DATA MINING AND STATISTICS: EXAMINING CRITICAL PATTERNS OF RESEARCH AND PRACTICE

by

AILING WANG

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

May 2006

ACKNOWLEDGEMENTS

I would like to acknowledge and thank many people, without whose help this study would not be a reality. First, I would like to thank my co-advisors, Drs. Mary M. Whiteside and James T. C. Teng. I thank them for their support and encouragement in all areas of my academic life, and for many thought-provoking conversations. I am overwhelmed by gratitude for their guidance and numerous hours spent on my dissertation, which was crucial to the completion of this work.

I would also like to acknowledge other members of my committee. I thank Dr. Mark Eakin for inspiring discussions in the earlier stage of the study, which were important for both the research topic and selecting the research methodology. I am extremely grateful for the advice and assistance I received during my stay at UT Arlington from Dr. Greg Frazier. Dr. Sridhar Nerur, a great, wonderful person brought me into the world of citation analysis. I thank all the members of the committee for their valuable feedback and comments.

Other professors to be acknowledged from UT Arlington: Dr. Doyle Hawkins Dr. Nancy Rows, Dr. R. C. Baker, and Dr. Xueming Luo, who gave me the best academic assistance whenever I needed it. In addition, I would like to thank my doctoral classmates and friends: Aakash Taneja, Anil Singh, Peishan Tsai, and George Mangalaraj, who gave up their studying time to help me with many issues from miscellaneous everyday life to academic activities.

Finally, my family deserves special thanks. To my mom, dad, and sister: it is largely because of your constant support and encouragement that I am completing my Ph.D. To my husband, Dustin, and my son, George: you have made my life wonderful and have provided me with inspiration in all walks of life.

April 15, 2006

ABSTRACT

DATA MINING AND STATISTICS: EXAMINING CRITICAL PATTERNS OF RESEARCH AND PRACTICE

Publication No. ______

Ailing Wang, PhD.

The University of Texas at Arlington, 2006

Supervising Professors: Drs. Mary M. Whiteside and James T. C. Teng

Data Mining (DM) has gained increasing attention in academia and growing importance in business over past decades. This is a study to examine the current status of DM in both research and practice with a focus on the role of statistics for the advancement of the field.

Being multi-disciplinary in nature, DM challenges researchers to work together and be aware of the progress made in other fields in order to contribute towards the development of the field. By employing citation analysis techniques and using publicly available citation data, I empirically examine how DM reference fields, such as statistics, machine learning, pattern recognition, and database systems, have shaped the

intellectual structure of DM and how these fields have communicated with and learnt from another.

Organizations are eager to employ this enabling tool to leverage business intelligence for more effective decision-making and improved operations. However, little empirical research explores the underlying factors that influence the success of a DM project. As a result, organizations lack guidance in their DM endeavor. Through field studies, I examine how DM practitioners do DM and what they perceive to be critical for the success of the projects. Based on the findings, theories, and prior empirical research, I further adapt two research models that relate the influence of both team member skills and important project process characteristics to DM implementation success. A survey instrument has also been developed, which could be used to empirically validate and confirm the relationships.

For researchers, results of the study should constitute a first step toward understanding how skill sets and project process characteristics affect DM implementation success. For practitioners, the results offer guidelines for conducting DM projects, particularly in the integration of statistics to achieve DM success.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Table Page

CHAPTER 1

INTRODUCTION

The ease of data collection and advances in Information Technologies such as storage capability, processing power, and access speed have enabled organizations to accumulate vast amount of data. In fact, it is properly said that we are now drowning in data. So there is a tremendous need to make sense out of these data and convert them into information and knowledge. The process, which has effectively helped organizations such as Johnson & Johnson, GE Capital, Procter & Gamble, and Harrah’s Casino to sustain strategic advantages in competitive environments, is known as Data Mining (DM).

Over the past few decades, DM has gained increasing attention in academia and importance in business. Ever since its inception, DM has relied on disciplines such as statistics, pattern recognition, machine learning, and database fields for theory and methodologies. Being multi-disciplinary in nature, DM challenges researchers to work together and be aware of the progress made in other fields in order to contribute towards the development of the field. The risk that the research in DM is heterogeneous and overlapped is considerable. In addition, the understanding of DM is inconsistent within the field. DM proponents have noted these problems anecdotally for a long time. However, few have empirically investigated these issues to validate these impressions or to expand an understanding of the underlying issues.

In business, DM is widely used as an enabling technique for knowledge creation and to help in enhancing the global competitive potential of companies. Organizations are eager to employ DM to leverage business intelligence for more effective decision-making and operations. However, though research in DM itself has made great achievements, in many cases, it still does not adequately fulfill the need of business clients (Kohavi and Rothleder et al. 2002). Organizations that have attempted DM initiatives have experienced serious complications and difficulties, and many failures have been reported. Organizations need guidance in their DM endeavor. Ironically, there has been little empirical research exploring the underlying factors that influence the success of a DM project. Business and academia definitely need to work together to bring rigor and relevance in the nature of research being done (Piatetsky-Shapiro 2001).

DM, being a relatively new field in both research and practice, needs a broad study to gain insights into the field and to develop a sound theoretical base. Towards this end, the main objective of this study is to examine the current-state of DM from both research and business perspectives. This research identifies and assesses how DM is influenced by statistics. Attention is particularly directed to the complementary relationship between statistics and DM, as a better integration of statistics and DM could increase the value of DM.

Specifically, this study employs three research methodologies. First, using citation analysis, it assesses the intellectual structure of DM and the dynamic interchange of ideas among DM reference disciplines. The findings facilitate an understanding of the stratification of DM as a research field in academia and how it is influenced by other disciplines in general, and statistics, in particular. This study also conducts case studies to identify how DM practitioners in real life are influenced by knowledge areas from statistics. Understanding the way a DM project is undertaken through case studies in firms provides an insight into the antecedents of DM implementation success. The lessons learned from the case studies along with the theoretical underpinning from the literature lead to a development of two research models and a pilot survey to validate them. The mapping of these findings is conducive to an understanding of the cohesion between academia and practice, and provides an opportunity to infer the available prospects in terms of both research and practice.

By employing these three research methodologies, this research examines the following research questions as formulated in Chapter III:

1. To what extent have statisticians and researchers in other fields influenced and shaped the academic field of DM?
2. How do statistics and other DM reference disciplines exchange ideas and learn from one another?
3. How is DM practice influenced by statistics?
4. What are the influences of the skill set and stage-efforts profile on DM implementation success?

This study contributes to the DM literature in several ways. It goes beyond anecdotal accounts of relationships among DM reference disciplines to examine empirically the dynamic interchange of ideas among them and offers a better understanding of the DM field. The lessons learned from case studies provide further guidelines for organizations in undertaking DM projects. In particular, it suggests to DM practitioners the skill requirements in successfully deployed DM projects. Finally, the model proposed in this thesis may be employed as a baseline for future DM implementation research.

This thesis is organized as follows. Chapter Two presents necessary background information about DM for readers to further understand this thesis. A literature review and research agenda follow in Chapter Three. Then the research questions are answered in Chapter Four (Citation Analysis of Data Mining Research), in Chapter Five (Data Mining Practices in the Retailing Industry), and in Chapter Six (A Preliminary Survey) respectively. The conclusions, suggestions, and limitations of this study are summarized in Chapter Seven.

CHAPTER 2

BACKGROUND

2.1 Data Mining Applications

The relentless quest for generating and leveraging business intelligence has brought considerable attention to DM. In an ever changing and turbulent business world, organizations rely on such intelligence to stay ahead of their competition. As a result, DM finds a wide range of applications in today's business. Driven by slim margins, the retailing industry probably saw its first DM application in the 1980s. Since then, DM has been widely explored in different business areas, such as logistic scheduling, production, assessing employees, and banking or financing, for more informed decision-making and improved business operations.

In marketing and the retailing industry, due to the fast development in database marketing, DM has been widely used in direct mail or catalogue: for instance, to categorize customers or products into different groups. The information derived from DM can be used in determining cross selling or in determining the customers to target for promotion. For example, rather than sending product information to all the customers with the risk of irritating some, the retailers can only send the promotion to those that are most likely to respond. In this process, DM results help retain profitable customers and, at the same time, remain cost effective and increase revenue.

DM is also used in market basket analysis to discover the relationship between personal characteristics and probability of purchase, such as who buys similar products or what products are often bought together. Beer and diapers are a famous example of this. Prior to the data analysis, it is difficult to link beer and diapers in the same shopping basket. Studies have found that young dads are sent to the grocery store on weekend days; they buy diapers for babies and at the same time buy beer for themselves. Such finding from market basket analysis can be used in inventory control or shop layout. For example, some frequently bought items need to be kept in inventory all the time if possible. Furthermore, those seldom bought but highly profitable items also should be stocked regularly, but less frequently. The results also suggest to the store manager that some products usually bought together should be put in nearby shelves to make shopping more convenient to customers, or put reasonably far away to increase possible sales on the way from shelf to shelf.

DM has also been successfully used in corporate analysis and risk management. Since the early 1990s neural networks have found a wide range of applications in the area of finance in organizations. Many other DM techniques have been used in finance or banking literature. To gain the most profit and assure the proper functioning of their businesses, top management teams are actively pursuing ways to better understand the performance of their company, thus taking all possible preventive measures early to minimize loss. Interesting problems include bank failure prediction, bankruptcy prediction, stock price prediction, forecasting interest rates, investment analysis, credit assessment, loan approval, corporate bond rating and risk management, among others.

Nowadays, DM is widely used in Customer Relationship Management (CRM) to increase customer satisfaction by significantly improving the service provided to customers. A good example of how DM can help a company gain competitive advantage through the building of good customer relationships is given in Loveman (2003). According to the author, the proper use of DM is the secret weapon of Harrah's great success over the past a few years. By mining, Harrah's found that 26% of the gamblers in Harrah's actually generated 82% of the revenue. Contrary to normal expectations, its best customers are former teachers, doctors, and others with more flexible time and a stable income. It has also been found that a simple offer of casino chips alone often results in a more positive customer response than an offer of a free hotel room, free steaks, and a small amount of casino chips. Harrah's also built a predicative model to forecast its customer's lifetime value and classified customers into three groups. The results of DM allowed Harrah's to distinguish the services provided for each customer group and to quickly differentiate itself from its competitors with great gains in revenue and in net income.

Another example of successful DM is the corporate strategy that changed First American Corporation (FAC) from a company operating under a letter of agreement with regulators in 1990 to a profitable leader in the financial services industry in 1999 (Cooper, Watson et al. 2000). To survive in the ever-changing banking industry, FAC started with a customer relationship-oriented strategy. The additional use of DM really supported the smooth implementation of the corporate strategy. For example, DM revealed that the average size of the balance in one's checking account could explain the difference between profitable and unprofitable Seniors Accounts products. This led to the redesign of a product that creates great benefit for both FAC and its customers.

The DM results also were used in distribution management of its service. For example, result shows that closing one branch, associated with high operational cost, and increasing the number of ATMs, with less operational cost involved, would increase profit and better meet customers' needs. A trade-off between one more branch and more ATMs was made, and this created over 20% return on investment. FAC also used traditional statistical techniques, such as experimental design and conjoint analysis, to derive the preference information for all its customers. Working on the information of the preference of high value customers, FAC redesigned marketing programs targeting this group and resulting in a 15% increase in revenues.

Electronic commerce is emerging as the "killer domain" for DM (Kohavi and Provost 2001). Amazon.com is a good example of how businesses operationalize DM results into a closed-loop CRM system. One may have such an experience when he/she is browsing its website. After a few moments of browsing, a pop-window appears and asks whether he/she is interested in a particular book. Very often it is a kind of book that the browser is looking for. How does this happen? Actually, when one clicks and clicks, the system is tracking his/her browsing patterns. Once the patterns are clear, the system knows what types of things the person is interested in based on its past experience. The system will then recommend something assuming that people who share similar browsing patterns would very likely want to buy the same kind of things. For more DM applications in business please refer to Apte, Liu et al. (2002), Hormonzi and Giles (2000), and Fayyad and Uthurusamy (2002).

Other industries such as health care also have an increasing awareness of the importance of DM. Physicians record their patients' information, such as the symptoms the patient has during each visit. An analysis of such a data set would suggest to the physicians the type of treatment that should or should not be given to another patient with similar symptoms. When the patient gets the prescription during his visit to a physician's office and goes to the pharmacy, both the patient and the physician's information are entered into the database in the pharmacy. Such data can be used by medicine manufactures to analyze the prescribing patterns of each physician, which helps in deciding the ways a new product should be suggested to the physician.

DM has been also used in exploring scientific data. For example, biologists have used it to investigate the DNA and protein sequences; meteorologists have used it to study geospatial data, climate data and the earth's ecosystems, to name a few. Another widespread application of DM is in national security by government. A detailed description of DM applications in these areas is beyond the scope of this paper. For more information, please refer to Brachman and Anand (1996), Piatetsky-Shapiro, Brachman et al. (1996), and Han, Altman et al. (2002).

2.2 Types of Data Mining

Due to the looseness in the definitions of DM and the heterogeneous research areas among various fields, researchers categorize DM in different ways. Some classify DM according to the algorithms used, while others classify DM according to the types of problems solved. For example, Groth (1998) defines three fundamental approaches to DM, namely supervised learning, unsupervised learning, and visualization studies.

Hand (1998) distinguishes two types of DM, namely models and patterns, where statisticians are typically pursuing the former type, while DM researchers from other fields are more actively finding patterns. Based on the goals of DM, Berry and Linoff, well-known DM consultants in business, categorize DM into hypothesis testing, undirected DM, and directed DM (Berry and Linoff 2004). While Piatetsky-Shapiro, Brachman et al. (1996) defined two types of DM as verification and discovery, where discovery is further divided into prediction and description. To be consistent with the DM literature in business areas, this study adopts Berry and Linoff's categorization, but incorporates hypothesis testing into directed DM.

Directed DM, equivalent to supervised learning, is a top-down or modeling approach that builds a model from the data set to describe one particular variable of interest in terms of other variables. This often takes the form of predictive modeling, where one knows exactly the predictive range of values. For example, to predict a customer's income status in one of the three known categories - high, medium, or low - we can build a model where the income status is the predicted variable, and all other variables that can help explain the difference in income status are predictors. Once new customer information comes in, the model can place the customer into the relevant category. The term supervised learning also applies when there are three possible income classes, and the model must place the subject into one of them. Hypothesis testing is a traditional statistical approach to data analysis, where the purpose of DM is to verify the hypothesis.

In contrast, undirected DM, equivalent to unsupervised learning, is a bottom-up approach where the data are explored. Many people believe this is what DM really is. Here, one does not specify the variable of interest, but wants to find relationships among all the variables. It is unsupervised learning because there are no pre-defined categories, but the user allows the data to suggest the answer. For example, if a retailer has information about his customers - age, marital status, number of children, number of cars owned, details of each transactions - he can use DM to group those customers together based on similar characteristics, while the dissimilar customers are placed in different groups. Such an analysis is typically performed in customer segmentation, and can be used to market a new product by sending the promotion to relevant customers instead of to all the customers. This can greatly increase the effectiveness of the campaign, leading to a better customer response with significant reduction in cost. Clustering, association rules analysis, and data visualization all are examples of undirected DM.

2.3 Data Mining Techniques

Earlier research on DM more focuses on optimizing algorithms and inventing new techniques. As a result, many DM techniques and methods have been made available for organizations to perform DM. DM techniques have been categorized in different ways. A popular categorization is given by Fayyad, Piatetsky-Shapiro et al. (1996) and Fayyad and Stolorz (1997), in which they categorize DM according to tasks into predictive modeling, clustering, data summarization, dependency modeling, change and deviation detection. These same authors also categorize DM techniques into five areas: decision trees and rules, non-linear regression and classification, example-based methods, probabilistic graphical dependency models, and relational learning models. Since the focus of this research is not on DM techniques/methods, in the remaining part of the subsection, a few popular DM techniques will be briefly described. The interested reader can refer to other technical documents, which are easily accessible.

Predictive modeling is either a regression problem if the predicted is a numeric variable or a classification problem if the predicted is a categorical variable. Specific techniques used in predictive modeling include Classification and Regression Trees (CART), regression, decision trees, metric-space based methods (such as the k-nearest-neighbor method), projection into decision regions (such as discriminant analysis, rule-based classifier), and neural networks for nonlinear decision surfaces. Genetic algorithms, rough set and fuzzy set approaches are also widely used in classification.

Decision trees are used to discover rules and relationships by systematically breaking down and subdividing the information contained in the data set by CART, ID3, C4.5 or CHAID algorithms. It can be traced back to the early 1980s in statistics (Breiman, Friedman et al. 1984), and later to machine learning communities (Quinlan 1986; Quinlan 1993), and is a technique widely used in classification and prediction problems when there are categorical variables, or it is reasonable to categorize variables.

Neural Networks (NN) use layers to train and model the data. A simple NN model consists of one input layer, one or more hidden layers, and one output layer. A popular architecture used in neural networks is feed-forward and back-propagation. By feed-forward, the data enters through the input layer and is transformed / weighted over hidden layers, and then eventually exits from the output layer without looping back. Back-propagation is used in the training phase by calculating the error in prediction, which is then used to adjust the weights to achieve a smallest prediction error for the training data set. NNs are sophisticated, non-linear methodologies capable of prediction in extremely complex scenarios by first learning from example; however, the big critique of NN is its "black box" approach, which makes the appropriate interpretation of the results difficult.

The DM techniques mentioned so far are mainly used in directed DM. The remainder of this section will discuss some techniques that are mainly used in undirected DM, such as data summarization, clustering, and data visualization.

Data Summarization is mainly about association rules analysis. In business, typically, it is called "Market Basket Analysis." Two concepts are involved in this technique. One is support, which is the joint probability that people will buy two particular items, say beer and diapers, together in one shopping visit. The second is confidence, which is the conditional probability that given a person has already bought diapers, he (or she) also buys beer in the same visit. Usually, some benchmarks (for example, 5% for support; 80% for confidence) are used, and they jointly determine whether a pattern should be reported. The big problem with this technique is it often reports too many patterns - some of them spurious patterns- which pose a big challenge for the identification of the especially interesting patterns.

Clustering maps data into several classes based on similarity metrics or probability density models. The clusters are not predefined, have no predicted variable and no predefined number of clusters to be formed, which differs from classification. To measure similarity, Euclidean distance is used most often, which, in three dimensional space, is the length of a straight line between two objects. Correlation or some sort of association can also be used. Major clustering methods include non-hierarchical methods, such as partitioning, density-based methods, and hierarchical methods, and agglomerative hierarchical methods, which construct a treelike structure on formed clusters: each object starts as one cluster; then the two closest objects are combined, ultimately ending with all objects in one cluster. Criteria such as single linkage, complete linkage, or Ward's method are used to determine the final solutions. An application of clustering in segmenting customers has been illustrated previously in the types of DM section.

Data visualization techniques use graphical tools to illustrate data relationships in a more readable way. Many of these tools allow visual interpretation of complex relationships in multidimensional data. Dependency modeling is used to describe significant dependencies or causal structures among variables. Density estimation methods and explicit causal modeling all fall in this category. Bayesian Networks are used often in causal modeling. A Bayesian network is a directed acyclic graph, where each variable is represented by a node, and each line with one-sided arrows represents a probabilistic dependence. The line starts with a parent, and the arrow points to the descendent(s) of the parent. All the variables and their joint conditional probability relationships are depicted in the graph.

