# Knowledge Discovery in Databases for Competitive Advantage

Mark Gilchrist
Nealanders International Inc
Toronto, Ontario, Canada

Deana Lehmann Mooers
Human Resources
Northern Transportation Company Ltd
Edmonton, Alberta, Canada
E-mail: dmooers@ntcl.com

Glenn Skrubbeltrang
Department of Accounting
Faculty of Business, Brock University
500 Glenridge, St. Catharines, ON, L2S 3A1 Canada
E-mail: gskrubbeltrang@brocku.ca

Francine Vachon (Corresponding author)
Department of Finance, Operations and Information Systems
Faculty of Business, Brock University
500 Glenridge, St. Catharines, ON, L2S 3A1 Canada
E-mail: fvachon@brocku.ca

Received: January 29, 2012 Accepted: February 20, 2012 Published: April 15, 2012
doi: 10.5430/jms.v3n2p2 URL: http://dx.doi.org/10.5430/jms.v3n2p2

**Abstract**

In today's increasingly competitive business world, organizations are using ICT to advance their business strategies and increase their competitive advantage. One technological element that is growing in popularity is knowledge discovery in databases (KDD). In this paper, we propose an analytic framework which is applied to two cases concerning KDD. The first case presents an organization at the analysis stage of a KDD project. The second one shows how a multinational company leverages its databases by mining data to discover new knowledge.

**Keywords:** Knowledge Discovery, Data Mining, Databases, Strategy, Competitive Advantage

## 1. Introduction

Rapid changes in the global business context and in information and communication technologies (ICT) have placed a great deal of pressure on business organizations. Strategy formulation requires an accurate picture of the competitive environment to quickly adapt and to secure new opportunities (Ofori & Atiogbe 2012). Unfortunately, the increased capacity of databases, and other technological advances, has resulted in companies being "data rich but information poor" (Hanna 2004b, p. 31). This huge quantity of data is impossible to comprehend without sophisticated and powerful IT. A survey found that databases were the top-ranked knowledge management tool (Mundra et al. 2011, p. 18). Antonova et al. (2011) found that 60% of 357 managers used databases as a tool for knowledge sharing. For business organizations, databases represent a critical knowledge management tool that must be leveraged to yield maximum strategic returns.

"Nowadays, having knowledge and utilizing it in organizations has become a process which can lead them to an advantage over competition" (Sanayei & Sadidi 2011, p. 235). Using knowledge discovery in databases (KDD) is an excellent way for businesses to leverage databases. KDD has become an integral part of developing, maintaining, and furthering a company's business strategy. KDD is defined as the "non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Mariscal, Marban & Fernandez 2010, p. 142). This process is made up of many iterative steps, one of which is data mining (DM) (Crespo & Weber 2005). 'DM' refers specifically to the mathematical models and computing tools used to analyze data (Mariscal et al., 2010).

Most literature in the KDD and DM domain concerns the technological aspects of KDD. However, IS research demonstrates that managerial issues are just as important to the success of ICT projects. This article presents an analytical framework that links KDD to firm strategies and merges both managerial and technological perspectives.

This research had the following objectives:

a) Review the literature on KDD from a multidisciplinary perspective.

b) Define the KDD process using this literature review.

c) Present two cases, University A and Walmart, that illustrate the KDD process.

d) Develop an analytic framework integrating key aspects of this literature.

e) Use this analytic framework to analyze the two cases.

## 2. Evolution of Knowledge Discovery in Databases

The concept of KDD emerged in the late 1980s out of several founding disciplines: statistics, machine learning, database management, information science and visualization (Hanna 2004b, Meisel & Mattfeld 2010)). Throughout the 1970s and early 1980s, new data storage technologies yielded increasingly sophisticated relational databases (Han & Kamber 2006). DM and KDD resulted from this evolution of information technology's functionalities. Interest in KDD, both academic and industrial, has exploded in recent years, as the methods used to collect and store data have become simpler and less expensive, allowing organizations to support massive databases (Olafsson et al., 2008).

## 3. The Process of Discovering Knowledge in Databases

The KDD process uses data collected from databases inside as well as outside organizations. These data are analyzed from different angles and perspectives to discover relationships. The goal is to leverage the data asset to improve decision-making. KDD can be defined as the process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. This requires a large, well-integrated database and a good understanding of the business context within which the DM is to be applied (Sankar & Pabitra 2004, p. 7).

Data mining (DM) represents the IT tools used for KDD. Whereas statistics are used to analyze data collected to answer pre-defined questions, DM is concerned with the "secondary analysis of large amounts of data that are collected as a bi-product of other systems" (Meisel & Mattfeld 2010, p. 1).

One of the most common methodologies for processing organizational data is called the "Cross-Industry Standard Process for DM," (CRISP-DM 1.0: Step-by-step data mining guide, 2010). The CRISP-DM provides a framework within which organizations can structure DM (Olson & Delen 2008, p. 11); it is a neutral process model that can be used with any tool or application by any industry. However, the use of the CRISP-DM is not increasing, because this method defines "what to do" but not "how to do it" (Mariscal et al. 2010, p. 139).

Following the CRISP-DM, several methodologies and process models were developed. Mariscal et al. (2010) surveyed these to propose a Refined Data Mining (RDM) process. The RDM process comprises 7 processes (Table 1, Appendix 1 following the References).

KDD goes beyond IT applications. Orienting the search for knowledge is crucial. Without a clear orientation based on domain knowledge, the software could come up with spurious associations or solutions that do not apply to the problem at hand (Cao & Zhang 2007, Mariscal et al. 2010). From the beginning, the project team must include managers who know the organization's domain and strategies, as well as IT experts.

## 4. Data Mining

DM consists in applying specific algorithms to discover useful knowledge. A model is essentially a score-producing mathematical equation. An algorithm is a segment of computer code that performs a model's calculations. Models can be formulated by various models. A predictive model, or statistical model of future behaviour, is used to improve the odds of obtaining a favourable result. Predictive analytics comprises a number of predictors, i.e., the variable factors most

likely to influence future events. To create a predictive model, data is collected for relevant predictors, statistical models are formulated, predictions are made, and then the model is validated or revised as more data is collected (Henze 2008).

