. They are driven by the necessity of getting closer to their customers and gaining their loyalty. Customers are extremely volatile; they have access to a huge variety of products and brands through the Internet, regular mail and other means. In this open environment, building customer loyalty becomes very challenging for most companies. New tools are needed in such situation. DM is among these tools that could provide more insights about customers and ways to increase rates.

Direct marketing is one of the most popular applications of Data Mining because the results can be readily applied to obtain a better campaign response and a higher return on investment. For example, a trained statistician or consultant can develop a response model using data from a past direct marketing campaign to predict those most likely to respond to the next campaign. By contacting only those most likely to respond, that marketer can substantially increase the percentage of responses and generate higher profits.

Direct marketing campaigns extend beyond the retail merchandising industry to other consumer and business services. The convergence of direct marketing and Internet-based electronic commerce provides a rich resource for future applications of data mining in this burgeoning arena.

Customer churn is a significant problem in many industries. It is generally considered to be far more cost-effective to retain existing customers than to secure new ones. Thus, it is desirable to identify customers who are at risk of changing their service to a new provider and offer them incentives to retain their existing service.

Identifying customers, who are at risk of changing their service, and subsequently determining the appropriate actions to take that will encourage customers to retain their service, presents a significant challenge to organizations responsible for customer care and customer retention. It is often unclear exactly what characteristics dissatisfied customers' exhibit and, therefore, difficult to predict exactly which customers will not renew their service. In addition, even when it is known that a specific customer is dissatisfied, it is very difficult to determine what incentives can be offered that will encourage him or her to retain his or her service. Of course, it is desirable to offer these dissatisfied customers incentives in as cost-effective manner as possible.

Customer historical data frequently contains information that can be extremely valuable to retention specialists. Historical data holds usage patterns and other important customer characteristics that, when discovered, can be used to identify satisfied and dissatisfied customers. Combined with historical information, identifying which of those customers renewed their service, which did not renew their service, and what incentives were offered to both groups, a predictive model can be built to predict which customers will not renew their service as well as make recommendations as to the most effective incentives to offer them.

The key to satisfying customers, optimizing their experience, and thereby quite possibly their loyalty, lies in understanding their individual preferences. As Dorian Pyle so eloquently notes in the Foreword, *customers have become a critical resource*. Twenty-first century customer relationships are based on mass-customization rather than mass-production. What one customer may consider being "attentive" customer service, another may deem "oppressive". Being able to recognize the differences between these two types of customers and making effective use of such information remains a challenge, but a challenge that Data Mining techniques are uniquely well-suited to address.
There are many Data Mining applications that can enhance an organization's customer services. They include:

- Customer acquisition profiling
- Customer retention profiling
- Customer-centric selling
- Targeting market

The following are some successful application areas:
- A pharmaceutical company can analyze its recent sales force activity and its results to improve targeting of high-value physicians and determine which marketing activities will have the greatest impact in the next few months. The data needs to include competitor market activity as well as information about the local health-care systems. The results can be distributed to the sales force via a wide-area network that enables the representatives to review the recommendations from the perspective of the key attributes in the decision process. The ongoing, dynamic analysis of the data warehouse allows best practices from throughout the organization to be applied in specific sales situations.

- A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product. Using a small mailing test, the attributes of customers with an affinity for the product can be identified. Recent projects have indicated more than a 20-fold decrease in costs for targeted mailing campaigns over conventional approaches.

- A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services. Using Data Mining to analyze its own customer experience, this company can build a unique segmentation identifying the attributes of high-value prospects. Applying this segmentation to a general business database, such as those provided by Dun & Bradstreet, can yield a prioritized list of prospects by region.

- A large consumer package goods company can apply Data Mining to improve its sales process to retailers. Data from consumer panels, shipments, and competitor activity can be applied to understand the reasons for brand and store switching. Through this analysis, the manufacturer can select promotional strategies that best reach their target customer segments.

- In the supermarket industry, Data Mining has been implemented in the form of personalized recommendation systems. Personalization involves filtering a list of consumer products according to an individual customer's profile that exists on Customer Purchase Database. The recommendation system then uses Market Basket Analysis and Clustering in order to match customers to new products. The final results are either directly sent to the customer (e.g. via Palm) or are used to redesign the layout of the supermarket.

Resources
[1] Awad E., Ghaziri H., Knowledge Management, Printence-Hall, 2004
[2] Groth R., Data Mining: Building Competitive Advantage. Prentice Hall PTR. 2000.
[3] Morten T. Hansen, Nitin Nohria, and Thomas Tierney "What's Your Strategy for Managing Knowledge?" Harvard Business Review, 77,2,1999,106-16. (Available: http://www.hbsp.harvard.edu/products/hbr/ marapr99/99206.html
[4] www.spss.com Crisp methodology
[5] Witten I., Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implantations. Morgan Kaufmann. 2000.