Besides the above-mentioned DM methods, there are many other statistical techniques widely used in DM, for example, logistic regression. Previously discussed DM applications in Harrah's entertainment and FAC also have used statistical experimental design and conjoint analysis techniques. With so many choices of DM techniques, it is worth noting that no one is best for all purposes. Each technique is typically suited for a particular type of problem and data. The choice of a particular combination of techniques to apply in a particular situation depends on the nature of the DM task, the nature of the available data, and of course the skills and preferences of the DM practitioner.

2.4 CRISP-DM Process Model

Implementing DM requires a mix of an appropriate use of the tool and human resources. The proper integration and management of techniques and human resources form a process in DM implementation. CRoss Industry Standard Process for DM (CRISP-DM) process model (Chapman, Clinton et al. 2000) is generally considered as a standard process model for the DM industry, which is initiated by NCR Corporation and SPSS Inc. According to CRISP-DM, the life cycle of a DM project typically consists of six phases including business understanding, data understanding, data preparation, modeling, evaluation, and deployment (See Figure 2.1).

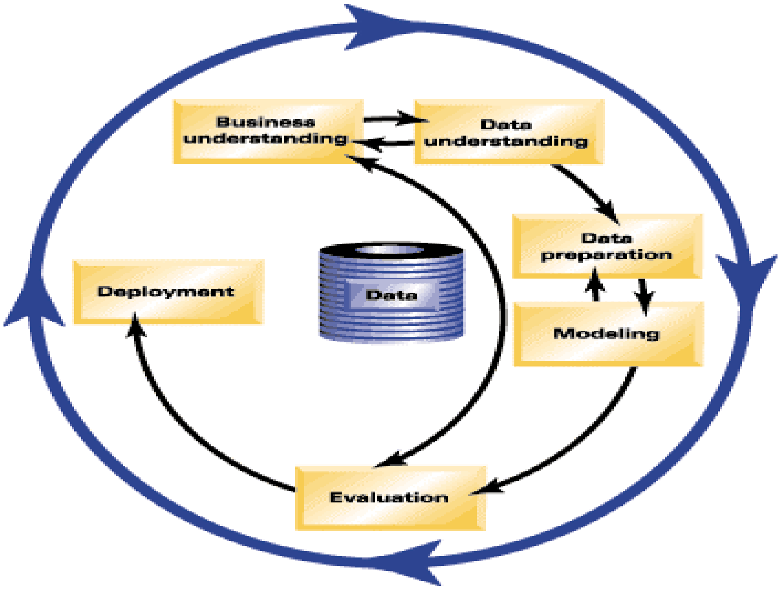


Figure 2.1 CRISP-DM Process Model (adapted from http://www.crisp-dm.org)

In the business understanding phase, researchers need to understand the project objectives and requirements from a business perspective, and then transform the problem into a properly formulated DM problem. In the data understanding phase, researchers need to obtain the initial data set and identify data quality problems. The main objective is to get familiar with the initial sample data set and see whether the data can meet the requirements for solving the business problem. If the data are sufficient to answer the business problem, then one needs to formalize a possible set of hypothesis and gain insights into the data set.

This phase is followed by the data preparation that includes data collection or extraction, data transformation, data cleaning, and other data quality related issues. The final product of this phase is a set of clean data that can be used for analysis in the DM process.

The next phase (at least for directed DM) is modeling. Researchers select and apply various modeling techniques, with appropriate parameters. The selection of the DM tools and techniques, to a larger extent, depends on the types of data and DM. As a result, the researcher may have to go back to the data preparation phase in order to get optimal results.

In the evaluation phase, the model or the results derived from earlier steps will be evaluated before business clients actually deploy them. Since the model or the results are derived from a data analysis perspective, they may not fit the business objectives. On the other hand, the results should be presented in a way that business clients can easily understand. Thus a properly critical evaluation of the DM results is especially important for business. A key point that needs to be determined is whether there are some important business issues that have not been sufficiently considered.

Finally, the DM results are deployed within the business process. Depending on the nature of DM, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable DM process.

Researchers further find that though modeling data or applying DM algorithms to extract patterns is a critical step, it is not the most time-consuming stage. In reality, a large portion of effort goes into how to properly formulate the DM problem, and the most time-consuming part is data preprocessing, typically around 60% to 70% of the effort (Cabena, Hadjinian et al. 1998), while only about 15% to 25% on actual modeling (Brachman and Anand 1996; Piatetsky-Shapiro, Brachman et al. 1996; Kohavi and Provost 2001).

The process model provides important guidelines for both DM practitioners and organizations in actual DM endeavors. As a result, many cases employing or adapting this process model have been reported under various contexts such as mining microarray gene expression data (Piatetsky-Shapiro, Khabaza et al. 2003), quality diagnostic models for the packaging manufacturing industry (Abajo, Diez et al. 2004), and deviation detection and analysis of warranty and goodwill cost statements in the automotive industry (Hotz, Grimmer et al. 2001). In many cases, the use of such a methodology in DM is observed to substantially shorten the time period the project lasts contrasting with in-house methods using conventional techniques (Adderley and Musgrove 2001).

CRISP-DM is a generic process model; under different contexts, users can instantiate this model to meet specific needs. It is worth noting that it is not that different from other DM process models, such as the KDD process model (Fayyad, Piatetsky-Shapiro et al. 1999), and the Virtual DM Cycle (Berry and Linoff 2004). A typical DM project is expected to have all these stages but not necessarily follow the same order, as DM is an interactive and iterative process (Piatetsky-Shapiro, Brachman et al. 1996). The results of each stage can be used either as inputs of the next step, or as feedback to the previous step, thus resulting in a process of continuous improvement.

CHAPTER 3

LITERATURE REVIEW AND RESEARCH QUESTIONS

Though the frequent use of the term DM is relatively new and can be traced back to the early 1990s, DM is not a new phenomenon. The idea of DM has been reflected, to some extent, in exploratory analysis of data. In literature, DM, in many cases, is associated with Knowledge Discovery in Databases (KDD). The term KDD was first used at the first KDD workshop in 1989 to, "emphasize the 'knowledge' is the end product of a data-driven discovery" (Fayyad and Stolorz 1997). In this sense, DM is one critical step of a KDD process (Fayyad, Piatetsky-Shapiro et al. 1999). However, in a broader sense, many researchers treat DM and KDD as synonyms (Chung and Gray 1999; Han and Kamber 2001; Piatetsky-Shapiro 2002). As a research field, DM and KDD go hand in hand. In this paper, I adopt a broad view of DM and use DM and KDD interchangeably.

Over the past a few decades, DM systems have experienced a breakthrough in the transition from being tool-driven to being application-driven. According to Piatetsky-Shapiro (1999), three generations of DM systems have taken place (Figure 3.1). The first generation emerged in the 1980s and is mainly used in the research lab to solve a particular problem. The users of the first generation system are researchers with highly specialized training in one or more DM techniques, such as C4.5, a popular decision tree algorithm. However, the real world is much more complicated than the environment in the research lab. Upon realizing the importance of data processing and how time consuming this task is, researchers actively started incorporating traditional statistical software tools with more powerful capabilities in data collection, data splitting, transformation, and sampling. This leads in around 1995 to the second generation of DM systems: suites. Some of these suites, even popular today, are SAS Enterprise Miner and SPSS Clementine. However, both first and second generation DM systems are tool-driven and require a higher skill set from the users with expertise in DM techniques or algorithms or former training in statistics. It is difficult to integrate such systems with a real-world environment, where the users are typically business people with major functions in either finance or marketing. These users want systems to be easy enough to be utilized in their routine functions. With the realization of the importance of DM as an enabler of competitive advantage and the potential benefit that DM could bring to businesses, organizations are reactively and proactively exploring DM. Fortunately, DM vendors and business people collaborate towards this end. As a result, some customized DM products for particular business problems appeared in the late 1990s. These application-driven products initiate the third generation of DM systems.

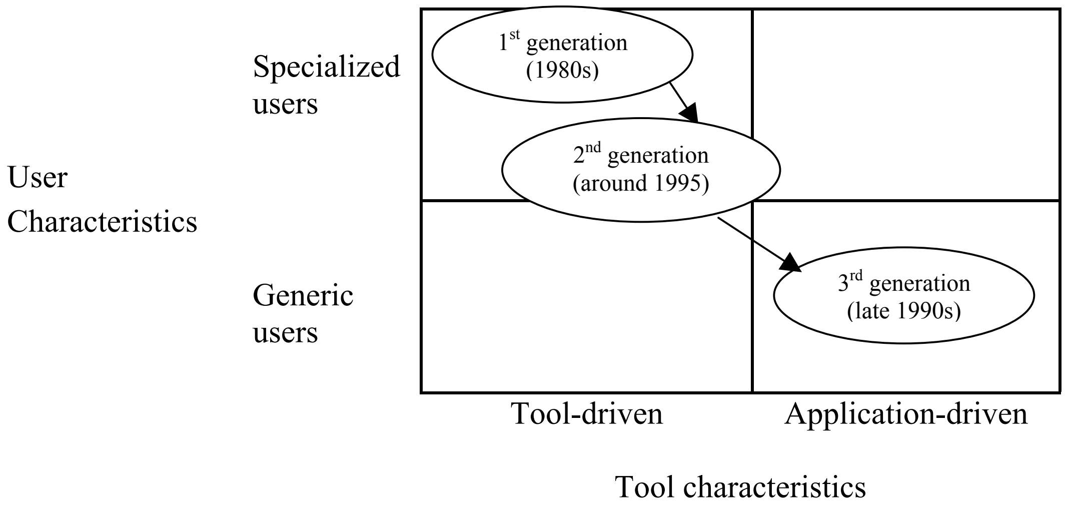


Figure 3.1 Evolutions of Three-Generations of Data Mining Systems

The main driving force for the fast growing field of DM is business needs (Kohavi, Rothleder et al. 2002). In addition, several other factors also have made significant contributions. First, the large volume of data made available due to an increased ability to generate data (e.g., bar codes for commercial products), and an increased ability to store data with better database management systems, data warehousing technology for collecting and cleaning transactional data for online access. It is generally believed that there is a symbiotic relationship between the activity of DM and the data warehouse, and DM is one of the major applications of a data warehouse (Brachman, Khabaza et al. 1996; Inmon 1996). The exponential increase in processing power of computers is another factor that makes the fast processing of this substantial amount of data a reality. As a result, DM is continuously attracting an ever-increasing number of researchers and practitioners.

Since the term DM is relatively new and research in this area is multidisciplinary, there are many extant definitions of DM provided by researchers from different fields under various circumstances. Some of these are listed in Table 3.1.

Table 3.1 A List of Extant Definitions of Data Mining

The extraction of previously unknown information from databases that may be large, noisy, and have missing data.	Chatfield (1995)
DM aims to discover something new from the facts recorded in a database.	Glymour, Madigan et al. (1996)
The application of specific algorithms of extracting patterns from data.	Fayyad, Piatetsky-Shapiro et al. (1996)
DM is about extracting interesting patterns from raw data.	Kleinberg, Papadimitriou et al. (1998)
DM is a term synonymous with data dredging or fishing and has been used to describe the process of trawling through data in the hope of identifying patterns.	Hand (1998)
DM is the process of extracting previously unknown, valid, and actionable information from large databases and then using the information to make crucial business decisions.	Cabena, Hadjinian et al. (1998)
The objective of DM is to identify valid, novel, potentially useful, and understandable correlations and patterns in existing data.	Chung and Gray (1999)
DM is the process of seeking interesting or valuable information within large databases.	Hand, Blunt et al. (2000)
DM is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules.	Berry and Linoff (2000)
DM is the process of discovering interesting knowledge from a large amount of data stored either in databases, data warehouses, or other information repositories.	Han and Kamber (2001)
DM is data processing using sophisticated data search capabilities and statistical algorithms to discover patterns and correlations in large preexisting databases; a way to discover new meaning in data.	Dictionary.com

This research attempts to assess the current status of DM, thus the scope of DM needs to be clearly defined. However, a casual look at the table suggests that the definitions of DM are loose and broad: confusion exists in the literature about what is meant by the term “data mining”. Towards this end, five characteristics are identified, based on previous descriptions on DM, and are considered to be typical in business DM activities (See Table 3.2).

Table 3.2 Characteristics of Data Mining

- It is often about secondary data analysis, in which the data have been collected for other purposes.
- The data may be typically stored in databases, data marts or in a data warehouse.
- The data size is relatively large.
- The purpose can be discovering hidden patterns in the data.
- It is generally conducted to support effective decision-making, to improve operational efficiency, or to create knowledge.

With these characteristics in mind, in the remaining review of literature, attention is particularly directed to:

1. DM and its reference disciplines
2. DM research methods and scope
3. Gaps in current DM research

3.1 The Influence of Statistics and other Disciplines on Data Mining

DM is a multidisciplinary field. Over the years, DM continuously has drawn on theories and technologies from statistics, database systems, pattern recognition, data visualization, high-performance computing, and machine learning from artificial intelligence. Researchers from these disciplines have contributed to the DM literature in different ways.

3.1.1 Data Mining and Statistics: Complementary

DM techniques offer a powerful complement to statistical techniques, especially in large data analysis (Garver 2002). In fact, DM expands the role and value of traditional statistics in business by popularizing its use, while the proper integration of statistics with DM can greatly improve the process of transforming organizational data into knowledge. DM benefits greatly from the wealth of experience that statisticians have acquired in using statistics, especially the art of statistical thinking in problem formulation, problem solving, result evaluation, and in applying the results of DM (Glymour, Madigan et al. 1997; Mackinnon and Glick 1999). Many DM techniques have a statistics origin. For example, the famous work of Classification and Regression Tree (CART) and Chi-squared Automatic Interaction Detector (CHAID) (Kass 1980) form the basis for the development of decision trees and rules. Probability theory and Bayes Theorem constitute the underlying theoretical foundation for association rule analysis and Bayesian networks; while statistical pattern recognition is fundamental to clustering techniques. Other areas of statistics, such as logistic regression, estimation, rational decision making, hypothesis testing, model scoring, Gibb's sampler, meta-analysis, model averaging, time series analysis, Monte Carlo Markov Chain, and multivariate adaptive regression splines (MARS), have also proven to be critical in the development of other DM techniques.

As data analytical tools, statistics and DM share many similarities, it is no wonder that, in the past, some statisticians considered DM as a simple scale-up of statistical techniques, thus a subject of statistics (Hand 1999; Mannila 2000). Specifically, Friedman (1997) raised the issue whether DM is an intellectual discipline, or whether it is a sub-field to statistics. However, some differences exist between traditional statistics and data analysis in DM must be noted (see Table 3.3). One point of departure is the enormous volume of data to which DM is applied compared with the data sets that are traditionally used in statistical approaches (Hand 2000). Traditional statistical techniques often deal with a smaller set of data and researchers collect data for the analytical purpose, while DM typically deals with a very large data set and the data has already been collected for other purposes and stored in databases (Hand 1999). Thus the analysis of data, unlike traditional statistics, is secondary. Another difference is traditional statistics follows a confirmatory-driven approach, while DM is a fundamentally exploratory-driven approach. Note that this exploratory-driven approach is one aspect of statistics but at one time, have negative connotations, called "data snooping" or "data dredging". These differences mandate the recognition that "statistics is not the only game in town" (Friedman 1997).

Table 3.3 Comparisons between Traditional Statistics and Data Mining

	Traditional Statistics	Data Mining
Data Size	Relatively small	Relatively large
Data Type	Mostly primary	Secondary
Typical Approach	Confirmatory-driven	Exploratory-driven

Failing to fully appreciate these differences between traditional statistics and DM in the past may, to a large extent, explain the fact that "Statistics, as a community, has not been actively pursuing data mining, even though a few statisticians have contributed a bunch to data mining" (Friedman 1997).

In recent years, the statistical community has had a major change in its attitude towards DM. More researchers in statistics regard DM at the interface of computer science and statistics (Glymour, Madigan et al. 1996; Hand 1998; Hand 1999). Particularly, in 1997 in the *Journal of the American Statistical Association*'s reprint of his presidential address, Kettenring clearly made this statement:

> *Let me highlight two opportunities that sit at the intersection of computer science and statistics. ... The second area is the current hot topic of data mining, which statisticians might reasonably think of as data analysis of very large databases. In fact, if you dig down and look at what is involved in data mining, you will find a variety of statistical components, such as statistical graphics and cluster analysis. Moreover, there is a great opportunity to bring a variety of statistical concepts to bear—modeling, sampling, robust estimation, outlier detection, dimensionality reduction, etc. Nevertheless, there are new opportunities as well, and we would be wise to pay very close attention and to become seriously involved with these developments*. (Ketterring 1997)

The changed attitude toward DM is also evident by the fact that the *Current Index of Statistics* includes the mainstream DM journals, such as *Data Mining and Knowledge Discovery*. Traditional statistical software vendors, for example, SAS

Institute Inc. and SPSS Inc., empower the software with DM capabilities and are marketing themselves as DM vendors. In its recent call for papers, *CHANCE Magazine* lists both genomics and DM as modern statistical topics.

3.1.2 Data Mining and other Disciplines

Other fields such as pattern recognition, machine learning and artificial intelligence, and database systems have also made substantial contributions to the development of DM. The field of pattern recognition has made an important contribution by illuminating the use of neural networks (Hirji 2001). The AI/ML community has contributed numerous efficient search algorithms, modeling techniques, primarily in decision tree and rational learning (Quinlan 1986; Quinlan 1993; Muggleton and Raedt 1994), and novice approaches for knowledge presentation. The database field also has a fundamental importance to DM (Fayyad and Stolorz 1997) by providing efficient and reliable storage, such as databases and data warehousing, and retrieval techniques, like Structured Query Language (SQL) and Online Analytical Processing (OLAP). Other contributions from the database discipline include various efficient and scalable association rule algorithms (Agrawal, Mannila et al. 1996), and major concepts like privacy preserving DM.

In light of the unique contributions of each discipline made to the DM literature, it is little wonder that researchers are ambivalent about the relevance of the domain of DM to their work. Like the case in the statistical community, confusions linger in the computer science and the engineering field about their role in DM. Some researchers consider the invention of new techniques or optimizing algorithms as the core of DM, thus regarding DM as a subject of computer science (Mannila 2000). Particularly, Quinlan (2000) believes the origins of KDD (DM) are in Artificial Intelligence/Machine Learning, saying “it would be a pity if KDD (DM) stays way from this root.” Piatetsky-Shapiro (1999) expects to see DM emerging into the database industry.

Although DM was multidisciplinary in its origin and needs building upon the theories and/or methodologies from its reference disciplines in order to make further advancement, it is evolving into a fast growing unique field. A recent survey by META Group (2004) projects that the DM industry will continuously increase at a rate of 10% in the next few years, with significant increase in the service industry.

3.2 Data Mining Research Methods and Scope

In this section, I present the results of a literature review based on 389 papers found by the key phrase “data mining” in a search on the database Business Source Premier. Through an in-depth examination of what has been done in the literature in terms of research scope and research methodology, this review suggests areas for future research.