*4.1 Data Mining Tasks*

Organizations can use various types of predictive models to extract useful information from a database (Henze, 2008). Here are some examples:

4.1.1 Data Pre-Processing

Data Pre-Processing consists in converting the data into a form suitable for DM. Databases contain errors that must be corrected prior to mining. Au et al. (2010) have developed a way to automate irregularity detection in large databases. Pre-processing also includes data transformation from one form to a more useful form—for example, from a numeric attribute to a nominal attribute (Hall et al. 2009, p. 14). Once the data is pre-processed, one of the following operations can be performed.

4.1.2 Classification

Classification consists in "assigning cases into categories based on a predictable attribute" (Tang & McLennan 2005, p. 6). The algorithm must "identify among attributes which are the most predictive ones" (Hall et al. 2009, p.11). The values of these predictive attributes are used to classify new cases and predict their behaviour. For example, classification is used for predicting the loan payment behaviour of prospective borrowers (Han & Kamber 2006). A neural network includes algorithms that enable the system to learn, such as systems that "advise" airline seat bookings (Dunham 2003, p. 63).

4.1.3 Regression

Regression is similar to classification, except that "the predictable attribute is a continuous number" (Tang & McLennan 2005, p. 8). Linear regression, a form of regression analysis, creates a line that approximates the trend of a grouping of variables plotted on a graph (Dunham 2003, p. 6). For example, a retailer can use regression to examine the effectiveness of sales and promotions. This knowledge will help the retailer plan more effective promotional campaigns (Han & Kamber 2006).

4.1.4 Cluster Analysis

Cluster analysis "assigns the objects into k groups such that each object in a group is more similar to other objects in its group as compared to other objects in other groups" (Chaoji et al. 2009, p. 201). It is used extensively in science and in marketing, for example, in picture processing, mode identification (Tan et al. 2010) bioinformatics, privacy protection, and security applications (Chaoji et al. 2009). Le Blanc and Rucks (2009) used this technique to classify university alumni into six groups, according to their university donation behaviour. They were able to figure out which characteristics were shared by members of each cluster, and used this information to better direct fund-solicitation efforts.

4.1.5 Association Rules Analysis

Association rules analysis "returns a set of rules determining the likelihood of one or more events or attributes occurring if another event or attribute occurs first" (Hanna 2004a, p. 136). A marketing example is so-called "market-basket analysis," or "determining what products are purchased by customers and using these associations to cross-sell or up-sell" (Olson & Delen 2008 p.4). Association rules are used to decide which products should be discontinued in order to maximize profits, taking into account cross-selling considerations (Wong, Fu & Wang 2005). In the medical field, Hu (2010) perfected an algorithm to detect associations between patient attributes (such as tumour size) and breast cancer recurrences.

4.1.6 Sequence Analysis

Sequence analysis is used to detect patterns in a discrete series. An example of a sequence would be a Web click sequence consisting in a series of URLs (Tang & MacLennan 2005).

4.1.7 Forecasting

This technique uses a dataset of values taken over time, for example, amounts of daily sales. Forecasting takes into account the order of values, for example, seasonal variations in sales or revenues (Tang & McLennan 2005).

4.1.8 Deviation Analysis

Deviation Analysis is used to detect the exceptional cases that are very dissimilar to the others (Tang & McLennan 2005). Banks use this technique for fraud detection.

4.1.9 Data Visualization

Based on the idea that "a picture is worth a thousand words," visualization tools make complex data easier to understand and interpret. This is especially important for scientific applications (Tang & McLennan 2005). It is gaining in popularity in business applications.

4.1.10 Customized Models

Customized models use organizational information to construct a factor-weighted analysis that will be significant to a specific organization.

*4.2 DM System Solutions*

Organizations wanting to implement KDD select a DM system that meets their needs. In general, all DM systems must meet the following requirements:

a. The adopted solution must process data from a variety of database formats and platforms (Hall et al. 2009);

b. it must be dynamic, i.e., adaptable to changes in customer behaviour over time (Crespo & Weber 2005);

c. it should conform to established DM standards (Tang & MacLennan 2005);

d. it must be scalable, i.e., adjustable to meet increasing database sizes (Horwitz, 2010); and

e. the organization must be able to afford the selected system (Horwitz, 2010).

DM systems can be categorized as one of the following:

4.2.1 Custom-Developed Solution

A package is tailored to meet an organization's needs. Custom-developed systems are not accessible to competitors (Hays, 2004). However, such systems are expensive and the risk of project failure is greater because the software has not been benchmarked with other businesses.

4.2.2 Licensed Product:

The client buys a few developer licenses plus a number of licenses for the organization—per workstation, per server or an "enterprise agreement" (Horwitz 2010). Developers build applications adapted to organizational needs by using an Application Programming Interface (API) (Tang & MacLennan 2005). A licensed product such as SQL 2012 offers a number of advantages: compatibility with other products by the same vendor, compliance with standards, better-quality product for a lower cost, and vendor warranty and support (Horwitz 2010). Licenses products have some disadvantages as well. Clients do not own the product's source code. Therefore, clients are tied to the vendor's decisions concerning product evolution. They may be forced to pay for upgrades when the vendor stops supporting the existing version, whether or not the old version was satisfactory. Moreover, if a vendor goes out of business or merges with another vendor, clients may be forced to adopt different software and develop new applications. Finally, clients must "contend with complex rules and special exceptions.… They must identify licensing options and choose between them, and assure license compliance" (Horwitz 2010, p. 3). Both underestimating and overestimating license requirements can prove very costly. Nevertheless, a ready-to-use licensed product can be a cost-effective solution for a small business.

4.2.3 Open-Source Software (OSS)

Open-Source Software (OSS) is "released under a license approved by the OSI [Open Source Initiative]. … OSS development is different from traditional software in that it frequently depends on volunteers coordinating their efforts, ... and the end product is often provided for free" (Stewart, Ammeter, & Maruping 2006, pp. 127-128). An example of a widely accepted DM OSS is Waikato Environment for Knowledge Analysis (WEKA) (Hall et al. 2009). WEKA "is an organized collection of state-of-the-art machine learning algorithms and data processing tools" (Frank et al. 2005). Organizations using OSS DM applications find many advantages. They encounter lower (or no) licensing costs. They own the source code, freeing them from vendors' decisions about product evolution. Although they cannot count on vendor support, OSS developers can find support online in a community of volunteers. The most successful OSSs are sponsored by either businesses or not-for-profit organizations (Stewart et al., 2006). For example, WEKA is sponsored by the University of Waikato, New Zealand, and the open-source BI software company Pentaho (http://www.cs.waikato.ac.nz/ml/WEKA/). Adoption of OSS has some disadvantages. A business interested in developing OSS applications must either secure its own team of experts or outsource application development to a specialized service provider. Vendor support or warranty are absent. Selecting a popular OSS such as WEKA mitigates these risks.