3.2.1 Data Mining Research Scope

In this study, DM research scope (from 1996 to 2005) is categorized into DM process, DM result, DM method, and general issues in DM. The DM process research studies business understanding, data understanding, data preparing, modeling, evaluation, and deployment, and their impact on DM project success. It also answers the question how a project starting from scratch progresses towards completion and during this process the factors that affect the project success. The DM result research has a primary focus on what DM brings or could bring to firms, and its impact on organizations. Both DM process and DM result research study the business issues within DM. In contrast, the DM method research studies new methods and new algorithms in DM, thus has a focus on hard issues within DM. The DM general issues research studies the underlying foundation of DM, its relationship with other disciplines, software tools, data issues, informing new application areas, its social impact etc., mainly anecdotal in nature. Table 3.4 summarizes the research scope in DM research.

Table 3.4 Frequency in Data Mining Research Scope

Research scope	Frequency
DM process	12
DM result	98
DM method	220
DM general issues	59

It is clear from Table 3.4 that a significant amount of previous research in DM has focused on developing new or optimizing algorithms/techniques for large data analysis to be more efficient, effective and scalable, as reflected by the fact that 57% of papers in the initial papers reviewed belong to this category. Previous research also has focused on DM result, while little attention was given to the process characteristics.

3.2.2 Data Mining Research Methods

In this study, DM research methods (from 1996 to 2005) are categorized into four areas: case study (including field study and interview), survey, conceptual work, and applications (applying one or more DM methods to address real-world issues). Since this study has a main focus on business issues in DM, I eliminate 279 papers with a focus on DM method and DM general issues identified in previous section from further discussion. For the 110 papers that remained, I examine the primary methods DM research has adopted. Table 3.5 shows the frequencies of the four research methods in DM research with a focus on business issues.

Table 3.5 Frequency in Data Mining Research Scope and Methods

	DM Process	DM Result
Case study	2	10
Survey	2	4
Conceptual Framework	8	26
Applications	0*	58

*No paper falls in this category for DM papers reviewed within Business Source Premier. But another search on DM proceedings such as ACM SIGKDD reveals some publications in this category, and some of them have already been reported in Chapter 2.

According to Table 3.5, due to the dramatic rise in DM in recent years, many DM applications result, while few survey studies have been conducted. As a relatively new phenomenon in both research and practice, the earlier stages of inquiry into DM have also given attention to the foundations of DM and, in particular, to process models, but most of them are anecdotal in nature. Combining both research methods and scope, a gap in current DM research is clear. Though the phenomenon of DM is already evolving into an increasingly important stance in business, the literature has not extensively examined the entire process aspect of DM by employing case study and survey methods.

One case study that comprehensively examines the DM process characteristics is reported by Hirji (2001). In this study, the author examines how the DM team perform their jobs and how they follow a five-stage process model proposed by Cabena, Hadjinian et al. (1998). It was found that unrealistic expectations about the DM results among team members, and data issues related to extraction, quality, and availability are two main problems in the implementation of this DM project. To accommodate these issues, the team has dropped some business problems, and has added new ones; the team also has had to modify the planned process model. The author also notes that the DM practitioner interacts with clients extensively and, besides the technical skills, he is also an experienced facilitator, which is critical for this project.

A few empirical studies have investigated factors that impact DM implementation success. In one empirical study, Nemati and Barko (2003) surveyed DM professionals to examine the current organizational DM practices. In particular, they investigated data issues, technological issues, organizational issues, people issues, and time, scope and resources issues and their impact on DM success. The authors found that all issues, except organizational issues, have important impact on organizational DM project success. The authors further attributed the missing link between project success and organizational issues (including aligned DM/business strategies, supporting business processes, and new incentive plans) to the inexperience of the respondents. In addition, the study suggests other factors not studied, such as the iterative process characteristics of DM, in-depth knowledge of the industry and business, the proper selection of DM tools and algorithms do affect project success. These findings are summarized in Nemati and Barko (2001) and Nemati and Barko (2003).

Another empirical study examines the effect of organizational attributes on the adoption of DM techniques. Based on a survey of firms in the financial service industry, Chang, Chang et al. (2003) found that the organizational size, attitude toward data resource, and style of decision-making significantly influence DM adoption in the organization. However, DM adoption is not significantly affected by competitiveness of outside environment and organizational culture in terms of marketing orientation and information orientation.

One comprehensive study of DM process employed both case studies and survey is reported by Davenport, Harris et al. (2001). In this study, the authors identified the primary success factors for organizations in the process to transform data into knowledge and then into actionable business results. It was found that business strategy, skills and experience, organization and culture, technology and data context constitute the underlying key contextual factors of this process. In transforming data into knowledge, the analytic process and the decision-making process are employed together; they are iterative and intertwined. The authors further argued that the context and transformation have little value if nothing changes in the organization as a result of them. And they proposed to measure the outcomes in the context of changes in behaviors, in new process, and financial results. In the process, four people groups and five key competencies were observed to have played important roles. The four people groups are: database administrator, business analyst and data modeler, decision maker, and outcome manager. The five competencies are: technology skills (these skills involve knowledge of the software and underlying systems used to extract, manipulate, analyze, and present data), statistical modeling and analytic skills, knowledge of the data, knowledge of the business, and communication/partnering skills.

To avoid possible omission of DM papers relevant to this research, other sources were also located, and it was found that one more research work needs to be mentioned here. In his unpublished dissertation, Sim (2003) followed other studies on critical success factors in information systems, in data management, and in data warehousing, and explored critical success factors of organizational DM projects. He examined the impact of action, dataset, communications, outputs, business mission, consultation, and business environment on DM project success and found that only dataset is a significant factor for DM success.

Although these pioneer studies in case and empirical DM research present evidence about the understanding of DM implementation and provide important insights on how DM has been done in a real-world context, there are some issues that need to be noted. First, all these studies focus on issues at an organizational level with an attempt to explain DM success. Ideally, one should take all organizational factors into consideration in DM success. However, the number of possible factors and the complex relationships among these factors make such a study complicated. Secondly, the factors identified in previous studies are not consistent, and sometimes, actually contradictory. Thirdly, all these studies suggest, to varying extent, that skill sets affect DM project success. However, no one systematically examined what skills and how the team's skills influence DM implementation success, and how DM practitioners can further impact DM project success. In addition, 1) in Hirji's study, the DM project is a simulated one, not to achieve an actionable DM outcome. Furthermore, the DM practitioner is from an external organization, which casts doubts on the validity of the findings for an organization with in-house DM practitioners; 2) in Sim's study, the sample size is relatively small (56).

Conceptually, numerous DM researchers and practitioners in different situations have expressed that a successful DM project requires a sufficient skill base of people who really understand the technology (Gehrke 2002), who have knowledge of statistics (Luan 2002), who have knowledge of databases/data warehouse (Luan 2002), and who have knowledge of business (Chou and Chou 1999). Another obstacle, according to Piatetsky-Shapiro (Koo 1998), in applying DM in a business environment, is a business factor, poor problem definition and poor management. Berry and Linoff (2004) also note "the key to (DM) success is incorporating DM into business processes and being able to foster lines of communication between the technical data miners and the business users of the results." There anecdotal accounts provide important information that needs further research effort. Companies investing in these DM initiatives have also noted complications involved in DM projects and the high risk of failure (Hofmann and

Tierney 2003). The complexity of a particular DM task calls for a systematic examination of the process.

3.3 Research Agenda

In summary, based on this literature review, three major gaps are evident in current DM research.

3.3.1 Data Mining Knowledge Network

As mentioned earlier, DM is a multidisciplinary field incorporating many disciplines. The researchers in these disciplines borrow validated ideas and knowledge from each other to fulfill their research objectives. This communication flow among the researchers in these disciplines results in the formation of a knowledge network which is crucial for the development of DM as a research field. Nonetheless, the research in DM is fragmented, which has posed the potential that researchers from a variety of disciplines make advancement in DM but may be unaware of each other's progress (ICDM 2004). There is still confusion about DM within the field. As a relatively new research field, a study of the intellectual structure is critical for the progress of the field. This leads to the first research question:

1. To what extent have statisticians and researchers in other fields influenced and shaped the academic field of DM?

DM serves as a bridge linking many previously unlinked areas. To assure the healthy development of DM as a field, many of the leading proponents of DM acknowledge the importance of integrating reference disciplines to the field for the advancement and maturity of DM (Goodman 1999; Goodman 2000; Gehrke 2002; Fayyad, Piatetsky-Shapiro et al. 2003). Formal communication among researchers in DM from different disciplines has been noticed from a glance at the literature, e.g., they co-author a research project. Nonetheless, misunderstandings still exist among these disciplines. For example, the misconceptions that machine learning researchers have about databases, i.e., databases are just static tables of simple data, not complex entities with transactions, security, and updates (Piatetsky-Shapiro 1991). Simultaneous DM and statistical/database/machine learning conferences have been arranged to promote healthy and fruitful intellectual communication and collaboration between researchers from various fields. According to Fayyad, Piatetsky-Shapiro et al. (2003), these efforts have not fully yielded notable collaborative outcomes. However, there is a need to empirically assess these efforts and their outcomes in order to explore areas for further improvements. While much conceptual discussion and numerous anecdotal accounts have been reported emphasizing the importance of collaboration in DM, so far, there is no study to empirically examine the interchange of ideas among DM reference disciplines. In particular, how statistics communicates with other disciplines within the context of DM in order to expand our understanding of DM. This leads to the second research question:

2. How do statistics and other DM reference disciplines exchange ideas and learn from one another?

3.3.2 The Influence of Statistics on Data Mining Practice

While DM is getting considerable attention in business, statistical techniques are widely in use, and in many cases, complement DM in addressing business issues. Statisticians with expertise acquired through many years of work in solving these real-world business issues are going to play an important role in the mining process. However, the differences between traditional statistics and DM demand changes in their behavior. There is a need to examine how DM practitioners, including statisticians, have integrated statistics into their DM practice, and what approaches they adopt reactively and proactively to solve DM questions and problems appearing during the implementation process. This leads to the third research question:

3. How is DM practice influenced by statistics?

3.3.3 A Profile for Successful Data Mining Implementation

With the growing realization that DM is a critical enabling technique for organizations to make more informed decision and to operate more efficiently, many organizations attempt to implement DM projects. However, implementing DM involves a group of dedicated people with varieties of expertise, an appropriate process to follow, and many other factors in the organization domain. While a few studies, with a primary focus on organizational DM, provide some ideas, no one has systematically studied skill sets and process characteristics that impact the DM implementation success from the DM practitioners' point of view. There are quite limited guidelines for organizations that attempt to implement DM internally. So far we still do not know what and how skills and process characteristics (namely, the stage-efforts profile, which is defined as the strength of effort, in terms of both time and resources, the DM project team spent on each stage of the process during the implementation of the project) actually impact the DM implementation success. All these issues mandate a reevaluation of the factors that influence DM implementation success. The fourth question is:

4. What are the influences of the skill set and stage-efforts profile on DM implementation success?

This research of the current-state of DM is two-fold: one is DM in research and another is DM in business. The above discussion of research gaps and the corresponding research designs are summarized in Figure 3.2.

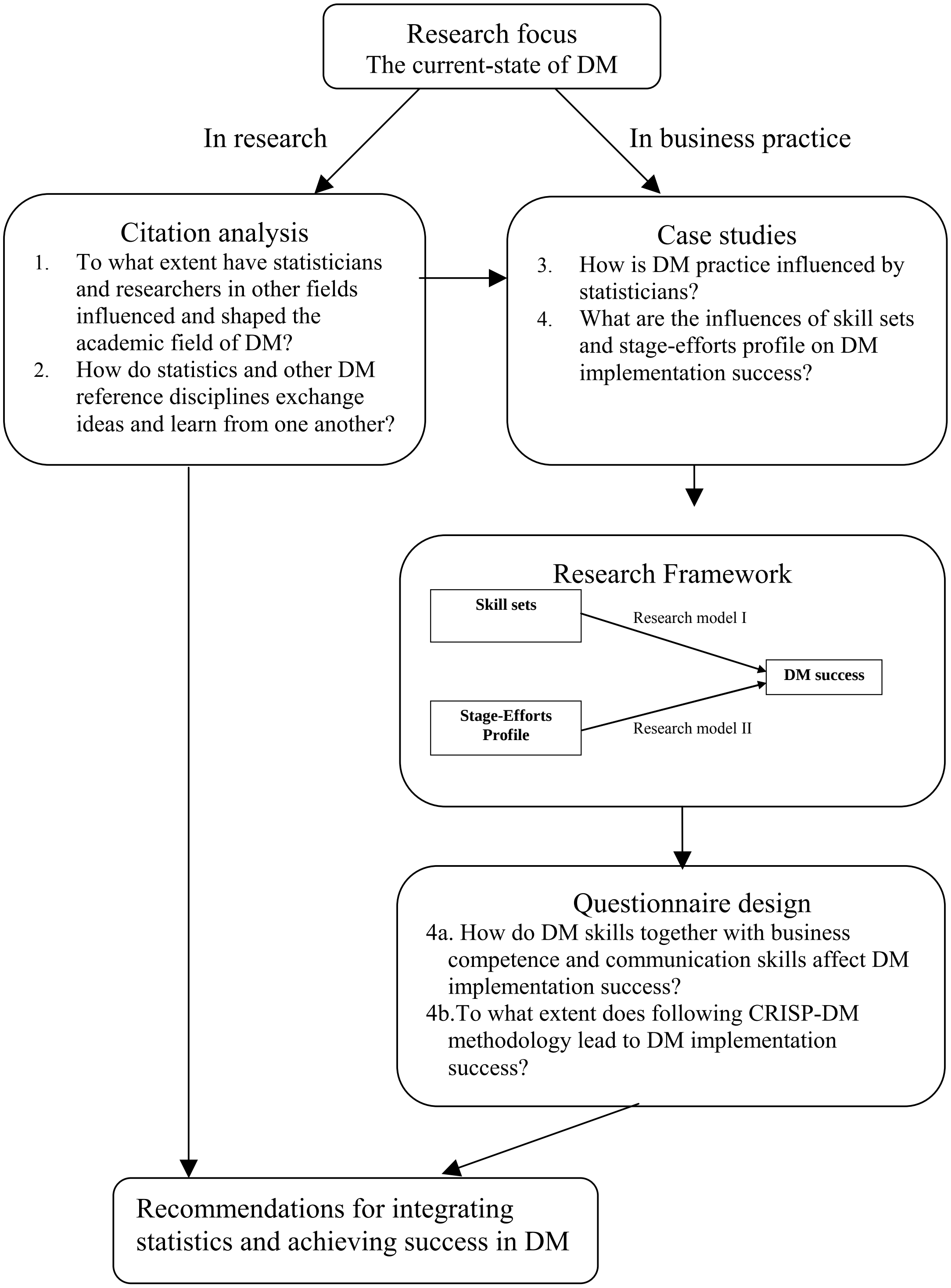

Research focus
The current-state of DM
In research
In business practice
Citation analysis
1. To what extent have statisticians and researchers in other fields influenced and shaped the academic field of DM?
2. How do statistics and other DM reference disciplines exchange ideas and learn from one another?
Case studies
3. How is DM practice influenced by statisticians?
4. What are the influences of skill sets and stage-efforts profile on DM implementation success?
Research Framework
Skill sets
Research model I
DM success
Stage-Efforts Profile
Research model II
Questionnaire design
4a. How do DM skills together with business competence and communication skills affect DM implementation success?
4b.To what extent does following CRISP-DM methodology lead to DM implementation success?
Recommendations for integrating statistics and achieving success in DM

Figure 3.2 Framework of Data Mining Research and Practice

CHAPTER 4

CITATION ANALYSIS OF DATA MINING RESEARCH

The DM field being multi-disciplinary in nature has been influenced by various reference disciplines. The collaboration and communication among these reference disciplines continuously have shaped the evolving field of DM. Author co-citation analysis and journal citation analysis are the two citation analysis techniques that are used to assess the intellectual structure of DM and the dynamic interchange of ideas among DM reference disciplines in the citation environment of DM.

The rationale behind using citation analysis methodologies is to answer the first and second research questions, namely, "To what extent have statisticians and researchers in other fields influenced and shaped the academic field of DM?" and "How do statistics and other DM reference disciplines exchange ideas and learn from one another?"

4.1 Introduction of Citation Analysis Methodologies

In Author Co-citation Analysis (ACA), the unit of interest is authors, who are placeholders for concepts that they champion through their writings (Culnan 1986; Culnan 1987; White and McCain 1998; Ponzi 2002). The intellectual structure of a field often comprises research strands or subfields that are influenced by a key set of authors. The foundations and core principles of each specialty within a discipline are determined by the informal and formal communications that occur between different authors. These authors may communicate with each other through informal interactions, co-authorship, or by referring to and building on their ideas. They enhance the body of knowledge, and thereby the field, by borrowing from the work of other authors and acknowledging this fact in the form of citation made to them. An examination of citation data helps elucidate the patterns of communications and the clusters (i.e. subfields) that these authors tend to form with each other (Crane 1972; White and Griffith 1981; McCain 1990). ACA has been used to address questions pertaining to: a) the evolution of thought and the emergence of revolutionary ideas within a field, e.g., Sircar, Nerur et al. (2001); b) the differences between research traditions in a discipline, e.g., Cottrill, Rogers et al. (1989); and c) the conceptual foundations of a field of interest, e.g., Culnan (1986), Culnan (1987), White and McCain (1998), Ponzi (2002). Thus, it is an appropriate methodology for unraveling the intellectual structure of DM in terms of its constituent specialties and the ideational relationships between them. In particular, it helps discern areas of research where statisticians have made significant contributions, while identifying those areas of research where there is a potential for bringing in statistical techniques.

In contrast to ACA, Journal Citation Analysis (JCA) uses the intricate patterns of communication that occur among journals as the basic block to understand the creation and diffusion of knowledge (Crane 1972; Zinkhan and Roth 1992; Pieters and Baumgartner 1999). Journals continually expand the frontiers of research either by creating new knowledge or by extending concepts that have been espoused in the literature. Citations of articles that appear in journals are systematically chronicled in indexes that can be readily accessed to obtain the number of citations they send to and receive from others. These data allow researchers to detect the underlying dynamics of the journal interactions and the level of influence that they exercise on each other as well as on the entire network (Pieters and Baumgartner 1999; Pieters and Baumgartner 2002). JCA has been very useful in assessing the prominence accorded to journals in a network, as well as in understanding the mutual influence between related disciplines (Pieters and Baumgartner 1999; Zinkhan and Leigh 1999; Pieters and Baumgartner 2002; Biehl and Kim 2005). Thus, a citation analysis of a network of journals meets one of our primary objectives in understanding the extent to which statistics and other DM reference disciplines, inspiring the evolving research traditions in DM, communicate with each other.

One assumption implicit in citation analysis methodology is that the frequency of citation/co-citation is a measure of prominence/ proximity for subjects under study.

4.2 Research Hypotheses

As mentioned earlier, DM is a multidisciplinary field incorporating database systems, Artificial Intelligence (AI) / Machine Learning (ML), pattern recognition, and statistics. The researchers in these disciplines pursue research in various topics within DM and borrow validated ideas and knowledge from each other, thus enabling DM to evolve into a new research field. Ever since its inception, the main thrust of DM research largely is on optimizing or developing new algorithms/techniques to make voluminous data processing more efficient and effective, an area in which researchers in AI/ML and database systems have been actively working. Statisticians with a wealth of experience they have acquired in using statistics for formulating and solving business problems, also have a great deal to offer to DM in terms of providing theories and methodologies, applying the techniques and evaluating the search (Glymour, Madigan et al. 1996). However, the slow recognition/ acceptance of DM as a new research area within the statistical community greatly inhibits statisticians from entering the DM field. In particular, as noticed earlier by Friedman (1997), "Statistics, as a community, has not been actively pursuing data mining, even though a few statisticians have contributed a bunch to data mining". Thus, we hypothesize

H-C1: Top DM researchers are more likely to have primary expertise in computer science or other related academic fields than in statistics.

It should be also noted that the multidisciplinary nature of DM has posed the potential that researchers from a variety of disciplines, such as pattern recognition, ML, and database field, make advancement in DM but are unaware of each other's progress. However, a collaborative effort of these disciplines will definitely add more value to the research in DM, and communication among these disciplines is very important. These disciplines can utilize the strength and overcome the weakness of each other. Statistics being a well-established field, if well integrated with other fields, can provide theoretical foundation and methodologies for these areas, and further enhance the cause of development of the DM field. This also helps statisticians to gain more experience by solving new problems, hereby developing statistics' own intellectual competence and body of knowledge. The more statistics communicates with others fields, the more it is beneficial to statistics itself, to other fields, and to DM research. To quantify the extent of communication, a measure of influence is used with a threshold value of 20% (In the citation analysis literature, researchers use arrowed line to connect two fields starting with the field doing the citation and ending with the field being cited. Only a percentage of citations to one another (after adjusting for the self-citations) of 20 percent or more is considered to be significant (Pieters and Baumgartner 2002). This leads to the following research questions:

H-C2: The number of references to statistics in the pattern recognition field after adjusting for self-citation is greater than the threshold influence.

H-C3: The number of references to statistics in AI/ML field after adjusting for self-citation is greater than the threshold influence.

H-C4: The number of references to statistics in the database field after adjusting for self-citation is greater than the threshold influence.

4.3 Data Collection

The data collection for both author co-citation analysis and journal citation analysis follows the standard methodologies of collecting the citation frequency using a bibliometric database, ISI's Web of Knowledge (http://isiknowledge.com), which includes the Science Citation Index-Expanded (SCI-Expanded) and the Social Science Citation Index (SSCI).

4.3.1 Data Collection for Author Co-citation Analysis

The first step in ACA is to choose authors who have made significant contributions to the DM literature. The construction of such a list of authors was conducted in two stages to assure as much possible that the right authors, whose investigations really form the basis of current DM research, were chosen. First, a search of the entire database for articles with key phrases like “data mining”, and “knowledge discovery” was conducted (other phrases were also tried; however, these two phases seem to give enough information). Then references were retrieved and the number of times each author was cited for each article was counted. Using ninety citations as the benchmark, we derived a list of the top sixty authors. To avoid the possible omission of some authors, the procedure was repeated specifically for the two journals: *IEEE Transactions on Knowledge and Data Engineering* (TKDE) and *Data Mining and Knowledge Discovery* (DMKD). These two journals are chosen because TKDE has a section on DM/knowledge discovery, and DMKD is entirely dedicated to DM and knowledge discovery as suggested by its title, signifying their importance in publishing DM research. A benchmark of seven citations lead to a total set of thirty four authors publishing DM articles in these key journals. Combining two set of authors lead to a total of seventy authors since twenty four authors were common to both lists. After a careful review of the list, one author was eliminated since he/she was cited because people used the database maintained by him or her. The inclusion of such an author is clearly not appropriate. Finally sixty nine authors were obtained for initial analysis.

The second step is to retrieve co-citation data. For each of the identified 69 authors, cited references were counted for the years from 1996 to 2004. A co-citation occurs between two authors, say A and B, when a citation of both the authors is made together by an author C. Then a shell script was used to obtain the co-citation counts, i.e. the number of papers that have cited the works of both authors for each author pair. An initial symmetric matrix of 69 x 69 with authors’ names on the rows and columns was constructed (sample co-citation matrix is reported in Table 4.1 and a full data matrix is reported in Appendix A). In Table 4.1, each cell except the diagonal represents the co-citation count of the row author and the column author. The diagonal values of the matrix are computed by summing the three highest co-citation counts for each author and dividing the result by two.

Table 4.1 Sample Co-citation Matrix

	Quinlan	Savasere	Scholkopf	Srikant	Toivonen	Vapnik
Quinlan	870.5					
Savasere	14	153.5				
Scholkopf	30	0	353			
Srikant	44	86	1	254		
Toivonen	30	75	3	98	241	
Vapnik	159	0	552	2	5	566

The third step is author refinement. To increase the robustness and stability of the results, mean co-citation counts were computed and only authors with average co-citation counts of eight or above for nine years of data were retained for final analysis (McCain 1990). This narrows the list to 51 authors. The authors chosen are clearly not exhaustive of the individuals who have contributed to or are currently active in the DM field but should be a good representative sample. Table 4.2 provides an alphabetical listing of the 51 authors together with their mean co-citation counts.

Table 4.2 51 Seminal Authors in Data Ming Literature (1996-2004)

Author	Count	Author	Count
Agrawal, Rakesh	71.9	Langley, Pat	34.7
Bayardo, Roberto J.	14	Lavrac, Nada	16.3
Brachman, Ron J.	11	Mangasarian, Olvi. L.	11.4
Breiman, Leo	84	Mannila, Heikki	33.6
Brin, Sergey	22.6	Michalski, Ryszard S.	44.8
Buntine, Wray*	19.3	Mitchell, Tom M.	21.9
Cheung, David Wai-lok	12.8	Muggleton, Stephen	29.4
De Raedt, Luc	14.3	Ng, Raymond T.	19.3
Dietterich, Thomas G.*	38	Pawlak, Zdzislaw	33
Domingos, Pedro*	20.9	Pearl, Judea	34.1
Duda, Richard O.	60.2	Piatetsky-Shapiro, Gregory	20.4
Ester, Martin	10	Provost, Foster	16.7
Fayyard, Usama M.	66.3	Quinlan, Ross J.	127
Feldman, Ronen	8.02	Rumelhart, David E.	53.5
Fisher, Douglas H.*	15.2	Salton, Gerard	20
Frawley, Williams J.	10.1	Savasere, Ashok	16
Friedman, Jerome H.	11.7	Scholkopf, Bernhard*	21.3
Goldberg, David E.	8.76	Skowron, Andrzej	15.3
Grzymala-Busse, Jerzy W.	12.9	Srikant, Ramakrishnan	28.7
Han, Jiawei	27	Toivonen, Hannu	24.6
Hand, David J.	18.7	Vapnik, Vladimir N.	49.3
Heckerman, David	21.9	Witten, Ian H.	18.9
Holte, Robert C.*	20	Zadeh, Lotfi A.	37.3
Jain, Anil K.	9.84	Zaki, Mohammed J.	17.2
Kohavi, Ronny	45.4	Ziarko, Wojciech	17.7
Kohonen, Teuvo	46.9		

* Authors were chosen based on the citations in DMKD and TKDE.

4.3.2 Data Collection for Journal Citation Analysis

To define the social network of journals, papers in DMKD, papers in TKDE under the data mining/knowledge discovery sections, and special issue papers on DM/KDD in *Journal of Intelligent Information Systems* were used to suggest the answers. A reference analysis of articles on DM with a key phrase search for the terms "data mining" and "knowledge discovery" suggests that these journals are the top three in the field, having published DM/KDD special issue papers from 1997-2004.

Then references for all these selected papers were retrieved to get the count of citations received by various journals occurring in these references. A benchmark of 19 citations was used to identify the list of 17 journals for the analysis, as shown in Table 4.3 (all other journals have a count below 15). As a magazine, *Communication of the ACM* (CACM) is included for this study due to its high standing and the availability of citation data (indexed in the database). *IEEE Transactions on Systems, Man, and Cybernetics* (TSMC) split into three journals in 1996. Since it is reasonable to assume that these three journals are going to attract similar authors and audiences, this journal was retained for the analysis, and the citations of the three journals were combined to represent TSMC after 1996.

The frequency of citation flows among the journals in the network was then obtained. These flows were computed for each journal in the list by examining citation data for all articles that are indexed between 1997 and 2004. All the journals are indexed during the chosen time period. Further, to allow for trend analysis, the citation data was collected separately for the two time periods, viz.: 1997-2000, 2001-2004.

Table 4.3 The Social Network of Journals Related to Data Mining

ACM Transactions on Database Systems (TODS)
Annals of Statistics (AOS)
Artificial Intelligence (AI)
Communications of the ACM (CACM)
Data Mining and Knowledge Discovery (DMKD)
IEEE Transactions on Computers (TCOMP)
IEEE Transactions on Information Theory (TIT)
IEEE Transactions on Knowledge and Data Engineering (TKDE)
IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)
IEEE Transactions on Systems, Man, and Cybernetics (TSMC)
Journal of Artificial Intelligence Research (JAIR)
Journal of the American Statistical Association (JASA)
Journal of the Royal Statistical Society: Series B (Statistical Methodology) (JRSS-B)
Machine Learning (ML)
Neural Computation (NC)
Pattern Recognition (PR)
VLDB Journal (VLDB)

Since the exchanges are typically unequal because citations between journals are not necessarily reciprocated, the citation matrix is asymmetric. An extraction of the asymmetric matrix showing the frequency of citations that a journal sends to and receives from others is presented in Appendix B. In Appendix B, the diagonals represent self-citations (i.e. the number of times a journal cites itself). The rows are the journals that are sending citations and the columns represent journals that are receiving citations, and each cell in the table is the frequency of citations that a row journal sends to a column journal.

4.4 Analysis and Results

4.4.1The Intellectual Structure of Data Mining

Three multivariate statistical techniques were employed to analyze the intellectual structure of citation data matrices and thereby describe DM. They are factor analysis, cluster analysis, and Multidimensional Scaling (MDS).

4.4.1.1 Factor Analysis

A principal component factor analysis with oblique factor rotation was performed. The Kaiser criterion (eigenvalue>1) was used to select the number of factors, and a benchmark of ?0.4 was used for factor loadings. Ten research specialties resulted, and all authors were loaded on one or more factors. These ten factors account for 82.3% of the variance with the first two factors accounting for about 45%. The factors were named based on a general assessment of the research areas represented collectively by the individuals with a loading of at least 0.7 on each factor (McCain 1990). The inferred ten DM research subspecialties are reported in Table 4.4.

Factors 1 and 2 represent two major tasks of DM in prediction and description. Of 51 authors, 32 were loaded on one of these two factors. The first factor appears to define the work on prediction & classification from both the machine learning and statistical communities and is considered supervised DM. The majority of this work focuses on either regression, if the predicted variable is numeric, or classification, if the predicted variable is categorical. Friedman, Breiman and Hand are representative statisticians; their work on Classification and Regression Trees (CART) and discrimination and classification is still in use. Kohavi and Quinlan represent the machine learning research community with work on feature subset selection and the famous work on ID3 and C4.5 decision tree algorithms respectively.

Factor 2 has a strong database and pervasive computing flavor, focusing primarily on association rule and data summary with an emphasis on efficiency and scalability. Some major DM concepts came from this group, such as privacy preserving DM, fast algorithms in knowledge discovery for large data sets, novel techniques in association rules and privacy preserving. The authors in this group collectively represent the significant role of database researchers in DM; their work is considered unsupervised DM.

Factors 3 to 9 represent important works on theoretical methodologies and theory development in DM/KDD. Factor 3 deals with rough sets about reasoning under uncertainty, knowledge representation, loading on Pawlak, Ziarko etc. Factor 4 has a strong neural networks orientation, represented by Rumelhart and Kohonen. Factor 5 covers authors who have made substantial contributions to causal modeling / Bayesian Networks. Factor 6 represents foundation work/theory development on DM/KDD, dominated by researchers from the AI/ML community, but with a synergy with database communities. It is of particular interest that statisticians are not a significant contributing factor for DM/KDD in terms of theory development, even though the ideas of statistics and statistical thinking are critical components in successful DM/KDD (Goodman 2000). It is worth to point out that Piatetsky-Shapiro is credited with the creation of the term “Knowledge Discovery in Databases”. Factor 7 includes research on inductive logical programming, represented by Muggleton, who introduced the concepts of inductive logical programming.

Table 4.4 Contributing Research Subspecialties

Factor name	Prediction & classification		Data summarization		Rough sets		Neural computation	
Eigenvalue	12.314		10.518		4.571		3.732	
Explained variance	24.145%		20.623%		8.963%		7.317%	
Loading	KOHAVI	.90	SRIKANT	.94	PAWLAK	.98	RUMELHART	.95
	QUINLAN	.90	AGRAWAL	.94	ZIARKO	.95	KOHONEN	.95
	BREIMAN	.89	MANNILA	.91	SKOWRON	.94	DUDA	.81
	HOLTE	.87	ZAKI	.90	GRZYMAL-ABUSSE	.92	ZADEH	.74
	DOMINGOS	.86	SAVASERE	.90	ZADEH	.46	GOLDBERG	.64
	PROVOST	.82	BRIN	.90			SALTON	.50
	DIETTERICH	.82	TOIVONEN	.89			HAND	.44
	LANGLEY	.77	BAYARDO	.87				
	WITTEN	.73	CHEUNG	.84				
	FAYYAD	.71	NG	.72				
	HAND	.69	HAN	.71				
	MICHALSKI	.67	FELDMAN	.71				
	MITCHELL	.63	PIATETSKY-SHAPIRO	.57				
	BUNTINE	.55	ESTER	.40				
	MUGGLETON	.47						
	FISHER	.46						
	FRIEDMAN	.42						
	DUDA	.41						

Factor name	Bayesian networks		DM/KDD foundation		Inductive logic programming	
Eigenvalue	2.722		2.152		1.897	
Explained variance	5.337%		4.220%		3.720%	
Loading	PEARL	.97	FRAWLEY	.84	MUGGLETON	.98
	HECKERMAN	.95	HAN	.75	LAVRAC	.93
	BUNTINE	.78	PIATETSKY-SHAPIRO	.73	DERAEDT	.91
	ZADEH	.48	ESTER	.72	MICHALSKI	.74
			FAYYAD	.71	QUINLAN	.61
			BRACHMAN	.69	LANGLEY	.53
			NG	.66	BREIMAN	.49
			AGRAWAL	.59	MITCHELL	.46
			FELDMAN	.49	DIETTERICH	.45
			SRIKANT	.44	FAYYAD	.44
			CHEUNG	.42	HOLTE	.44
			FISHER	.40	FISHER	.44
					PROVOST	.41

Table 4.4 – *Continued.*

Factor name	Support Vector Machine		Quantitative learning		Information retrieval	
Eigenvalue	1.618		1.288		1.168	
Explained variance	3.173%		2.525%		2.290%	
Loadings	VAPNIK	.93	JAIN	.86	SALTON	.87
	SCHOLKOPF	.93	FRIEDMAN	.79	MITCHELL	.57
	MANGASARIAN	.82	GOLDBERG	.51	WITTEN	.57
	DUDA	.53	DUDA	.40		
	FRIEDMAN	.41				

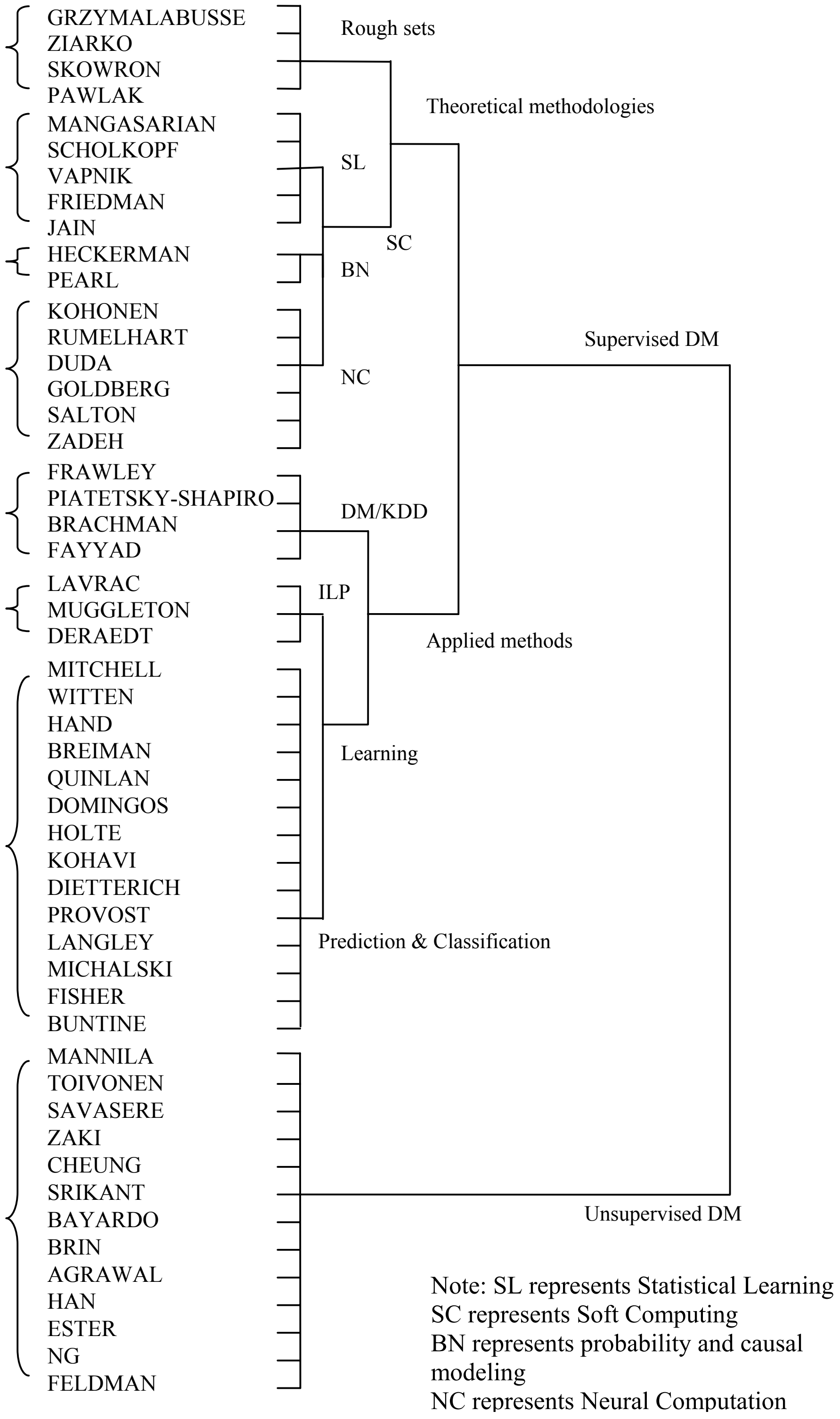
Factor 8, on support vector machine (SVM) or kernel learning, is represented by Vapnik, who is credited with the creation of "Statistical Learning Theory" (1998) and the support vector machine, and Scholkopf, who contributes substantially to SVM literature by introducing novel algorithms, such as the kernel-based learning algorithm. Most authors who are loaded on this factor have a strong background in mathematics or statistics. Factor 9, represents a family of quantitative learning techniques on classification, clustering, and pattern recognition, represented by Jain and Friedman. As a statistician, Friedman is well known in both statistics and DM fields for a body of work including classification and regression trees, multivariate adaptive regression splines (MARS) and project pursuit regression. Methods from this factor complement predictive modeling in solving supervised DM problems.