*4.3 Requirements for Data Storage and Processing*

As well as DM software, KDD requires capacity for data storage and processing. Two solutions are presented here:

4.3.1 Full provisioning

All the capacity of an external disk is provisioned to the DM application (Chang 2011). This confers some advantages with respect to data security and confidentiality (only one disk to protect and back-up) (Armbrust et al. 2009). It greatly simplifies licensing agreements. However, on average, storage is underutilized (Armbrust et al. 2009, Chang 2011, Nurmi et al. 2009). "Storage utilization rates achieved by most companies is 40% or lower" (Chang 2011). However, during peak periods, utilization may exceed existing capacity (Armbrust et al. 2009, Nurmi et al. 2009). Full provisioning forces users to anticipate peak utilization. This leads to unused processing and storage capacities, and additional costs, such as power consumption and building amortization (Armbrust et al. 2009, Chang 2011, Nurmi et al. 2009).

4.3.2 Cloud computing

A datacenter hardware and software constitutes a cloud. Cloud computing refers to a pay-per-use service or utility computing (Armbrust et al. 2009). It has several advantages. Because customers can upscale or downscale computing resources as needed, this eliminates the problem of over-provisioning. It allows users to start on a small scale and increase capacity as needed. It allows them to pay only for the resources they actually use (Armbrust et al. 2009). This affords a special opportunity for organization to temporarily increase processing capacity devoted to DM applications that perform complex operations. These calculations can be executed much faster through parallel processing. However, cloud computing is not a panacea. Armbrust et al. (2009) note some obstacles. The most important is the problem of maintaining data confidentiality and auditability. The first issue can be solved with IT such as encryption. The problem about auditability is more difficult, as data may be stored in a foreign country and subject to different laws and standards. Armbrust et al. (2009) also note possible difficulties with data transfer bottlenecks.

This overview of the KDD process and DM tools is followed by the KDD analytic framework.

**5. The KDD Project Analytical Framework**

The proposed analytical framework merges managerial and technological perspectives as follows:

*5.1 Strategy*

KDD is only beneficial in as much as it allows a firm to support its business strategy (Hirji 2001). The Porter Model (as cited by Allen et al. 2008) is a well-known model used to classify business strategies to gain a competitive advantage in three categories:

a) Cost Leadership is aimed at reducing costs and charging customers lower prices than those of other competitors.

b) Product Differentiation concentrates on offering different or unique products that set the firm apart from the competition.

c) Focus or Niche Strategy "targets a specific, often narrow segment of the market" (Allen et al. 2008, p. 40).

To succeed, there must be an interconnection between IT strategy and business strategy (Rivard et al. 2004). This applies to KDD as well.

*5.2 KDD Objectives*

What are the objectives of a KDD process? Do these objectives support the business strategy in achieving business advantage? (Hirji 2001) Both CRISP-DM (3.) (Note 1) and the RDM Process (Table 1, Appendix 1) mention the importance of understanding the business problem and objectives. KDD objectives must be achievable, concrete and realistic (Hirji 2001, Mariscal et al. 2010).

*5.3 Degree of Advancement of KDD Process*

Mariscal et al. (2010) identified 7 processes, further subdivided into 17 sub processes (Table 1). By identifying which processes were completed and which are in progress, a researcher will learn more about the advancement of a KDD project.

*5.4 KDD Financial Feasibility*

Like all other investments, the scope of a KDD system depends on the availability of financial resources.

*5.5 Organizational DM Expertise*

Lack of familiarity with a new IT increases a project's riskiness (Dennis, Wixom & Roth 2012). Availability of internal expertise is a key component of the project life cycle selection (Table 1).

*5.6 DM Tasks*

Which DM tasks (4.1) are most likely to achieve KDD objectives and gain a competitive advantage?

*5.7 DM System Solutions*

Which DM system solution (section 4.2) best meets organizational requirements?

*5.8 Data Processing and Storage*

Which data processing and storage solution (section 4.3) best meets organizational requirements?

**6. The KDD Process at Two Existing Organizations**

High learning organizations (HLO) leverage knowledge to enhance competitive advantage (Mishra & Bhaskar 2011). The analytical framework was tested on two HLOs: University A and Walmart. The first case concerns an institution at the analysis stage, just starting the KDD process,. The second presents a giant corporation that has implemented KDD successfully—Walmart. These two cases were selected for the following reasons.

University A had to increase alumni donations to achieve its strategic goals. Until then, it had yet used ICT or databases to support its business strategy. This case illustrates the critical analysis process of a KDD project. Conversely, Walmart has historically been a company that has pioneered database applications, and is considered a mature organization in terms of the KDD process. The Walmart example illustrates a soundly conceptualized and developed KDD system as well as the benefits that can ensue. It seemed logical to assess two organizations that were at diametrically opposite ends of the KDD process.

*6.1 The Advancement Office at University A*

Section 6 will investigate how KDD—or, more precisely, an extension of DM termed 'predictive modeling'—can be used to supplement and support organizational strategy at University A (UA), as well as at UA's Advancement Office (UAAO). The institution's name is withheld.

As a diverse and inclusive family, UA contributes positively to the country and beyond through its imagination and innovation (UA's Mission Statement). UA was founded 45 years ago as a small regional institution with 125 students. Today it is a well-known university with an enrolment of 20,000. Despite its impressive progress, UA remains a small player, compared to other universities. UA is working to develop a specific "brand image"—one that highlights its innovative research, academic programs, as well as student activities—to distinguish itself from other institutions.

The 2008 recession adversely affected not only businesses and individual consumers, but higher learning institutions as well. Universities and colleges face cutbacks in private and corporate donations and some government grants. At many schools in Canada, including UA, alumni have not established a tradition of educational donations, such as can be found in the United States; such institutions must therefore compete for dollars from non-government sources.