Finally, the grouping of Salton, Mitchell, and Witten is somewhat surprising. Salton worked mainly on information retrieval; Mitchell and Witten researched machine learning/ text mining. A detailed examination of literature that cites each author paired with Salton suggests that the citers have already borrowed the advanced techniques from other fields and applied them to information retrieval.

4.4.1.2 Cluster Analysis

Consistent with ACA literature, cluster analysis based on a hierarchical agglomerative method was performed. The result from an eight-cluster solution using a cut-off distance value of 1.0 is reported in Figure 4.1.

Two final clusters emerged: one is unsupervised learning, another is supervised learning. For unsupervised learning, research primarily focuses on association rules mining, with a strong flavor of algorithms and a high emphasis on efficiency and scalability as actively pursued by database researchers. For supervised learning, research primarily focused on machine learning (including classification and regression tree, and inductive logic programming (ILP)), together with DM/KDD theory and theoretical methodologies (including rough sets, neural computation (NC), probability and causal modeling (BN), and statistical learning (SL), where the later three are grouped into soft computing (SC)). This is a cluster that is occupied by statisticians and AI/ML researchers.



Note: SL represents Statistical Learning
SC represents Soft Computing
BN represents probability and causal modeling
NC represents Neural Computation

Figure 4.1 Research Streams in Data Mining

4.4.1.3 Multidimensional Scaling

The two-dimensional map representing the conceptual structure of the field was derived using the ALSCAL non-metric MDS procedure in SPSS, the Euclidean distance model, and a cut-off value of -1.0 for missing values (in order to accommodate negative correlation values), as reported in Figure 4.2. The boundaries around authors indicate clusters, the composition of which was identified through cluster analysis. Both the stress value and the percentage of total variance explained measures were used to measure the "goodness of fit". With a stress value of 0.098 (<0.2) and a percentage of total variance explained of 96.018%, the two-dimensional solution suggests a good fit of the data.

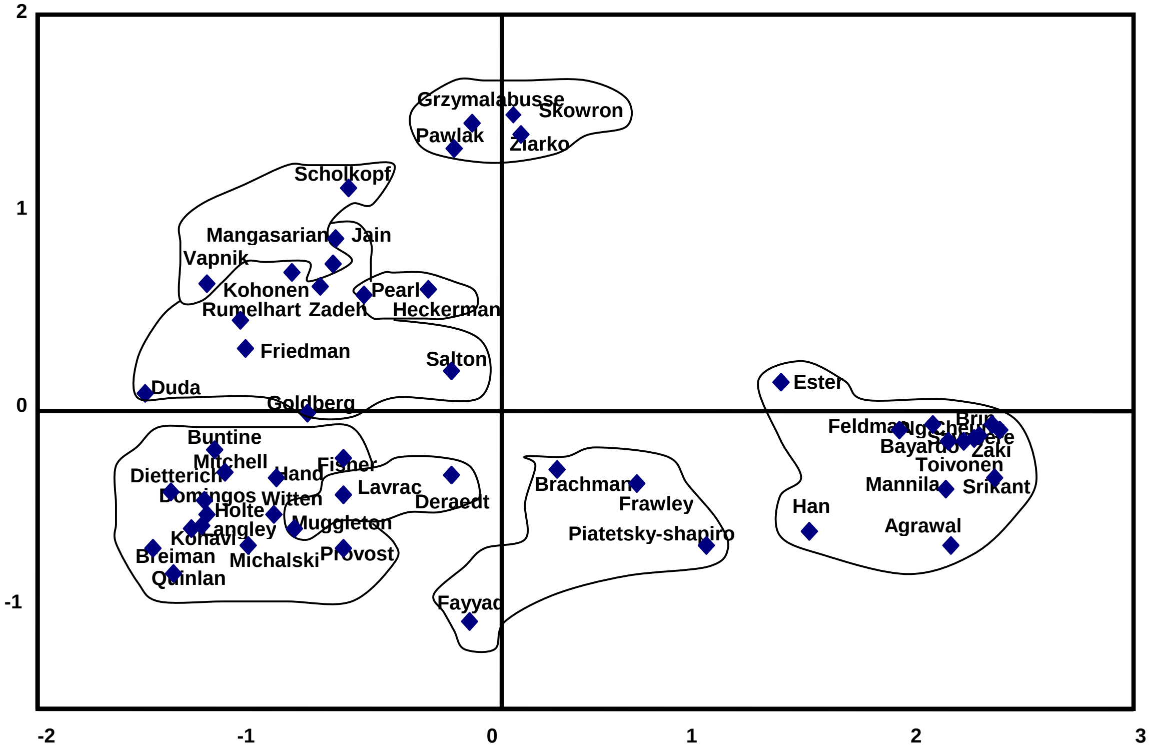


Figure 4.2 MDS Map on Co-citation Analysis

In Figure 4.2, the clusters are oriented along a horizontal and a vertical dimension. The horizontal axis appears to separate authors into supervised and unsupervised learning, moving from left supervised learning to right unsupervised learning. The vertical axis seems to separate pure theoretical methodologies and applied methods, moving from top pure theoretical methodologies to bottom applied methods. Noticeably, three statisticians, Breiman, Friedman, and Hand, and two researchers with a strong mathematical/statistical background, Vapnik and Mangasarian are all placed in the clusters on the left-hand side, indicating the substantial contribution of the theoretical methods from the discipline of statistics to prediction & classification. The majority of machine learning researchers fall into the lower-left quadrant, database researchers fall into the lower-right quadrant, while researchers in the AI discipline mainly fall into the upper-left quadrant, or as a bridge linking machine learning researchers with database researchers. These findings support the multidisciplinary nature of DM and its role in bridging previously unlinked disciplines, such as machine learning and database.

A casual look at the MDS map again shows the absence of pure theoretical methodologies in unsupervised DM, as demonstrated by the almost blank area in the upper-right quadrant. Statisticians fail to appear in unsupervised DM, indicating a need for more cross-fertilization of ideas between statistics and database; this poses rich opportunities for researchers from both of these areas. These gaps also provide important insights for researchers who joined late, or will join later, about the rich set of potential research directions.

4.4.2 The Dynamics in Data Mining

This section is to examine the dynamics in the knowledge network of DM. In particular, it examines the similarity of these journals, and how the journal clusters exchange and borrow from each other. To derive journal clusters, consistent with the literature (e.g., Pieters and Baumgartner (1999) and Pieters and Baumgartner (2002)), I used the dimensional values for the journals from an estimated log-multiplicative model as inputs to an agglomerated hierarchical clustering procedure. To discern the change of the perceived similarity among journals, the same procedure was repeated separately on the two time periods.

The following log-multiplicative model, as proposed in Pieters and Baumgartner (1999) and Pieters and Baumgartner (2002) as estimated:

$$\log F_{sr} = u + u_s{}^S + u_r{}^R + \delta_{sr} + \sum_{m=1}^{M} \xi_s{}^m \psi^m \xi_r{}^m$$

The expected cell frequency, F_{sr}, is expressed in terms of a constant (u in the equation), log-linear parameters (represented by the $u_s{}^S$ and $u_r{}^R$ terms), self-citation denoted by δ_{sr}, and the mutual influence that the journals have on one another (expressed by $\sum_{m=1}^{M} \xi_s^m \psi^m \xi_r^m$). This model has a sound statistic foundation. For a detailed view of this method, see Goodman (1991) and Clogg and Shihadeh (994). The log-linear parameter $u_r{}^R$ can be used to assess the relative prominence of a journal in its citation network. To determine the number of dimensions (i.e. the dimensions of association between sending [S] and receiving [R]) that would be the “best fit” for the data, Bayesian Information Criterion (BIC) (Pieters and Baumgartner 1999; Pieters and

Baumgartner 2002) was used. Log-multiplicative models with four and five dimensions of association between R and S were found to be the most appropriate fit for the two time periods, respectively.

Then an agglomerated hierarchical clustering procedure was performed on the dimensional values for the journals from the estimated log-multiplicative model. The cluster analysis results in Figure 4.3 indicate the cohesion of the journals in the DM communication network. The result shows that the three statistical journals (ASA, JRSS-B, and AOS) consistently form one cluster and all computer science journals are grouped into AI, database, general computer science, machine learning, and pattern recognition clusters. It seems that the grouping is reasonably stable over time as evident from the structures of the two clusters. However, a shift of DMKD from AI/ML to the database field occurs. This finding reflects the origins of DM/KDD in AI/ML, and its having evolved into the database field. A new alignment of patterns with the database field also has taken place. The result also suggests TSMC always clusters with TPAMI/PR.

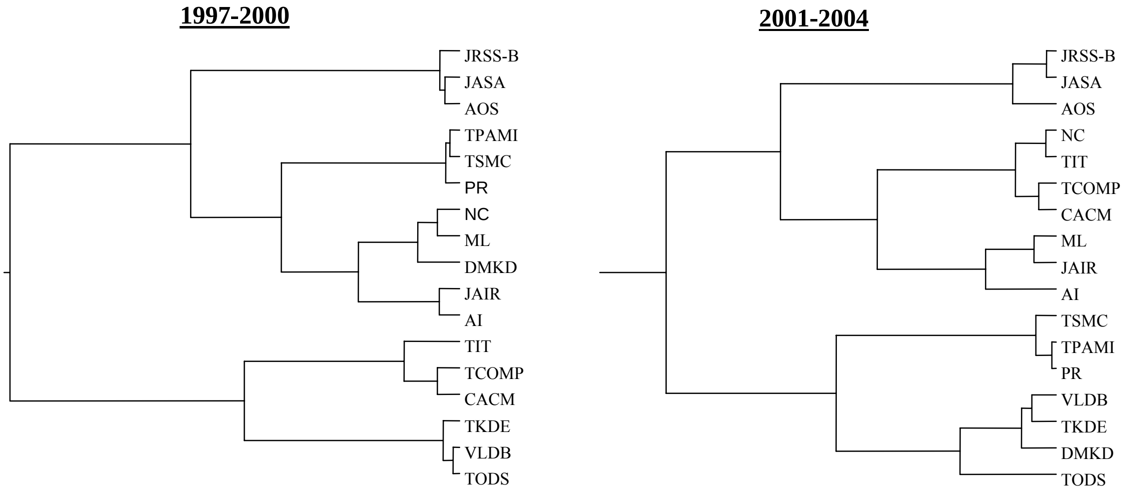


Figure 4.3 Cluster Analysis

Now let’s focus on the examination of the citation flows among journal clusters. As mentioned earlier, the journals chosen are representative of the fields they belong to; they are all top journals in their fields. Based on the cluster analysis results and an examination of the aims and scope of each journal, five reference disciplines to DM were defined (the same for both time periods in order to perceive the trend). The journals representing statistics are AOS, JASA, and JRSS-B. TPAMI, PR, and TSMC represent pattern recognition. VLDB, TODS, and TKDE represent the database field. TCOMP, CACM, and TIT represent general CS. AI, ML, NC, and JAIR represent AI/ML, while DMKD is intentionally separated from the above disciplines to see more clearly the exchange of ideas among the five reference disciplines. The citing and cited data of journals belonging to the same reference discipline were then aggregated to obtain inter-cluster (cross discipline) citation flows for each time period. The detail citation flows are summarized in Table 4.5, where numbers in parentheses indicate the number of journals in each cluster.

Table 4.5 Citation Flows between Clusters of Journals

	Patterns	Statistics	AI/ML	DB	CS	DMKD	Frequency of citing
1997-2000							
Patterns (3)	7205	311	816	88	778	1	9199
Statistics(3)	95	7247	44	7	118	0	7511
AI/ML(4)	405	258	3828	49	373	15	4928
DB(3)	124	14	193	814	344	6	1495
CS(3)	79	279	71	49	5628	0	6106
DMKD(1)	29	24	71	32	23	9	188
Frequency of being cited	7937	8133	5023	1039	7264	31	29427
2001-2004							
Patterns (3)	9081	467	1110	122	835	51	11666
Statistics(3)	79	5923	91	0	99	5	6197
AI/ML(4)	447	383	3693	60	302	48	4933
DB(3)	221	45	295	922	296	36	1815
CS(3)	95	285	78	74	8017	6	8555
DMKD(1)	18	34	68	27	32	29	208
Frequency of being cited	9941	7137	5335	1205	9581	175	33374

Note: Patterns represents the Pattern Recognition field; AI/ML represents the Artificial Intelligence/Machine Learning field; DB represents the Database field; CS represents General Computer Science areas

To allow for easy visualization, the data were further transformed and presented in Figure 4.4, where the numbers shown in parentheses in each ellipse are the percentages of self-citations to all citations; an arrowed line connects two clusters, starting with the cluster doing the citing and ending with the cluster being cited; and numbers shown along each arrowed line denote a percentage of citations to other journal clusters after adjusting for the self-citations. A percentage is reported only if it is at least 20 percent (Pieters and Baumgartner 2002).

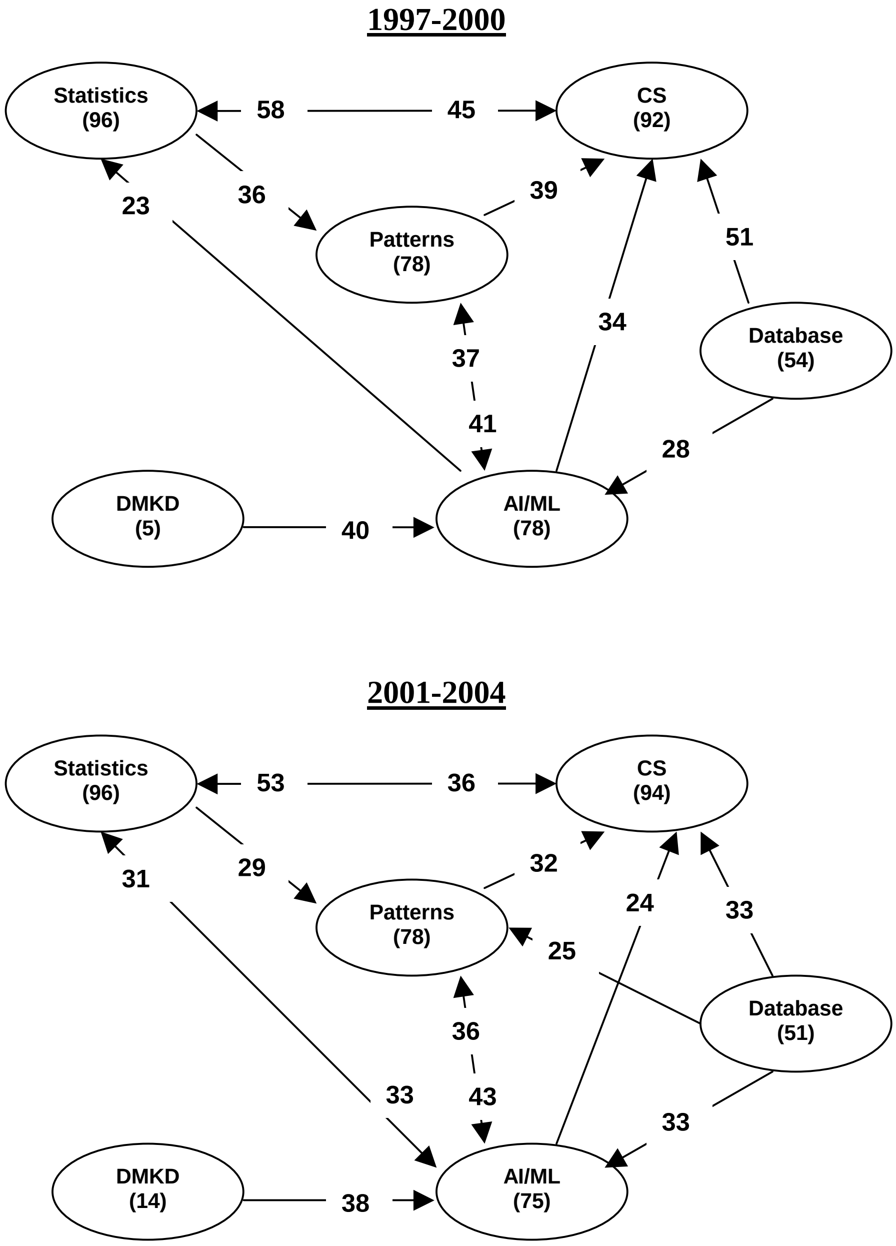


Figure 4.4 Citation Flows in the Data Mining Network

A few points are evident from the comparison of citation flows for the two time periods. First, statistics and the database field have only a limited exchange of ideas. This is true for both time periods. Second, the exchange between statistics and the AI field goes from one direction (AI to statistics) to two directions. Third, the communication between pattern and AI/ML is sound and stable over time. Fourth, the relative impact of CS seems to be decreasing, even though it is still heavily cited by other fields. Interestingly, CS cites statistics at a rate greater than twenty percent. However, a detailed examination suggests many of these citations were between TIT and AOS, which means that an absence of TIT could possibly change the citation pattern between CS and statistics. Therefore, readers should be careful when interpreting this finding. Fifth, the self-citation of each discipline is essentially stable over the time periods with the notable exception that the self-citation of DMKD increases from 5% to 14%. As a result, the relative percentages of citations to each of its reference disciplines decrease. In particular, the citation in DMKD to pattern recognition journals shows a 6% decrease. There is one exception to this: in the second time period, the citation in DMKD to statistical journals shows a rise of 6%. Sixth, there are very limited citations to the database field from other disciplines. The results also suggest that AI highly influences DMKD in both time periods. Other disciplines are also necessary components for DMKD to mature; citations from DMKD to these disciplines range from 10-19%. No discipline cites DMKD beyond 5%. However, all five reference disciplines' citations of DMKD slightly increase in the second time period.

4.4.3 Hypotheses Testing

A classification of these authors based on their primary expertise shows that among these 51 authors, three are identified as statisticians/mathematicians and two others hold a Ph.D in Mathematics or Statistics but currently work in a CS department, one is a mathematical psychologist, 13 are identified as database researchers, and all others fall into AI/ML areas with backgrounds in engineering, computer science, and physics. To conduct the first hypothesis test,

$$H_0 : P \leq 0.5$$
$$H_1 : P > 0.5$$

where P is the probability that a randomly selected researcher does not have primary expertise in statistics, an upper-tailed nonparametric binomial test is adopted. Given the total sample size is 51, the number of observed researchers having primary expertise in statistics/mathematics is 5, and the other 46 authors having primary expertise in other fields. The binomial test suggests that there is sufficient evidence the likelihood of top DM researchers having their primary expertise in computer science or other related academic fields is greater than the likelihood of DM researchers having primary expertise in statistics/mathematics (P-value = 1.2E-09). Thus, hypothesis HC-1 is supported.

To further examine the implication of this finding, using various web resources, information on the formal training for 39 authors was obtained (for 20 of them, we only know the discipline in which they got their highest degree). Of these 39 authors, 19 received early training in at least two disciplines. The information on their academic background is summarized in Table 4.6. Thirteen of the authors actually have training in both computer science and statistics.