UA's Advancement Office (UAAO) serves UA by fostering the commitment and generosity of a growing circle of alumni, friends, and supporters. UAAO staff members conduct extensive research to ensure that the right people are asked for the right donation. UAAO is committed to coordinating the efforts of the community to increase funds raised for UA. In order to determine where best to allocate its scarce resources, UAAO is focusing investment in staffing, infrastructure, and reengineering its processes.

The fundraising environment is increasingly competitive. Many universities are looking for new ways to locate and successfully solicit every possible donor. When launching a campaign, they do not want to waste donation dollars soliciting either already committed donors or people unlikely to donate (Le Blanc & Rucks, 2009). Already lucratively applied in business, predictive modeling has been making its way into non-profit and fundraising organizations. Could UA now or in the future benefit from a predictive modeling exercise? If so, what are the requirements, both in the database and in the expertise needed, for maximizing UA's return on investment?

As part of their analysis process, UAAO employees determine whether predictive modeling can serve UA's organizational strategy for raising funds. They identified organizations similar to UA that have benefited from predictive modeling. Syracuse University Library (SUL) has seen positive initial results that stem from using a database of potential donors to increase the success of fundraising activities (Griffen, 2005). By determining what information was available in the database, and who could best facilitate the extraction, manipulation, and analysis of the information,

SUL created a collaborative work that identified donors across many different university units. Jacksonville University used customized statistical modeling to identify its best donor prospects before launching a capital campaign. The model identified about 50 previously untapped prospects with a high propensity to make major gifts, paving the way for a special appeal. The Texas A&M Foundation used a combination of prescriptive and customized statistical modeling to better focus an appeal for charitable gift annuities. The Foundation screened 62,000 prospects and identified one-third of them as good candidates for this targeted campaign (Henze 2008).

The use of predictive models has thus proven successful in strengthening fundraising campaigns. What UAAO needs to determine is whether or not it has the database requirements to justify investing in a predictive modeling process. Since its inception, UA has awarded degrees to 70,000 individuals. However, not all alumni become donors.

What information does UA require to meet the end that the UAAO is trying to achieve? Is UA trying to understand what motivates people to donate? What motivates people to donate more than once? Who are the best prospects for UA's upcoming campaigns? The answer to all these questions can be found in the relationships that exist in both internal and external databases. By using donor behaviour as criteria, UA can identify industry benchmarks to determine where UA should be to effectively use a predictive modeling exercise. In this sense, UA adopts an external perspective to examine itself, as one institution among a group of peers. The selection of criteria can be an intuitive process which should include the internal stakeholders with domain knowledge of fund raising at UA. However, such research should also include indicators of giving trends and capabilities that other organizations have researched. Table 2 lists some of the preliminary criteria that would be part of the benchmarking process (Appendix1). Assigning weights to these criteria will allow the development of a model to assess potential donors. Therefore, it is crucial for UA to determine whether or not the list of preliminary criteria is exhaustive and then determining whether or not UA's database contains these criteria. This is important because finding donors that are not yet giving can have dramatic effects over time. Because donation patterns tend to conform to Pareto's Law, whereby a minority of people donate the majority of gifts, finding a new donor can lead to larger benefits in the future (Clotfelter 2001). The data is going to determine the effectiveness of any type of predictive modeling exercise.

After speaking with industry consultants (Henze 2008), two key components must be supplied before any predictive modeling exercise can be effective. First is the contact information. Many insights can be gleaned based on where people are located, for example, determining whether the donor lives in a well-to-do neighbourhood. In order to assess the area where a person lives, the postal code must first be accurate. UA needs to identify the population of alumni and donors that are traced to accurate addresses. This would allow UA to separate those who moved since graduation and spend time determining accurate addresses. The second component is the giving history in the past 12 to 36 months. This will enable the predictive model to find the donors who will make up the concentration as stated in Pareto's Law (Le Blanc & Ruck, 2009). It will also help determine the donors who are reaching a point of critical mass, a point where a more intimate approach to engaging the donor will bring about an increase in the size of donation.

UAAO needs to make decisions early in the process concerning project life cycle. Several companies offer proprietary software suites that provide excellent DM capabilities. Adopting a software suite from the same provider as the vendor of UA database system would insure file compatibility between the DM algorithms and the existing databases. However, UA's databases are fragmented among several applications, most of them legacy systems. Moreover, as a public learning institution, UA's financial resources are very limited. An OSS such as WEKA would be an option worth considering. Recently, UA had great success replacing its licensed course management system with the Sakai course management OSS (www.sakaiproject.org). This conversion yielded UA cost savings, greater flexibility and increased scalability. This positive experience and internal expertise developed with the Sakai project should encourage UA to consider DM OSS. Cloud computing is another option worth exploring, because of flexibility and low capital expenditures.

The expertise requirements will stem from the database evaluation. To create the best predictive model, the KDD experts will need a solid understanding of fundraising and of what the organization is looking for in its donors. Two basic approaches are either to look internally or to look externally. Depending on UA's timeline, KDD experts will most likely need to utilize an external organization to facilitate the predictive modeling exercise, for example, by supplying accurate donor addresses. These companies are already heavily invested in the basic assets required to run this type of database cleansing.

Essential to the KDD project, the data-gathering task is greatly facilitated by the implementation of UApeople.com, a private social network for UA alumni. It comprises four components:

a. LEARN: where alumni learn the latest UA news;

b. CONNECT: where alumni can stay in contact with or locate former classmates;

c. GIVE: a secure donation webpage;

d. GET BACK: where they have access to group discounts with UA business partners.

UApeople.org makes it beneficial for alumni to keep in touch with UA. By keeping their information up-to-date, users perform part of the data-cleansing task. UApeople.org allows UA to identify potential donors and their interests. Registration is voluntary: around twenty percent of new graduates join the network.

For the second data component, history of giving, data miners must ensure that all donor information is accurate, non-redundant, timely and secure. The donor component of UApeople.org facilitates this task. Cheque and cash donor information must be preserved as well. However, as a public institution, UA has a duty to ensure that all personal information is collected in accordance with privacy protection legislation.

*6.2 Walmart*

Walmart played a pioneering role in the business uses of DM. This company is recognized as "best-in-class" in leveraging their vast databases to support their strategy of low cost leadership. "The success or failure of strategies is linked, to a great measure, on how they are implemented" (Waweru 2011, p. 49). This case examines how KDD contributed to the success of the company and what lessons can be learned from their example by other organizations.

In the 1970s Walmart executives saw the advantage of using technology and data to expand their competitive advantage. By 1973, 22 stores were supplying data to the head office in Bentonville, AK on a nightly basis. This data helped Walmart better track its inventory levels and thus reduce the likelihood of overage or outage (Ortega 1998, p. 78). By 1979, Walmart owned a computer network that allowed data flow between stores, headquarters, and distribution centers. Information on sales in each department in every store, and inventory levels in stores and warehouses, was collected and analyzed (Ortega 1998, p. 100).

By the 1980s, Walmart led the retail industry in the use of technology and data to increase its competitive advantage. At the time, they decided to expand those capabilities to include vendor relationships, starting with their biggest vendor, Proctor and Gamble. Traditionally, the relationship between supplier and customer had been somewhat adversarial. However, executives at both companies now realized that their strategic goals were the same; hence, they were driven to develop a more collaborative relationship (Bianco 2006, p. 179). Developing synergies between the two companies in the areas of finance, logistics, and technology resulted in computer links between the two organizations that enabled them to implement a system of "continuous replenishment" that resulted in cost savings for both partners (Bianco 2006, p. 180).

In 1991, Walmart was ready to roll out a system born from the connection with Proctor and Gamble—Retail Link—to the rest of their suppliers. At an initial cost of $4 billion, Retail Link now provides all Walmart's suppliers with information, updated eight times per day, about every product they sell in a Walmart store. A vendor can now access a record of every sale of each item at each Walmart store during every hour of every day in the past two years (Fishman 2006, p. 75). The system analyzes mountains of data; it anticipates the demand for items based on sales history, it checks inventories and then, if needed, it automatically generates an order, which is then transmitted to the nearest distribution centre (Bianco 2006, p. 180). The efficiency of the system is demonstrated by the Bentonville Distribution Center. From here, Walmart is able to distribute goods at approximately the same rate at which they are being sold at the 127 stores serviced by that particular centre (Bianco 2006, p. 181). Access to information, however, does not come without a hefty price tag. The responsibility for keeping items in stock now falls to the supplier (Fishman 2006, p. 75), and 70% of the products in a given Walmart store will be sold to consumers before Walmart has paid the vendors of these products (Bianco 2006, p. 181). A direct result of this has been that Walmart can maintain lower inventory levels than any of its competitors (Bonacich & Hardie 2006), and at the same time spend less on distribution. As early as 1983, Walmart was spending less than 2¢ per sales dollar on distribution, compared to its competitors, who were spending an average of 5¢ per sales dollar (Ortega 1998, pp. 130-131).

Today Walmart is a leader in the retail industry, due in no small part to its use of data and technology as a core competency. Walmart guards it as a highly valued business asset. Very little reliance is placed on third-party software and all programming is done in house with no outsourcing at all (Sullivan 2004) (Fishman 2006, p. 237). Secrecy and security is a means to protect every ounce of competitive advantage. All suppliers that sell computer equipment or software to Walmart are bound by non-disclosure agreements (Hays 2004). Walmart also controls the information they provide to the outside as well. In 2001, Walmart stopped sharing data with the national consumer sales data clearinghouses that was passing sales data information on to Walmart's competitors (Hays 2004).

In 2010, Walmart ranked first in the world among listed companies with global sales of 421.8 billion USD, 9000 stores in 15 countries (Sharma 2011) and 2.1 million employees (www.walmartstores.com). The Information Systems Division

(ISD) of Walmart is committed to increasing the competitive advantage of the business: "Technology touches every part of our business every day, from data centers to self check-out lanes, satellite communications, handheld devices and electronic product codes. As the world's largest retailer and one of the most admired U.S. companies, Walmart is looking for the best and the brightest to help us blaze a trail in innovation and technology" (www.walmartstores.com/Careers).

The amount of data collected and used by Walmart is staggering. The David Glass Technology Centre in Bentonville, AK, just down the road from Walmart's corporate headquarters, houses one of the three largest private databases in the world (Fishman 2006, p. 138). This centralized database houses over 500 terabytes of information (Petrovic & Hamilton 2006, p. 133); the internet has less than half that amount of data (Hays 2004). The Walmart database collects information on each item scanned at every register in every Walmart store. Every day, Walmart uploads 20 million point-of-sale transactions to a centralized database (Wong et al. 2005). This information assists management in measuring the efficiency and productivity of each checkout person—they can tell, for example, that an average of 6-8 items are scanned and sold per minute at any given register (Fishman 2006, pp. 12-13). The organization uses this information to perform market-basket analysis. "By knowing customer behaviour based on past records, increased profits from cross-selling and other business strategies can be achieved" (Wong et al. 2005, p. 81). Not only does Walmart gather and analyze sales data, they also collect and study information in roughly 10,000 categories, including ethnicity demographics, weather patterns, and local sports team preferences. This data is then used to project sales trends and volumes for each individual Walmart store (Bianco 2006, p. 181). Using predictive technology Walmart has been able to start "predicting what is going to happen instead of waiting for it to happen" (Dillman, as cited by Hays, 2004). When Hurricane Frances was bearing down on Florida in 2004, KDD specialists at Walmart's Technology Centre were able to quickly analyze other pre-hurricane sales data and subsequently load up trucks with flashlights, extra beer, and Strawberry Pop-Tarts, all of which sold to expected volumes (Hays 2004).

Technology and the use of data are key components in Walmart's business strategy. Not only can they review what has happened, they can use data to forecast future trends with considerable accuracy. As time goes on, Walmart continues to add technical capabilities to its repertoire. They are now working on launching Radio Frequency Identification (RFID) tags to all shipments to and from their distribution centres.