Table 4.6 Frequency in Academic Training

Training	Counts
Computer Science	27
Statistics/Mathematics	14
Engineering	11
Physics	6
Mathematical Psychology	2

To conduct the tests for the second set of hypotheses HC-2-HC-4, we formalize the problem into a two factor design, thus the problem can be solved by using Bonferroni technique for multiple comparisons. As shown in Table 4.7, one factor is time with 2 levels, another factor is academic field with 3 levels. The numbers within each cell are obtained from the inter-cluster citation flows data in Table 4.5. Since the journals chosen are all top journals in their fields, it is appropriate to use these data as the cross discipline citation flows.

Table 4.7 Two Factor Cross Citation Results

	Pattern Recognition	AI/ML	Database systems
1997-2000	311/1994 (P_{11})	258/1100 (P_{12})	14/681 (P_{13})
2001-2004	467/2585 (P_{21})	383/1240 (P_{22})	45/893 (P_{23})

Note: within each cell, a/b can be interpreted this way: b is the number of total citations that a column field made to other fields; a is of b the number of citations that a column field made to statistics. P_{ij} is the expected cell proportion, where i=1 and 2, and j=1 to 3.

Thus simultaneous tests of the second set of hypotheses can be achieved by using the family of comparisons for the following hypotheses for each time period:

$$H_0 : P_{ij} \leq 0.2$$
$$H_1 : P_{ij} > 0.2$$

where P_{ij} is the expected proportion of discipline j in time period i that cites statistics (after adjusting for self citations), i =1 to 2 and j=1 to 3. Under the null hypothesis, all P_{ij} are less than or equal to 2.

I adopt a critical value approach to hypothesis testing with a simultaneous significance level of α = 0.05. Under the null hypothesis the sampling distribution of the sample proportion ($\hat{p}_{ij} = y_{ij}/n_{ij}$) has mean P_{ij} and standard error can be derived using the following formula:

$$\sigma_{\hat{p}_{ij}} = \sqrt{P_{ij}(1 - P_{ij})/n_{ij}}$$

The distribution is approximately normal since n_{ij} is relatively large ranging from 893 to 2585 and P_{ij} = 0.2. The test statistic can be derived using the following formula:

$$Z_{ij} = (\hat{p}_{ij} - P_{ij})/\sigma_{\hat{p}_{ij}}$$

and the critical value for this set of tests is $Z_{1-\alpha/6}$= 2.394. The hypothesis testing results are summarized in Table 4.8.

As shown in Table 4.8, hypothesis HC-3 is supported for both time periods, the reference to statistics in AI/ML field is above the threshold influence. However, both hypotheses HC-1 and HC-2 are not supported in either time period. But further examination shows that the magnitudes of reference from both pattern and DB fields to statistics increase from the first time period to the second time period. In particular, the pattern field has a citation to statistics that is close to the threshold value. DB field indeed rarely referred to the statistics field.

To capture the possible variation of the true proportion of each cell mean, a two-sided confidence intervals estimation technique was employed. Construct simultaneously 95% confidence intervals using the following formula,

$$\hat{p}_{ij} \pm Z_{1-\alpha/12}\sqrt{\hat{p}_{ij}(1-\hat{p}_{ij})/n_{ij}}$$

The results are summarized in the right-hand side of Table 4.8.

Table 4.8 Results of Statistical Inference

		y_{ij}	n_{ij}	$\hat{p}_{ij}$	$\hat{p}_{ij} - P_{ij}$	$\sigma_{\hat{p}_{ij}}$	Hypothesis Test	Estimation	
							Test statistics	LCL	UCL
1997-2000	HC-2	311	1994	0.156	-0.044	0.009	-4.916	0.135	0.177
	HC-3	258	1100	0.235	0.035	0.012	2.864*	0.201	0.268
	HC-4	14	681	0.021	-0.179	0.015	-11.707	0.006	0.035
2001-2004	HC-2	467	2585	0.181	-0.019	0.008	-2.459	0.161	0.201
	HC-3	383	1240	0.309	0.109	0.011	9.584*	0.274	0.343
	HC-4	45	893	0.050	-0.150	0.013	-11.177	0.031	0.070

* significant at 0.05 significance level.

4.5 Limitations

There are some limitations with the citation analysis methodology. First, the data in the Web of Science database are not comprehensive. Some journals, conferences proceedings such as ACM SIGKDD and VLDB, and other references related to DM were not incorporated into the database. This is inseparable from the immature nature of the field, as numerous other journals across various disciplines also publish DM/KDD articles, and the term “data mining” is loosely used across these disciplines. The results would be more robust if all journals, conferences proceedings, and other references related to DM were indexed in the Web of Science database, as this would yield a more accurate picture of research in DM.

Second, as with all other citation analysis research, the data are noisy. For example, in the determination of the co-citation counts in ACA, no matter what reasons two authors were cited together by a third author, the method assigns equal weight to all citations. In other words, all citations are considered to be equal. Another drawback with ACA involves the selection of authors. In the database, authors are indexed by surname and initials, and the authors compiled for this study are the first authors of cited references. There is one possibility that the result set may contain some spurious references. However, this problem is not serious and this potential gets substantially reduced when using mean co-citation counts to increase the robustness and stability of the results (McCain 1990). In addition, to avoid the confounding of results, the appropriateness of some authors, such as Friedman, Feldman, and Goldberg, were examined because there are well-known researchers in others areas with the same last name and first name initial, and their inclusion was found to be proper and valid. Another possibility is authors who were not in the first authorship were excluded; in the future, a study of co-author citation might complement this study.

To guarantee the quality of the data, I manually checked and verified references for some journals for their abbreviations. For example, sometimes, people may cite

JRSS-B as “J Roy Stat Soc”. Journal citation analysis did not take the number of articles published, the average citations per article, or the journal age into consideration. Despite these limitations, this study provides a first step toward prescribing the dynamics within the knowledge network of DM. The results provide a good approximation and yield insights that can significantly enrich our understanding of the conceptual underpinnings of the field.

CHAPTER 5

DATA MINING PRACTICES IN THE RETAILING INDUSTRY

Driven by real-world applications and business needs, DM is gaining ever increasing importance in businesses. Business leaders have realized that DM is an important enabling technique for organizations to pursue productivity, efficiency and effectiveness. There are many successful stories; however, there are sad stories too. Companies investing in these initiatives have already noted the complications involved in DM projects and the high risk of failure (Hotfmann and Tierney 2003). To gain insights into the relatively new phenomenon in business and the role of statistics in DM process towards a building of a general theory on DM implementation, a case study methodology is chosen. The rationale behind using case study methodology is to answer the third research question, namely, how DM practitioners do DM and how the practice is influenced by statistics and what DM practitioners perceive to be important for DM implementation.

5.1 Case Study Methodology

The case study research methodology is deemed appropriate to assess the extent that statistics has been integrated into DM practice and explore the factors that are perceived to be important by DM practitioners in DM implementation success. The study is exploratory in nature (Bonoma 1985; Behbasat, Goldstein et al. 1987; Yin 1994), and the phenomenon needs to be observed in a natural setting (Benbasat, Goldstein et al. 1987). In addition, a case study is useful for generating theories for research areas where little theory has been developed (Eisenhardt 1989). I incorporated suggestions of Eisenhardt (1989) in using a case study to build theory grounded in data during the implementation of this project. I started with no theory and no hypothesis to test.

5.1.1 Site Selection

The research setting *was* the retailing industry adopting a multiple-case design to follow a replication logic (Yin 1994). The retailing industry was selected because it is one of the first business areas to have adopted DM to better meet the needs of customers and make more informed business decisions, and this is an area with more widespread applications and fruitful outcomes. At the time of study, the retailing industry was undergoing substantial changes towards a more effective CRM.

I deliberately, but not randomly, selected two retailers (Glaser and Strauss 1967): one is information-oriented (thus with a technology and data analysis culture), another is not (thus with an anti-technology culture); one with high end merchandise (thus with fewer competitors in the market), the other not (thus facing severe competitiveness from outside environment). In addition to these variations, company size and customer size were also taken into consideration. The primary criterion that guided the selection of corporations was that they were attempting DM projects.

5.1.2 Data Collection and Analysis

Data were collected through semi-structured interviews. Each interview took approximately 1.5 hours. Given the importance of the case study protocol (Eisenhardt 1989), open-ended interview questions were carefully designed before each interview.

Slight changes were allowed during the interview. The initial protocol included constructs derived from the six stages of CRISP-DM and constructs that are perceived to be critical during the implementation of DM projects. The actual items for the role of statistics in a business DM context are more ad-hoc in nature, while there are some theoretical underpinnings for items that are perceived to be critical. Issues relating to the measures of DM success were also addressed. Before the start of the actual interviews, a pilot interview was performed with two DM academicians, who have extensive consulting experience with statistical analysis in various functional areas of business. Their feedback was used to refine the protocol. The detailed interview protocol is in Appendix C.

Each interview was conducted by two investigators who are statisticians: with one primarily responsible for the interview and the other for taking notes, filling in gaps in the questioning, and recording tapes. Each interview began with questions about the norms and history of an informant's organization. Interviews were tape-recorded, which helped to capture all data, regardless of its apparent importance at the interview. Immediately after an interview, the investigators cross-checked facts and shared their impressions. The tapes were carefully transcribed, and issues of disagreement between two researchers were resolved before the data were analyzed. Useful data from such records as conference presentations were also gathered.

Following suggestions in the literature (Bonoma 1985; Behbasat, Goldstein et al. 1987; Yin 1994), analysis was performed both within and across cases. This approach allows the researcher to examine the issues in a single context, as well as to replicate the findings from another. The within case analysis primarily focused on identifying research issues within each site, and the cross-case analysis aims to find emerging patterns.

5.2 Data Sources

The final project was carried out at two retailers. Company A is one of America's largest department store, catalog, and e-commerce retailers. The company has a large body of customers, a few hundred million. In contrast to company A, company B is a premier luxury retailer. Since the company markets luxury merchandise, the customer base is not big, a few hundred thousand. Company B has an anti-technology culture, as "*Nobody here uses any technology and nobody talks anything about statistics*", "*They do not see the numbers, they do not want to see the fact*". Compared to a more customer-centric business model in company A, company B is a more product-focused organization. As noted by the interviewee in company B, "*Our strength is our merchandise, not marketing. Marketing is number two; merchandise is definitely number one.*" The basic characteristics of the two companies are summarized in Table 5.1.

Table 5.1 Companies' Characteristics

	Company A	Company B
Number of employees	A few hundred thousand	A few ten thousand
Company culture	Data/Information-oriented	Fact-based culture
Business model	More customer-centric	More product-focused
Customers size	A few hundred million	A few hundred thousand
Competitiveness of outside environment	Severe	Mild

DM practitioners were our interviewees. The person we interviewed in company A is a senior statistical analyst and has primary expertise in statistical modeling. It is interesting to know that the person interviewed in company B though titled Statistician, really has an MBA with a focus on MIS. All the interviews were conducted within the same time period.

5.3 Case Analysis

Across the retailing industry, companies have come to realize that customers are central to their business and that customer information is one of their key assets. CRM challenges the retailing industry in several ways such as customer acquisition and retention, customer buying habits and preferences, and personalization. DM, as a fundamental component in business decision-making infrastructure, can assist retailers in transforming customer data into customer knowledge, which, in turn, assist in developing unique customer-firm relationships. In this section, I discuss how DM practitioners implement DM to fulfill the project goals. In particular, I illustrate their DM practice by following the CRISP-DM process model, examine what they perceive to be critical for project success, and how they reactively and proactively overcome the problems presented during the process.

To guarantee that both the informants are on the same page when talking about DM, first, I examine their understanding of DM. All informants consider what they do to be DM and their views of DM generally can encompass both types as reflected by these quotes:

Data mining, in a general sense is to use data to learn something.

Data mining is mining (the) database to keep searching for trend.

The first quote is from company A which takes a more directed approach to DM. The latter one is from company B where a more undirected approach to DM is taken.

5.3.1 Statistics in Data Mining Practices

All informants were excited to see their use of statistics transformed into DM success. As pointed out by one interviewee: "*With strong statistics or computer science background, it is so helpful in transforming data into information, into knowledge, making data actionable*." Table 5.2 summarizes successful DM projects involving statistics based on the issues addressed and the statistical techniques used. It is clear that statistics has been widely used across a variety of business activities in the retailing industry.

Table 5.2 Data Mining Projects and Techniques

Project	Techniques
Catalog response/return prediction	Logistic regression
Inventory planning/store placement	Fitting probability distributions
Size forecast in sales data	Data summary
Senior citizen account	Exploratory data analysis
Customer score card	Survey, time series analysis, exploratory
Do people who shop at multiple channels spend more?	Exploratory, statistical inference
Does free shipping cause higher return rate than 20% off promotion	Case control, experimental design
Shipping handling rate	Simulation
Which subdivisions are most similar based on the buying patterns of the customers?	Association analysis

Table 5.2 – *Continued*.

Promotion mailer	Association analysis
Customer segmentation	Data summarization, Collaborative filtering, experimental design
Revenue forecast	Regression

Our informants were also enthusiastic about the impact of their DM activities on their organization. With a large volume of customers on hand, one problem the retailers face is to whom the catalog should be mailed. The retailers expect to target a smaller set of customers, so that cost would be less and effort could be minimized, but at the same time, it still remains cost effective and increase revenue. Catalog response prediction addresses such an issue. The whole idea involves the use of a gain table, constructed using the scores from a logistic regression model. By doing this, customers are grouped based on the likelihood they will respond to the catalog. Customers with higher likelihood are assigned into top groups; thus they are the targets for the catalog mailing. In addition to cost cutting and improved response rates, this method significantly reduces the risk of irritating some customers not showing interest in the catalog. More importantly, revenue can be increased up to a few million dollars.

In addition to these measurable outcomes, our informants were glad to find out their DM activities have the capabilities to lead to a change of the policy in their organization. According to one interviewee, the high return rate associated with catalog orders used to be a big issue in the company. Earlier, when customers placed an order, they could choose either to have it sent to a local store or directly to their home. In many cases, if they chose to send it to a local store, they also could choose to pay cash when they came to pick up the order. But for those who never come to pick up the product, their order would just return by default. So, as the first step the DM practitioners tried to identify those people who are likely to return (who placed the order but did not pay for it or who have a high likelihood of never to come to pick them up and pay), so that they basically could avoid sending catalog to these customers. This was achieved by using the same catalog data and techniques as in the previous example, but instead of predicting who will buy, they predict who are likely to return. The model also takes this into consideration: customers who only made one order and happened to return it do not provide as much evidence as customers who made ten orders and returned them all. The model was used quite well to suppress people from getting catalogs if they score too high on the likelihood of returning merchandise. As a result of the DM practitioners' endeavors, the company eventually changed the whole return policy. For example, if the customers choose to pick up at the stores, they have to provide a credit card number or pay by check ahead of time.

5.3.2 Data Mining Process Characteristics

5.3.2.1 Business Understanding

Both DM practitioners agree, most of the time, they are given the problem. Before they start with the project, they need to understand their business to translate the original business problem into a detailed DM problem, defining the inputs, the results and the actions indicated by the results. With the strong need to understand CRM for their business, in their spare time, practitioners spend substantial time reading database/direct marketing books. They also resort to other sources, such as Internet site.

As one interviewee commented, "*I talk on the Internet a lot to professional people, involved a lot of CRM communities, to try to make me updated*." Due to the differences in the culture of the companies, DM practitioners rely on different sources to overcome these issues. In company A, statisticians count on a coordinator who helps translate the problem from the client and understands what data analysis can support. In contrast, the DM practitioner in company B has to figure out the objectives, technology, inputs, output, pretty much everything on his own. All that the manager provides is just the name of the project.

5.3.2.2 Data Understanding and Preparation

DM practitioners typically use data either in a bunch of tables, mainly in raw transaction format, as is the case in company A, or in an enterprise date warehouse, as is the case in company B. Getting data ready is the most-consuming stage of a DM project. To statisticians in company A, it is about 60-70% of the effort; while about 80% in company B. These percentages confirm data preprocessing is the most time-consuming part in DM (Brachman and Anand 1996; Piatetsky-Shapiro, Brachman et al. 1996; Cabena, Hadjinian et al. 1998). DM practitioners definitely struggle a lot with keeping track of many data fields. To overcome this, they put together a data dictionary. With their substantial hands-on experience with DM, statisticians in company A suggest that data preparation is one point where statisticians can make a breakthrough to "*getting data in a format where it can fall in the algorithms*".

Immersed with large data sets, getting a manageable data size is a problem. Often the problem can be solved by employing a stratified sample rather than a simple random sample. A simple random sampling mechanism may easily eliminate the rarities, which are most likely the targets for which the firms are looking. DM practitioners do have to be concerned to some extent with data quality. As one interviewee commented, "*We usually do have a kind of simple data glitch*."

5.3.2.3 Modeling

It is clear from Table 5.2, under different business contexts, DM practitioners have employed different statistical and DM techniques, both modeling and exploratory data analysis. All DM practitioners definitely try a variety of different techniques to fulfill the project goals. A majority of the DM endeavor in company A employs numerous relatively complicated statistical techniques, such as logistic regression, fitting probability distributions, and experimental design; while in company B, in many cases, the DM practitioner explores and summarizes the data using association analysis. Sometimes, he also employs statistical modeling techniques.

When employing these statistical modeling techniques, informants prefer to use them in a simple way, and by the standard of the methodology illustrated in traditional statistical textbooks their practices would be considered as doing wrong things. For example, in the previously mentioned catalog response prediction, statisticians started with a few hundred possible independent variables. They just arbitrarily chose around 10 variables based on their experience. Indeed, as noted by our statisticians, "*Precision is not really important for such big data size*". Especially in the fast changing business world, "*to find the right thing to do is more important than doing things right*". The DM practitioner in company B made the statement "*my theory is to make things simple*" as managers perceive models SIMPLE, and UNDERSTANDABLE are very important. Other comments include: “*Nothing is complicated*”; “*I know it is not easy to communicate with senior management. I never transform variables, because nobody will understand it*”, “*I will sacrifice R-square to fulfill my management team, their expectations*”.

We also observed the DM practitioners in the two companies constructed their initial customer pool in different ways. Statisticians in company A typically start with all customer data, while the DM practitioner in company B usually only works with customers who have spent with the company beyond a certain benchmark. This difference can best be explained by their organizational characteristics. Since company A is targeting common people, it is interested in all customers. As long as the customer spends money, management feels happy. While company B is selling high end merchandises; it eliminates customers who spent less.

The tools DM practitioners used can be categorized into two types: one is traditional statistical software tools; another is specialized DM software tools, such as SAS Enterprise Miner. Statisticians at company A usually stick to comparatively traditional statistical tools, like SAS STAT, and they actually doubt the incremental value of a DM software tool, as stated, “*We would rather have five more statisticians who are good at traditional statistical tools*.” While company B makes an expensive specialized DM software tool available to the DM practitioner, SAS Enterprise Miners, it is interesting to know that, at the point this research was performed, the DM practitioner just used it to get the optimal R-square to check how well the simplified model is doing.

Immersed with large volume of data, informants definitely felt the need to improve their capabilities of actual data analysis. For some projects, their current analytic approaches did not address the issues involved in the projects, thus they have to seek for new ways. In addition to traditional statistical techniques, they also have attempted popular DM tools, such as association rule analysis, neural networks and genetic algorithms.