### *6.3 Comparative Analysis of the Two Cases*

Table 3 in Appendix 1 applies the analytical framework to compare the two cases. UA is at the very start of its KDD process. Moreover, its resources are very limited. One solution is to partner with another university (without sharing data) to develop a WEKA KDD application (4.2.3). On the other hand, Walmart is an exemplar of successful KDD development (Table 1). This is not an overnight success: Walmart has been collecting its point-of-sales data since 1973. On UA's side, UApeople.org allows the Advancement Office to maintain ties with alumni and to keep data about them current. Both cases demonstrate the importance for organizations to build, maintain, and protect their data assets (4.1.1, 4.3)

Table 2 presents attributes that may prove useful for UA's KDD project. However, even if privacy considerations prevent UA from obtaining income or ethnicity information, a valid postal code combined with Census and other public databases will yield valuable information, such as average income by neighbourhood. For its part, Walmart is less interested in its clients' personal information than in what products sell, when, and with what other products. This demonstrates the importance of knowledge domain elicitation and clearly defining strategic objectives (5.2 & Table 1). Each case's objectives dictate different DM tasks (4.1, 5.3 & Table 1)

The two cases demonstrate the importance of decisions taken at the analysis stage (Table 1). Walmart never invested in technology for technology's sake. On the contrary: the feasibility of each project was thoroughly investigated (5.4). As a smaller organization, and like most businesses, UA cannot afford a custom-developed solution (4.2.1). An OSS like WEKA could be a good solution for UA, considering its previous success with Sakai (4.2.3). Cloud computing could be a viable option if measures are taken to guard personal information confidentiality (4.3.2). Walmart ensures that all data is stored on its own infrastructure for maximum protection against competitors (4.2.1).

## 7. Recommendations for Industry

The objectives of the KDD project must be in line with a firm's strategy. Managers may use the analytical framework to determine whether or not the KDD system supports the corporate strategic goals.

Spending-limited resources should be strategic, as well as taking into account the resources required to process massive amounts of data. Cloud computing presents the advantage of lower infrastructure costs and great flexibility. Before deciding between a licensed or open access solution, businesses must take into account such factors as product compatibility, scalability, available expertise and costs. Managers must have realistic expectations of what can be

achieved with their investment and their current databases. Researching publicly available data sources can be a starting point to expand resources and avoid effort duplication.

Due to the sensitive nature of the data assets and their importance in gaining a competitive advantage, both collection and preservation of data should be considered during the KDD process. After the initial investment in a KDD system, preserving the data is crucial in order to sustain its usefulness. Security measures must be put in place, especially if cloud computing is considered. The adopted solution must take into account confidentiality and security, as well as scalability and costs.

## 8. Conclusion

In this paper, we developed an analytical framework which was used to study the technical capabilities and strategies of two very different organizations. This framework is useful to both researchers and managers.

KDD in the 21st century helps organizations further their business strategy. It can enhance the competitive edge of businesses willing to invest the required time, money and effort. A vast amount of data is collected across various organizations, with capabilities increasing exponentially. These data are reviewed and analyzed in great depth to extract useful and actionable information. With technological advances, the future of KDD looks bright. The possibilities seem endless.

However, there are risks associated with this evolving technology. Consumer associations warn that collecting and using this data may seriously impact consumer privacy (Hays 2004). It is important for information privacy legislation to keep up with technological advances without, however, needlessly impeding legitimate business needs.

## References

Allen R.S., Helms M.M., Jones A., Takeda M.B. & White C.S. (2008). Porter's Business Strategies in Japan. Business Strategy Series, 9 (1), 37-44. http://dx.doi.org/10.1108/17515630810850109

Antonova , A., Csepregi, A., & Marchev A. Jr. (2011). How to Extend the ICT Used at Organizations for Transferring and Sharing Knowledge. The IUP Journal of Knowledge Management, (9) 1, 37–56. Found in ABI/Inform Global.

Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R.H., Konwinski, A., …. & Zaharia M. (2009). Above the Clouds: A Berkeley View of Cloud Computing. Technical Report No. UCB/EECS-2009-28. Electrical Engineering and Computer Sciences, University of California at Berkeley. [Online] Available: http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html

Au, S.T., Duan, T., Hesar, S.G., & Jiang W. (2010). A Framework of Irregularity Enlightenment for Data Pre Processing in Data Mining. Annals of Operations Research, 174 (1), 47-66. http://dx.doi.org/10.1007/s10479 008 0494 z

Bianco, A. (2006). Wal-Mart: The Bully of Bentonville. New York: Doubleday.

Bonacich, E., & Hardie, K. (2006). Wal-Mart and the logistics revolution. In N. Lichtenstein (Ed.) Wal-Mart: The Face of Twenty¬¬-first-century Capitalism (pp. 163 - 187). New York: The New Press.

Cao, L., & Zhang, C. (2007). The Evolution of KDD: towards Domain-Driven Data Mining. International Journal of Pattern Recognition and Artificial Intelligence, 21 (3), 1-16.

Chang, G. (2011, March 04). Thin provisioning optimizes storage utilization and reduces costs. Networkworld. [Online] Available: http://www.networkworld.com/news/tech/2011/030411-thin-provisioning.html

Chaoji V., Al Hasan, M., Salem, S., & Zaki M.J. (2009). SPARCL: An Effective and Efficient Algorithm for Mining Arbitrary Shape-Based Clusters. Knowledge Information Systems, 21 (2), 201 229. http://dx.doi.org/10.1007/s10115-009-0216-0

Clotfelter, C. T. (2001). Who are the Alumni Donors? Giving by Two Generations of Alumni from Selective Colleges. Duke University: NBER.

Crespo, F., & Weber R. (2005). A Methodology for Dynamic Data Mining Based on Fuzzy Clustering. Fuzzy Sets Systems, 150(2), 267-284. http://dx.doi.org/10.1016/j.fss.2004.03.028

Dennis A., Haley Wixom B., Roth R.M. (2012). Systems Analysis & Design, 5th Edition. Hoboken, NJ: John Wiley & Sons Inc. 563 p.

Dunham, M. (2003). Data Mining, Introductory and Advanced Topics. New Jersey: Pearson Education Inc.

Fishman, C. (2006). The Wal-Mart Effect. New York: Penguin Books.

Frank, E., Hall, M.A., Holmes, G., Kirkby, R., Pfahringer, B., & Witten I.H.(2005). WEKA: A machine learning workbench for data mining. In O. Maimon & L. Rokach(Eds), Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers (pp. 1305-1314). Berlin: Springer.