5.3.2.4 Evaluation and Deployment

DM practitioners check overfitting by using holdout/ test samples and go back to the business problems and check whether their model helps answer the questions. Projects in company A are considered successful if they result in some management actions (e.g., store replenishment, policy change), lead to better decision making (e.g., which type of promotion to use), or inform strengths, weaknesses, opportunities, and threats (SWOT). However, for company B, it is not really easy to evaluate the performance of the DM activity. For example, what part of revenue is coming from a particular effort. Here the problems typically were not as well-defined as those in company A. To make thing even worse, company B has an anti-technology culture. As mentioned by our interviewee, "*My boss never talks about variables. She could not get what an independent variable is, what a dependent variable is. So we are just using inputs or outputs. Whenever she has a problem to understand, I just suppress data into a black-box*", *"The management would think of anything else other than understand* *it"," The overall mining criterion is the model makes sense.*" As a consequence, the DM practitioner in company B had to spend considerable time on organizing the problems and thinking of a better way to present the results, so that the managers could understand. Many times, the DM practitioner has to resort to telling stories, "*It looks like my daily life is making a story. By the story, I can tell and help understanding*", "*To make the story is the most difficult. It is my job to guess business opportunities, and the possible return for the investment*", "*The managers are interested in what the results tell them, not the results themselves.*"

5.4 Discussions and Limitations

Through an exploratory study, this research takes a first step in examining the way DM practitioners have integrated statistics into their DM practice. This study provides new insight into how DM practitioners solve real-world business DM problems, and several issues observed in this study need to be highlighted.

It finds that DM practitioners typically follow a process in performing DM projects. In the process, statistical techniques are being widely used in various business DM contexts, and managers value statisticians. This confirms the important role of statistics in business for leveraging competitive advantages. DM practitioners value statistical techniques/thinking, though it was observed that the company without a long-lasting data analysis culture more felt the need of expensive analytical tools. On the other hand, well-trained statisticians flexibly apply their expertise, and in many cases, they simplify the techniques but still accomplish the project goals.

Beyond this, DM practitioners find business competence and communication skills are among the biggest obstacles that, if not handled well, would be a disaster to their projects. In the statistical community, it is well known that an approximate solution for the right problem is much better than an exact solution for a wrong problem. Examples have shown us that there are many good statisticians and competent analysts whose work is essentially wasted because they are solving problems that do not help the business (Berry and Linoff 2004). As business problem solvers to their organizations, DM practitioners need to form a proper business relationship with their clients (Green 1989; Bassellier and Benbasat 2004). DM practitioners perceive communication with appropriate languages to be important. To achieve this, many times, DM practitioners must downplay their DM skills and use fewer technical terms.

With almost no exception, DM projects done in the two companies are initiated by the management team. However, problem formulation and problem solving are intertwined, and in a real DM project, a large portion of effort goes into how to properly formulate the DM problem (Fayyad, Piatetsky-Shapiro et al. 1999). The inputs of DM practitioners definitely help in both formulating the problem and problem solving, thereby saving important and valuable corporate resources. This happens in company A. In company A when statisticians were "brave" enough to suggest a problem for the manager, the manager agreed and supported the team in solving the problem, leading to a more productive and effective change as mentioned in ?5.3.1.

Though the findings of the case studies allow a better understanding of how DM is influenced by statistics in the retailing industry, the research sites, cannot be used to generalize to the overall population of organizations who employ DM. Large sample data collection efforts will be needed to validate the findings more generally. This is the main motivation for the coming chapter.

CHAPTER 6

A PRELIMINARY SURVEY

In this chapter, I endeavor to determine the influence of both skill sets and stage-efforts profile on the DM implementation success. I first review the IS literature and synthesize the findings to develop a model consisting of skills set as the factors affecting DM project success. This model highlights the importance of team's competency in terms of communication skills, DM skills, and business competence for the successful implementation of DM project. I further identify the role played by data extraction and infrastructure support for the success of the project. I also derive a model employing six stages of the CRISP-DM methodology as a baseline to identify the effect of stage-efforts profile on DM success. Finally, I propose a set of hypotheses and the development of a survey instrument that can be used to empirically validate the research models.

6.1 IS Literature

Studies exploring the importance of skill requirements in the domain of system analysts are intense (Senn 1978; Goldenstein and Rockart 1984; White 1984). Over the years, the studies on system analysts seem to come to a point of agreement that that the analyst's job requires a greater mix of skills (Strout 1971). More recently, the literature concerning skills requirements for systems analysts, while not discounting technical or system skills, appears to increasingly stress the importance of communication skills and business competence (Lee, Trauth et al. 1995; Todd, McKeen et al. 1995). To ensure that IT capabilities are integrated into the business effectively (Rockart, Earl et al. 1996), it is very important for the analysts to form a proper business relationship with business clients (Senn 1978; Green 1989; Bassellier and Benbasat 2004). Conflicts between them may have serious consequences that can be very costly, such as poorly developed systems, failed projects, behavioral dysfunctions (mistrust, avoidance, rejection) and negative user satisfaction (Green 1989).

The literature also suggests that the mix of skills needed by IS personnel varies with position (Zmud 1979). There has been very little published research, however, that looks at data analysts' skills in general, and DM practitioners, in particular. The question of what skills are demanded for DM practitioners to fulfill their job requirements will be addressed in this chapter.

In the IS implementation literature, plenty of studies have been attempted to investigate the factors that influence project implementation success in various contexts, such as decision support systems (Sanders and Courtney 1985), data warehousing (Wixom and Watson 2001), business process reengineering (Grover, Jeong et al. 1995; Teng, Jeong et al. 1998). In their data warehousing success study, Wixom and Watson found that team skills, together with resources and user participation are critical in order to achieve the success at a project level. In a recent case study, Chenoweth, Corral et al. (2006) propose that the interaction of technology and social context is the key to data warehouse success. It further finds that management championing is not that critical in project success. Users can champion the project just as successfully as management. White and Leifer (1986) examined information systems development success from project team participants' perspectives and identified five skills perceived to be important by the team participants: business competence, communication skills, technical expertise, analytic skills, and organizational skills. However, the team participants perceive 1) neither management support nor client involvement important; 2) technical skills are necessary, but not sufficient. While in another study on business process reengineering, Grover, Jeong et al. (1995) found that the social component has the most potential influence, technical component having the least potential influence on success.

6.2 Research Models

In this study, the DM team is considered as a problem solving communication network. Within this network, project team members transform vague ideas, concepts, and requirements given by the business clients, and utilize the necessary tools and techniques to give them the information they want. The team needs to have members with diversified background and knowledge as their functional diversity will increase the amount and variety of information available to solve the problem. This increased information helps project team members to understand the process more quickly and completely from a variety of perspectives, and thus improve their performance. In profiling reengineering projects, Teng, Jeong et al. (1998) found the reengineering project stage-efforts profile influences the reengineering project implementation success. Building upon the literature and observations from case studies, as shown in Figure 6.1, the research framework relates two independent variable sets regarding project profile, skill sets and stage-efforts profile, to DM implementation success.

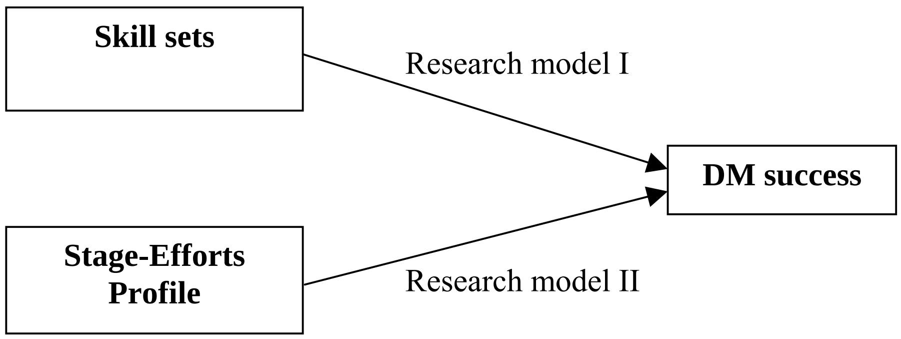


Figure 6.1 Research Framework

Findings from case studies and the literature were synthesized into a model consisting of skills as the factors affecting the success of DM project. As shown in Figure 6.2, research model I relates three independent variables: DM skills, business competence, and communications skills, moderated by data extraction and infrastructure support to DM implementation success. This model highlights the importance of the skill set of team members to the successful outcome of the project. It should be noted that the purpose of this model is not to explain as much as possible of the variations in DM project implementation success. Instead, it attempts to investigate the strength of associations of skill sets with the success.

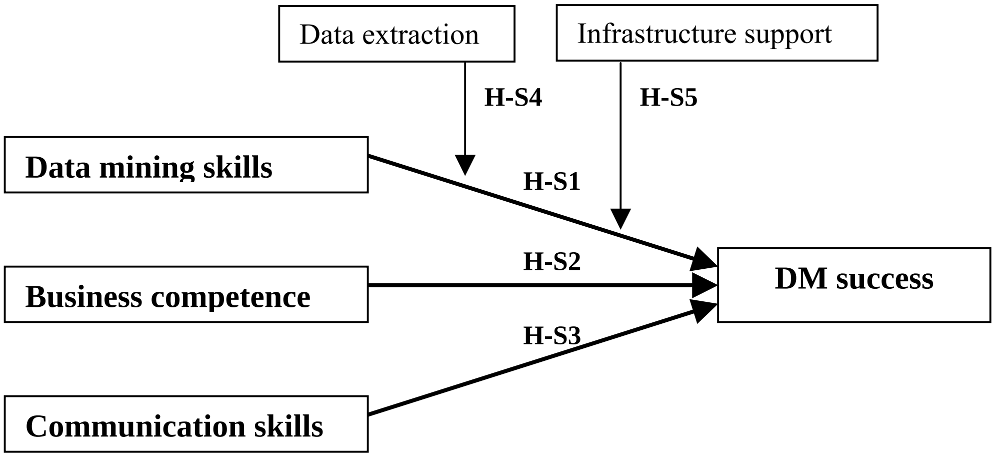


Figure 6.2 Research Model I

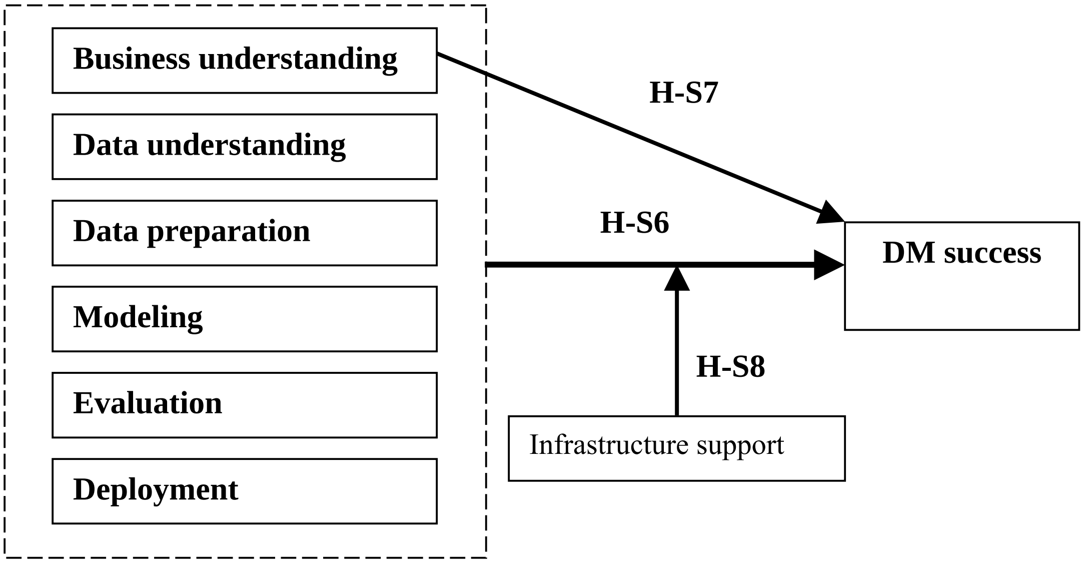


Figure 6.3 Research Model II

Given the wide acceptation of the CRISP-DM model, I derive a model consisting of all six stages of this model. As shown in Figure 6.3, this model highlights the importance of effort at each stage, and its effort on the successful outcome of the project, moderated by infrastructure support.

6.3 Descriptions of Variables and Research Hypotheses

6.3.1 Data Mining Implementation Success

The dependent variable in both models is DM implementation success.

6.3.2 Data Mining Skills

DM skills are defined as how well the DM team knows, chooses and uses the DM techniques/tools to solve the data analysis problems. DM practitioners need technical skills in order to perform their job. It is expected that the project team in a successful deployed DM project will have strong expertise in DM analytical skills. Thus,

H-S1: The DM team's level of DM skills is positively associated with DM implementation success.

6.3.3 Business Competence

Business competence is how well the DM team understands the business domain in which the DM project is deployed and the connections between DM effort and the business. Knowledge in functional areas helps DM practitioners to recognize the context and clearly understand users' needs before starting the implementation of the project, and to transform data into proper knowledge. Thus,

H-S2: The DM team's extent of business competence is positively associated with DM implementation success.

6.3.4 Communication Skills

Communication skills are how well the DM team demonstrates the appropriate communication behavior, speaking the language of business and interacting with business clients. Good communication skills should take into account language variation and appropriateness in contexts of use (Hymes 1971). As mentioned earlier, the IS literature suggests that communication competence is important for implementation success. There is ample evidence that a lack of communication will harm efforts at IS project implementation. It is expected that a similar relationship exists between communication competence and DM implementation success. Thus,

H-S3: The DM team's level of communication skills is positively associated with DM implementation success.

6.3.5 Moderators

There are two variables included in research model I that are hypothesized to moderate the relationship between the independent and dependent variables described above. One is data extraction, another is infrastructure support. Data extraction is defined as how well the DM team is able to use tools to extract data from databases for the DM purpose. Infrastructure support refers to the capability for a Data Warehouse, the availability of software / hardware, etc. that the DM team can access during the implementation of the DM project. To successfully complete DM project, there is no doubt the team members need to be very skillful. However, being skillful alone is not enough; there are other factors that also affect DM project outcomes. Without good data extraction to get the data to begin the project, or without sufficient infrastructure support, it would be difficult for the project team to accomplish the project goals. Thus, the following research hypotheses are proposed:

H-S4: The DM team's level of data extraction skills positively moderates the effect of the DM skills on DM implementation success.

H-S5: The infrastructure support the DM team received positively moderates the effect of the DM skills on DM implementation success.

Given the popularity of the CRISP-DM methodology, it is expected that the more the DM team follows it, the more likely the team can achieve DM implementation success. Thus,

H-S6: The strength of efforts that the DM team devotes to the six stages of the DM process is positively related to DM implementation success.

Two major steps in the business understanding stage are understanding the clients' requirement and converting the business problem into a DM: IS success literature highlights the importance of understanding the clients' requirements. DM literature also emphasizes the value of formulating appropriate DM questions. In a real DM endeavor, people actually spend a large portion of efforts on properly formulating the problem (Fayyad, Piatetsky-Shapiro et al. 1999). Given this, the stage of business understanding in the six stages of DM process seems more critical than other stages in order for the DM project to be successful. Thus,

H-S7: The strength of efforts devoted to the business understanding stage has significantly greater effect on DM implementation success than other stages.

Following a good process model is, no doubt, important. However, if the project team were hampered by the availability and capability of infrastructure support, they could not do a good job. Thus

H-S8: The infrastructure support the DM team received positively moderates the strength of efforts the DM team follows the CRISP-DM methodology on DM implementation success.

6.4 Instrument Development

Instruments were developed by following some generally accepted instrument development guidelines. All are measured based on a five-point scale.

6.4.1 Measures of Data Mining Implementation Success

DeLone and McLean (1992) suggested six major dependent variables to measure information systems success. Their comprehensive taxonomy for six dimensions or categories includes system quality, information quality, use, user satisfaction, individual impact, and organizational impact. However, DM goals should not be stated in broad, general terms, as they are hard to measure. Instead, it would be easier to monitor progress in achieving more specific ones (Berry and Linoff 2004). Based on this, DM implementation success will be measured using five relatively objective items. They are: the outcomes of the project have fulfilled the clients' expectations, the outcomes of project have been implemented and generated measurable benefits, the project has led to improved operational procedures, the project has spawn additional DM projects, and the project has fulfilled the goals of the project.

6.4.2 Measures of Data Mining Skills

In the literature, several studies present measures of technical skills. However, due to the changing nature of technology and the variety of technical positions available, a standardized test to evaluate technical skills is not preferable (Lee, Trauth et al. 1995). This study proposes to use self-reported measures of the DM skills at the project team level. Although objective measures are often preferable, self-reported measures have been successfully used in a number of studies and found to be comparable with objective measures. To capture the DM skills at a DM project team level, five items were developed. They are: knowledge of DM techniques (decision trees, neural networks, etc.), knowledge of statistical techniques (logistic regression etc.), familiarity with DM software, familiarity with statistical software, and comfort with learning new data analysis technology/tools.

6.4.3 Measures of Business Competence

Items for business competence are derived from earlier work. The most comprehensive scales to measure business competence of IT professionals have been developed so far is by Bassellier and Benbasat (2004). In their model, organization-specific knowledge and interpersonal and management knowledge together constitute business competence. To be consistent with others in the literature, two constructs from their organization-specific knowledge perspective are adopted to measure the business competence of the project team.

The first construct is organizational responsibility, and three items will be used. We propose to ask the respondent to measure the extent the team takes actions to stay informed about business developments not directly related to DM, the extent the team participates in business activities that are not directly related to DM, and the extent the team are concerned by the overall performance of the business organization.

The second construct is IT-business integration, the need for IT professionals to act as business problem solvers (Bashein and Markus 1997) and to integrate business development with IT capability. Four items will be used: how experienced is the team at recognizing potential ways to exploit new business opportunities using DM; how experienced is team at analyzing business problems in order to identify DM-based solutions (understand situations, getting the "big picture", identifying underlying root problems, etc.); how experienced is the team at evaluating the organizational impacts of DM solutions; and the team's level of knowledge of the alignment between business goals and DM projects goals in the organization as a whole.

6.4.4 Measures of Communication Skills

Five items of communication skills are derived from literature. Two items are adopted from the Bassellier and Benbasat (2004) study. They are: In general, how well can the team communicate about the DM project in non-technical language and within a business context to non-IT specialists, and how effectively can the team communicate with people at different levels of the organization (e.g., with clients). In the literature, communication decoding and communication encoding have been found to be important in the context of complex systems implementations (Robey, Ross et al. 2002). Thus three items, an overall measure, one item for communication decoding, and one item for communication encoding, in the Ko, Kirsch et al. (2005) study are also adopted. They are: The team can express ideas clearly to clients; when interacting with clients, the team members are good listeners; and the team generally can collaborate well with clients.

6.4.5 Measures of Moderators

Data extraction will be measured by three items: familiarity with data extraction tools such as SQL, knowledgeable about extracting data that the team wants, and the ability to find relevant data that the project needs. Infrastructure support will be measured by the capability of a data warehouse and the availability of software/ hardware.

The above-mentioned constructs and corresponding items used to measure these constructs of the DM project team are summarized in Table 6.1.