Grace, D. & Griffin, D. (2006). Exploring conspicuousness in the context of donation behaviour. International Journal of Nonprofit and Voluntary Sector Marketing, 11: 147–154. http://dx.doi.org/10.1002/nvsm.24

Griffin, G. J. (2005). Who's Your Donor? A Practical Approach to Building a Revenue-Producing Library Prospect Database. The Bottom Line: Managing Library Finances, 18 (3), pp. 138-145. http://dx.doi.org/10.1108/08880450510613614

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten I.H. (2009). The WEKA Data Mining Software: An Update. ACM SIGKDD Explorations Newsletter, 11(1), 10-18. http://dx.doi.org/10.1145/1656274.1656278

Han, J., & Kamber, M. (2006). Data Mining: Concepts and Techniques . San Francisco: Morgan Kaufman, 770 p.

Hanna, M. (2004a). Data Mining in the e-Learning Domain. Campus Wide Information Systems, 21 (3), 132 138. http://dx.doi.org/10.1108/10650740410544036.

Hanna, M. (2004b). Data Mining in the e-Learning Domain. Campus Wide Information Systems, 21 (1), 29-34. http://dx.doi.org/10.1108/10650740410512301

Havens, J.J., O'Herlihy, M.A., & Shervish, P.G. (2006). Charitable Giving: How Much, By Whom, To What, and Why. In The Nonprofit Sector: A Research Handbook, Second Edition. Walter W. Powell and Richard Steinberg (eds.) Yale University Press. [Online] Available: http://www.bc.edu/content/bc/research/cwp/publications/by-year/publications-2006.html

Hays, C.L. (2004, November 14). What They Know About You: An Obsessive Monitor of Consumor Behavior. New York Times. p. BU1. Retrieved from http://search.proquest.com/docview/92770930?accountid=9744

Henze, L. (2008) Using Statistical Modeling to Increase Donations. Target Analytics White Paper. January 2008. Blackbaud Analytics. Accessed 14 June 2011. 5 p. retrieved from: www.blackbaud.com/files/resources/downloads/WhitePaper_TargetAnalytics_StatisticalModeling.pdf

Hirji K.M. (2001). Exploring Data Mining Implementation. Communications of the ACM. 44 (7). 87-93. http://dx.doi.org/10.1145/379300.379323

Horwitz, R. (2010). Licencing SQL Server 2008 R2. Licensing Directions on Microsoft. Directions on Microsoft. Updated 25 October 2010. Accessed 19 December 2011. 44 p. http://www.directionsonmicrosoft.com/licensing/30-licensing/1776-licensing-sql-server-2008-r2-.html

Hu, R. (2010). Medical Data Mining Based on Association Rules. Computer and Information Science, 3 (4), 104-108. [Online] Available: http://www.ccsenet.org/cis

IBM Corporation (2010). CRISP-DM 1.0: Step-by-step data mining guide. IBM White Paper YTW03084GBEN IBM Corporation, Somers, NY, May 2010, Accessed 17 June 2011. [Online] Available: ftp://public.dhe.ibm.com/common/ssi/ecm/en/ytw03084gben/YTW03084GBEN.PDF

Le Blanc, L.A., & Rucks, C.T. (2009). Data Mining of University Philanthropic Giving: Cluster-Discriminant Analysis and Pareto Effects. International Journal of Educational Advancement, 9 (2), 64-82. http://dx.doi.org/10.1057/ijea.2009.28

Mariscal, G., Marban, O., & Fernandez C. (2010). A Survey of Data Minig and Knowledge Discovery Models and Methodologies. The Knowledge Engineering Review, 25(2), 137-166. http://dx.doi.org/10.1017/S0269888910000032

Meisel, S., & Mattfeld D. (2010). Synergies of Operations Research and Data Mining. European Journal of Operational 206 (1), 1–10. http://dx.doi.org/10.1016/j.ejor.2009.10.017

Mishra, B., & Bhaskar, U. (2011). Knowledge management process in two learning organizations. Journal of Knowledge Management, 15(2), 344-359. http://dx.doi.org/10.1108/13673271111119736

Mundra, N., Gulati, K., & Vashistsh R. (2011). Achieving Competitive Advantage Through Knowledge Management and Innovation: Empirical Evidences from the Indian IT Sector. The IUP Journal of Knowledge Management, 9 (2), 7-25.

Nurmi, D., Wolski, R., Grzegorczyk, C., Obertelli, G., Soman, S., Youseff , L., & Zagorodnov D. (2009). The Eucalyptus Open-Source Cloud-Computing System. In Proceedings of the 2009 9th IEEE/ACM International Symposium on

Cluster Computing and the Grid (CCGRID '09). (pp. 124-131). Washington, DC, USA:IEEE Computer Society. http://dx.doi.org/10.1109/CCGRID.2009.93

Ofiori D. & Atiogbe E. (2012). Strategic Planning in Public Universities: A Developing Country Perspective. Journal of Management and Strategy, 3 (1), 67:82. http://dx.doi.org/10.5430/jms.v3n1p67

Olafsson, S., Li X., & Wu S. (2008). Operations research and data mining. European Journal of Operational Research, 187 (3), 1429-1448. http://dx.doi.org /10.1016/j.ejor.2006.09.023

Olson, D. L., & Delen D. (2008). Advanced Data Mining Techniques. Berlin: Springer.

Ortega, B. (1998). In Sam We Trust. New York: Times Books, Random House.

Petrovic, M., & Hamilton G. (2006). Making Global Markets: Wal-Mart and Its Suppliers. In Nelson Lichtenstein (Ed.) Wal-Mart: The Face of Twenty-First-Century Capitalism (pp. 102-141). New York: The New Press.

Rivard S., Aubert B.A., Patry M., Paré G., Smith H.A. (2004). Information Technology and Organizational Transformation. Oxford UK: Elsevier Butterworth-Heinemann. 321 p.

Sanayei A., & Sadidi M. (2011). Investigation of Customer Knowledge Management (CKM) Dimensions: A Survey Research. International Journal of Business and Management. 6 (11). http://dx.doi.org/10.5539/ijbm.v6n11p234

Sankar, K. P., & Mitra, P. (2004) Pattern recognition algorithms for Data Mining. New York: Chapman & Hall/CRC.

Sharma, M. (2011, June 8). ID-Only Shopping Encourages Wal-Mart in India Where Retail is Off Limits. Bloomberg News. [Online] Available: http://www.bloomberg.com/news/2011-06-08/id-only-shopping-encourages-wal-mart-to-open-stores-in-india.html

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. Journal of Data Warehousing, 5(4), 13—22.

Stewart, K.J., Ammeter, A.P., & Maruping, L.M. (2006). Inpacts of License Choice and Organizational Sponsorship on User Interest and Development Activity in Open Source Software Projects. Information Systems Research, 17(2), 126-144. http://dx.doi.org/10.1287/isre.1060.0082

Sullivan, L. (2004, September 27) Wal-Mart's Way. Information Week. (1007), 36-50. [Online] Available: http://search.proquest.com/docview/229138777?accountid=9744

Tan J., Zhang J., & Li, W. (2010). An Improved Clustering Algorithm Based on Density Distribution Function. Computer and Information Science, 3 (3), 23-29. [Online] Available: http://www.ccsenet.org/cis

Tang, Z.H., & MacLennan J. (2005). Data Mining with SQL Server 2005. Indiannapolis, IN: Wiley Publishing Inc. 460 p.

Walmart Corporation. (2008). Information Systems (ISD). Wal-Mart Corporate Careers. Updated 1 /21/ 2008. [Online] Available: http://walmartstores.com/Careers/7688.aspx , accessed 10 June 2011.

Waweru, M.A. (2011). Comparative Analysis of Competitive Strategy Implementation. Journal of Management and Strategy, 2(3), 49-61. http://dx.doi.org/10.5430/jms.v2n3p49

Wong, R.C.W., Fu, A.W.C., & Wang, K. (2005). Data Mining for Inventory Item Selection with Cross-Selling Considerations. Data Mining and Knowledge Discovery, 11 (1), 81-112. http://dx.doi.org/10.1007/s10618-005-1359-6

**Notes**

Note 1. Numbers in parentheses refer to a section or subsection from this article.

Table 1. The Refined Data Mining Process (Mariscal et al. 2010)

| Processes | Sub Processes |
|---|---|
| 1. Analysis Process | Life Cycle Selection: At this stage, the organization decides which methodology is most appropriate for the project. They must also decide whether they have the expertise to complete the project or whether some or all of the project's IT aspects should be outsourced. |
| 2. Domain Knowledge Elicitation | Essential to understanding the problem and the data. It also limits the search space. |
| 3. Human Resources Identification | A team including both business managers and IT experts must be formed early on. The business managers contribute business-specific knowledge. |
| 4. Problem Specification | Concrete KDD objectives must be established to avoid finding answers to the wrong questions. |
| 5. Data Prospecting | Data is collected from internal databases as well as from outside the company. |
| 6. Data Cleaning | Includes looking for and correcting mistakes, sampling, sorting out outliers, and possibly balancing (Mariscal et al. 2010). Sources of data errors include data entry mistakes, faulty sensor readings, malicious activities, and outdated data (Au et al. 2010). Some data may be redundant (Meisel & Mattfeld 2010). |
| 7. Development Process | 7.1 Preprocessing.<br>7.2 Data reduction and projection.<br>7.3 Selection of the data mining function.<br>7.4 Selection of the data mining algorithm.<br>7.5 Building model (application of data mining algorithms).<br>7.6 Improving model.<br>7.7 Evaluation.<br>7.8 Interpretation of results.<br>7.9 Deployment. Applying the discovered knowledge.<br>7.10 Automation. Performed to let non-expert data mining users apply previously obtained models to new data. (Mariscal et al. 2010, p. 162). |
| 8. Maintenance Process | 8.1 Establish on-going support. |

Table 2. Preliminary Criteria for Predictive Modeling

| | | | |
|---|---|---|---|
| Age | Board member of a cultural institution | UA faculty or UA emeritus faculty | Business title |
| Income, Assets | Citizenship | Connection to campaign | Degree earned |
| UA Employee | Employment status | Ethnicity | Event attendance |
| Faculty member at other institution | Family members in the database | Giving Frequency | Gender |
| GPA | Had a mentor | Donation category (cash, in kind, …) | Wealth |
| Involvement as student | Marital status | Post-degree involvement | Postal Codes |
| Previous giving tendencies | Religion | Reunion year | Satisfied with degree, experience |
| Bequest to UA | Volunteer status | Where money was donated | |

The preliminary criteria in Table 2 were collected from interviews and from Cotfelter 2001; Grace & Griffen 2006 and Havens, O'Herlihy, Shervish 2006.

Table 3. Comparative Analysis of the Two Case Studies

| Analysis Criteria | University A | Walmart |
|---|---|---|
| Porter's Strategy | Product Differentiation: UA works at distinguishing itself from other universities. | Cost Leadership: With lower costs, Walmart sells for lower prices than its competitors while generating high profits. |
| KDD Objective | Find donors will identify with and support with UA new projects. | Lower inventory and distribution costs. Anticipate demand for products. |
| Degree of Advancement of KDD Process | Life Cycle Selection and Domain Knowledge Elicitation. UA is at an early decision stage. | Development and Maintenance processes. KDD is well established. The company refines, improves, and maintains its KDD processes. |
| Financial Feasibility | Limited financial resources, which limits choices. | Virtually no financial limits on its IT investments. However, each project must be rigorously justified. |
| DM Expertise | UA's IT team has limited experience in DM applications. External experts will by required. | Walmart has developed over the years a high level of KDD and DM expertise within the organization. |
| DM Tasks | Cluster Analysis, Classification, Forecasting | Market-Basket Analysis, Forecasting Deviation Analysis |
| DM System Solution | OSS-based Application: OSS could significantly lower projects costs. UA's IT team already familiar with risks and benefits of OSS. | In-house Developed Solution: Competitors do not have access to Walmart's data, nor its KDD programs. |
| Data Processing and Storage | Cloud Computing would allow project to start on small scale and upscale as needed. | Full Provisioning: Walmart maintains absolute control over its data resources. |

Table 3 compares the two cases according to the proposed analytic framework (section 5).