Table 6.1 Constructs and Measures of Survey Instrument

Instrument	Items	Source	Item loading
DM Success	The outcomes of the DM project have fulfilled the clients' expectations.	Newly developed	
	The outcomes of the DM project have been implemented and generated measurable benefits.		
	The DM project has led to improved operational procedures.		
	The DM project has spawned additional DM projects.		
	The DM project has fulfilled the goals of the project.		
DM Skills	Our team members are knowledgeable about DM techniques.	Newly developed	
	Our team members are knowledgeable about classical statistical techniques.		
	Our team members are familiar with DM software.		

Table 6.1 – *Continued.*

	Our team members are familiar with statistical software.		
	Our team members are comfortable with learning new data analysis technology/tools		
Business Competence	Our team takes actions to stay informed about business developments not directly related to DM.	Adapted from Bassellier and Benbasat (2004)	0.88
	Our team participates in business activities that are not directly related to DM.		0.75
	Our team is concerned about the overall performance of our business organization.		0.76
	Our team is experienced at recognizing potential ways to exploit new business opportunities using DM.		0.79
	Our team is experienced at analyzing business problems in order to identify DM-based solutions (understanding situations, identifying underlying root problems, etc.).		0.79
	Our team is experienced at evaluating the organizational impacts of DM solutions?		0.84
	Our team is knowledgeable about the alignment between business goals and DM projects goals in the organization as a whole.		0.76
Communication Skills	Our team can express our ideas clearly to our clients.	Adapted from Ko, Kirsch et al. (2005)	0.91
	When interacting with clients, our team members are good listeners.		0.94
	Our team generally can collaborate well with our clients.		0.81
	Our team generally can communicate well about DM project in non-technical language and within a business context to non-IT specialists.	Adapted from Bassellier and Benbasat (2004)	0.78
	Our team generally can communicate effectively with people at different levels of the organization (e.g., with clients).		0.83
Data Extraction	Our team members are familiar with data extraction tools such as SQL.	Newly developed	
	Our team members are knowledgeable about extracting data that we want.		
	Our team members are able to extract relevant data that we need.		

6.4.6 Measures of Stage-Efforts

The CRISP-DM is suggested to be the de facto standard for the DM industry. This study employs CRISP-DM as a baseline model to evaluate the strength of efforts that the project team has attempted to complete the stage based on a five-point scale. The detailed descriptions of each stage are summarized in Table 6.2.

Table 6.2 Stages and Tasks in CRISP-DM Process Model

Stages	Main tasks
Stage 1: Business understanding	• Understand the project objectives and requirements from a business perspective. • Convert this knowledge into a DM problem definition. • Develop a plan to achieve the objectives.
Stage 2: Data Understanding	• Collect and familiarize with the data. • Verify data quality. • Explore data for initial insight and hidden information.
Stage 3: Data Preparation	• Clean, format, integrate, and transform data. • Select tables, records, and attributes. • Construct the final dataset for input to models.
Stage 4: Modeling	• Select modeling technique • Generate test design • Build model • Assess model accuracy and reliability (e.g., with lift chart and confusion matrix)
Stage 5: Evaluation	• Evaluate the extent that the model meets business objectives.
Stage 6: Deployment	• Report final results to customers • Plan deployment (e.g., integrate with legacy systems, and implement a repeatable DM process) • Plan monitoring and maintenance

6.5 Pilot Survey

For the purpose of this research, initial reviews with and a pilot study on DM academicians and practitioners have been conducted to examine the face value of this questionnaire. As a first step toward validating the measurement instrument, the scale items were reviewed by three faculty members and two DM practitioners. They were asked to fill out the survey, making note of any ambiguous or confusing questions or instructions. I spent an average of one hour discussing the instrument with them. Various changes were made to the structure and format of the questions at this stage. Their feedback also led to rewording a number of questions to improve their clarity. The final version of the survey contained 25 scale items plus closed-ended demographic questions (see Appendix D).

The survey is planned to be distributed following guidelines for online survey methodologies (Schaefer and Dillman 1998). It aims to target practitioners who have hands-on experience in the implementation of DM projects. The check of the validity of the proposed research model and propositions will be done at a later stage and remains as future work.

6.6 Limitations

Like most studies of this kind, this study also suffers from limitations. First, the subject of this study is the project team; however an individual will answer the questionnaire. There is a possibility that the person does not know all the things about the company and the project. In this study, some measures will be taken to assure the respondent is really representative of the team and knows things going on in the organization. The second limitation is about measurement. Some well-validated measures for constructs, e.g., communication skills and business competence are adopted from the literature. However, some instruments have been derived, such as DM skills, since in the literature, no such measure exists. Another drawback is that the communication skills were measured between the team and the clients, while ignoring the possibility of poor communication within the team that could confound the perception of the respondent on the communication skills of the team. Multiple measures were used to reduce the measurement bias as much as possible. Future studies should further validate these measures and their effectiveness in a DM context.

CHAPTER 7

CONCLUSIONS

This dissertation empirically examines the critical patterns in both DM research and practice using three research methodologies: citation analysis methodology, case study methodology, and a preliminary survey methodology. Although each methodology has its limitations, the three together can complement each other. The use of citation analysis and case study methodologies allow viewing the issues from two perspectives, namely academia and business practice. The additional use of the survey instrument presents the chance to test and generalize the above findings in other contexts, not limited to the retailing industry. This thus helps to triangulate a rich spectrum of findings and leads to a comprehensive understanding of the critical patterns of DM research and practice.

This study assesses the extent that statisticians and researchers in other fields have influenced and shaped DM research, examines the exchange of ideas between statistics and other fields that have created DM, and investigates how DM practitioners do DM and the way statistics has influenced DM practice. It finds that more eminent DM researchers are currently working in computer science related fields, only a few of them in statistical field. However, a closer look at their background suggests that many of these researchers had former training in statistics. Statistics education is of considerable importance for eminent DM researchers. Pattern recognition and AI/ML fields heavily borrow and learn from statistics; while this is not the case for database systems.

Through two in-depth field interviews, this research empirically examines the way a DM project is undertaken from the DM practitioners' perspective. The research reveals that statistics has been applied in a wide range of DM activities and has substantial impact on DM in the marketing and the retailing industry. In fact, the companies value statisticians as vital human resources in their DM endeavor. It also suggests that no matter how complex the analytical software, DM practitioners including statisticians prefer to employ methods in a simple way. We further finds that without devaluating their analytical expertise, DM practitioners are pressing for knowledge in their functional areas and skills to properly communicate with their managers. These finding provide important insights into the antecedents of DM implementation success. These findings, plus theory and literature, finally lead to a research framework relating both skill sets and the stage-efforts profile to DM implementation success.

7.1 Contributions of this Study

The thesis contributes to the DM literature in several ways. First, it goes beyond conceptual relationships among DM reference disciplines to examine empirically the dynamic interchange of ideas among them. As mentioned in the literature review section, past knowledge of the intellectual structure of DM and the interchange of ideas among DM reference disciplines was mainly from anecdotal accounts. This study provides empirical support for many of these accounts.

Second, this research examines the state of art of DM practice in the retailing industry and provides new insight into how DM practitioners solve real-world business DM issues. It also takes a first step in examining the way DM practitioners have integrated statistics into their DM endeavors. Given the fact that the DM field is still short of theories, a study like this certainly expands our understanding of this relatively new phenomenon in business and would be useful in generating theory for the field.

Third, it suggests to DM practitioners the skill requirements in successfully deployed DM projects, and provides further guidelines for organizations that attempt DM projects, particularly in the integration of statistics to achieve DM success. Extant studies on skill sets in project implementation span across different contexts. Though none of the skills items is really new, this is a study that makes the first attempt to put them together under the domain of DM implementation. DM implementation literature touches skills but at an organizational level. However, this study investigates all these issues from the perspective of DM practitioners with a focus on a DM project team level.

Fourth, the models proposed in this thesis may be employed as a baseline for future DM implementation research. Research model I provides a first step toward prescribing the effect of skill sets on DM implementation success. It can help in gaining a better understanding of this relatively new phenomenon in practice, and lays the foundation for future research on implementation issues. Research model II provides a first step toward prescribing the stage-efforts profile for DM implementation success. It suggests to the DM team how they can optimize their effort in each stage within the limits of time and budget and still accomplish the project goal.

7.2 Future Research Agenda

As a relatively new phenomenon in both research and practice, DM has brought forth numerous opportunities as well as challenges. The business world is also changing rapidly. This research explores opportunities for further research to investigate DM implementation. Future studies on DM implementation might follow the following research agendas:

First, continue research on important factors that affect the successful DM project implementations in a business context. DM is gaining increasing importance in business. Organizations need guidance on how to improve DM for effective decision-making and efficient operations in light of issues at both the project team and organization level.

Second, intensify DM research into various functional areas of business and assess its performance. Performance analysis is fundamental in each sector of each organization. The proper integration of DM into other business processes will certainly add more value to organization.

Third, there is a wide range of applications of DM in the security context. There are also many privacy and ethical concerns associated with DM, particularly, related to data collection. The impacts of these issues on the adoption of DM techniques and the actual effectiveness of DM techniques are not clear. DM practitioners need to better address these concerns and understand the contexts in order to successfully apply DM.

Fourth, given the multidisciplinary nature of DM, how to provide the best education to our next generation DM researchers and practitioners certainly remains an issue.

APPENDIX A

RAW DATA OF AUTHOR CO-CITATION ANALYSIS

Appendix A Raw Data of Author Co-Citation Analysis

id	1	2	3	4	5	6	7	8	9	10	11	12	13
1:AGRAWAL	422												
2:BAYARDO	118	118											
3:BRACHMAN	33	8	78.5										
4:BREIMAN	81	10	17	716									
5:BRIN	189	58	6	13	182								
6:BUNTINE	6	1	4	103	1	167							
7:CHEUNG	117	14	7	9	42	0	111						
8:DERAEDT	36	5	4	27	2	16	0	159					
9:DIETTERICH	20	4	7	279	2	32	1	25	369				
10:DOMINGOS	22	5	2	75	6	17	2	7	65	158			
11:DUDA	44	4	13	222	14	33	4	1	76	73	459		
12:ESTER	81	0	5	10	12	1	15	1	1	2	17	110	
13:FAYYAD	257	29	78	216	42	40	36	29	62	55	102	44	453
14:FELDMAN	56	8	6	5	15	2	21	2	1	2	1	7	23
15:FISHER	22	2	6	40	6	11	5	4	14	9	29	12	45
16:FRAWLEY	48	2	14	9	8	2	6	2	3	2	2	19	56
17:FRIEDMAN	3	0	0	102	0	4	9	1	49	11	13	1	6
18:GOLDBERG	20	4	0	15	11	2	2	2	9	5	26	1	14
19:GRZYMAL-ABUSSE	12	0	2	13	0	0	0	0	4	6	9	0	30
20:HAN	193	17	15	41	31	3	37	7	7	7	20	42	108
21:HAND	29	3	11	139	4	19	6	5	18	23	62	1	64
22:HECKERMAN	35	5	6	40	11	94	1	0	9	14	60	4	46
23:HOLTE	13	5	2	92	6	26	2	9	46	50	43	0	59
24:JAIN	8	0	0	12	1	6	4	0	12	5	83	5	7
25:KOHAVI	48	13	8	240	11	45	7	10	166	91	107	7	140

id	27	28	16	17	18	19	20	21	22	23	24	25	26
1:AGRAWAL	56	22	48	3	20	12	193	29	35	13	8	48	33
2:BAYARDO	8	2	2	0	4	0	17	3	5	5	0	13	4
3:BRACHMAN	6	6	14	0	0	2	15	11	6	2	0	8	14
4:BREIMAN	5	40	9	102	15	13	41	139	40	92	12	240	137
5:BRIN	15	6	8	0	11	0	31	4	11	6	1	11	14
6:BUNTINE	2	11	2	4	2	0	3	19	94	26	6	45	11
7:CHEUNG	21	5	6	9	2	0	37	6	1	2	4	7	4
8:DERAEDT	2	4	2	1	2	0	7	5	0	9	0	10	1
9:DIETTERICH	1	14	3	49	9	4	7	18	9	46	12	166	30
10:DOMINGOS	2	9	2	11	5	6	7	23	14	50	5	91	7
11:DUDA	1	29	2	13	26	9	20	62	60	43	83	107	328
12:ESTER	7	12	19	1	1	0	42	1	4	0	5	7	14
13:FAYYAD	23	45	56	6	14	30	108	64	46	59	7	140	58
14:FELDMAN	53.5												
15:FISHER	5	124											
16:FRAWLEY	10	8	76.5										
17:FRIEDMAN	1	9	1	121									
18:GOLDBERG	11	11	0	9	52								
19:GRZYMAL-ABUSSE	0	1	2	5	1	187							
20:HAN	17	18	30	13	8	9	196						
21:HAND	2	7	0	9	24	3	22	144					
22:HECKERMAN	3	11	3	2	8	2	9	16	279				
23:HOLTE	1	8	6	3	6	6	6	18	11	161			
24:JAIN	9	12	0	75	25	0	16	8	3	0	92.5		
25:KOHAVI	56	22	48	3	20	12	193	29	35	13	8	48	

id	27	28	29	30	31	32	33	34	35	36	37	38	39
1:AGRAWAL	35	26	5	305	67	30	36	158	57	35	123	33	196
2:BAYARDO	13	1	0	60	10	4	4	35	1	13	13	4	23
3:BRACHMAN	14	5	1	18	28	5	15	9	4	17	25	10	46
4:BREIMAN	94	44	29	28	130	59	80	12	34	68	37	69	884
5:BRIN	9	3	0	75	5	14	3	37	2	11	32	10	17
6:BUNTINE	33	14	2	7	39	12	37	1	2	108	8	12	122
7:CHEUNG	2	0	1	38	6	1	2	22	0	4	14	10	16
8:DERAEDT	14	71	0	38	45	14	151	3	3	7	8	8	96
9:DIETTERICH	68	23	13	19	81	31	64	2	10	26	8	38	293
10:DOMINGOS	57	12	1	5	36	29	12	2	11	13	10	36	149
11:DUDA	90	11	22	11	56	52	19	22	17	113	13	21	241
12:ESTER	1	0	3	11	9	7	0	95	0	2	9	6	14
13:FAYYAD	102	38	20	97	112	34	69	55	81	41	86	60	432
14:FELDMAN	4	2	0	24	5	10	3	9	3	8	12	9	13
15:FISHER	82	8	2	6	81	20	19	14	7	12	9	8	84
16:FRAWLEY	12	6	1	11	24	1	12	14	11	5	33	4	49
17:FRIEDMAN	7	0	7	3	5	21	0	6	8	7	0	2	50
18:GOLDBERG	12	1	0	1	16	14	6	0	1	7	3	4	39
19:GRZYMAL-ABUSSE	10	3	1	4	36	8	1	0	189	3	7	3	51
20:HAN	15	6	3	57	32	18	5	53	18	14	41	9	87
21:HAND	25	10	9	20	10	21	8	2	6	19	9	28	84
22:HECKERMAN	34	2	1	11	11	11	13	6	5	393	13	9	70
23:HOLTE	46	15	5	7	49	22	22	0	9	21	12	40	160
24:JAIN	17	1	1	0	6	12	2	12	0	8	0	2	21
25:KOHAVI	133	21	16	22	55	46	29	9	32	37	14	43	352

id	40	41	42	43	44	45	46	47	48	49	50	51
1:AGRAWAL	13	49	146	3	30	281	196	11	29	60	183	35
2:BAYARDO	3	10	28	1	1	52	56	1	4	4	51	2
3:BRACHMAN	18	21	7	1	2	12	7	1	3	10	5	9
4:BREIMAN	190	19	10	51	13	22	21	268	87	65	8	16
5:BRIN	4	72	74	1	3	93	81	5	22	11	60	3
6:BUNTINE	50	0	1	4	0	3	4	21	10	9	1	3
7:CHEUNG	1	5	40	0	0	57	31	0	2	1	48	3
8:DERAEDT	9	0	7	2	0	7	28	8	8	2	3	0
9:DIETTERICH	52	11	2	40	6	5	9	116	48	18	6	7
10:DOMINGOS	14	13	2	9	5	5	4	36	29	7	4	4
11:DUDA	277	68	3	99	8	7	6	312	46	162	4	5
12:ESTER	2	11	8	0	0	9	9	2	2	2	4	3
13:FAYYAD	47	26	39	12	35	72	55	34	51	59	36	42
14:FELDMAN	5	24	7	0	0	27	14	2	3	3	11	1
15:FISHER	29	16	2	3	1	5	4	3	6	16	2	6
16:FRAWLEY	11	9	4	0	1	17	7	1	5	14	7	12
17:FRIEDMAN	14	0	0	17	2	0	1	65	3	5	4	0
18:GOLDBERG	39	13	0	4	0	2	1	16	4	20	4	1
19:GRZYMAL-ABUSSE	3	1	0	0	72	1	1	6	1	24	1	112
20:HAN	9	17	39	4	3	90	43	8	18	15	31	12
21:HAND	32	4	1	5	1	5	10	37	15	18	4	2
22:HECKERMAN	16	13	2	4	0	7	5	18	2	33	0	2
23:HOLTE	24	9	1	3	3	4	4	14	26	8	5	6
24:JAIN	12	18	0	6	0	1	1	24	5	11	3	0
25:KOHAVI	66	10	6	18	11	15	13	83	54	14	14	11

id	1	2	3	4	5	6	7	8	9	10	11	12	13
26:KOHONEN	33	4	14	137	14	11	4	1	30	7	328	14	58
27:LANGLEY	35	13	14	94	9	33	2	14	68	57	90	1	102
28:LAVRAC	26	1	5	44	3	14	0	71	23	12	11	0	38
29:MANGASARIAN	10	0	1	64	0	7	1	1	19	1	22	3	30
30:MANNILA	305	60	18	28	75	7	38	38	19	5	11	11	97
31:MICHALSKI	67	10	28	130	5	39	6	45	81	36	56	9	112
32:MITCHELL	30	4	5	59	14	12	1	14	31	29	52	7	34
33:MUGGLETON	36	4	15	80	3	37	2	151	64	12	19	0	69
34:NG	158	35	9	12	37	1	22	3	2	2	22	95	55
35:PAWLAK	57	1	4	34	2	2	0	3	10	11	17	0	81
36:PEARL	35	13	17	68	11	108	4	7	26	13	113	2	41
37:PIATETSKY-SHAPIRO	123	13	25	37	32	8	14	8	8	10	13	9	86
38:PROVOST	33	4	10	69	10	12	10	8	38	36	21	6	60
39:QUINLAN	196	23	46	884	17	122	16	96	293	149	241	14	432
40:RUMELHART	13	3	18	190	4	50	1	9	52	14	277	2	47
41:SALTON	49	10	21	19	72	0	5	0	11	13	68	11	26
42:SAVASERE	146	28	7	10	74	1	40	7	2	2	3	8	39
43:SCHOLKOPF	3	1	1	51	1	4	0	2	40	9	99	0	12
44:SKOWRON	30	1	2	13	3	0	0	0	6	5	8	0	35
45:SRIKANT	281	52	12	22	93	3	57	7	5	5	7	9	72
46:TOIVONEN	196	56	7	21	81	4	31	28	9	4	6	9	55
47:VAPNIK	11	1	1	268	5	21	0	8	116	36	312	2	34
48:WITTEN	29	4	3	87	22	10	2	8	48	29	46	2	51
49:ZADEH	60	4	10	65	11	9	1	2	18	7	162	2	59
50:ZAKI	183	51	5	8	60	1	48	3	6	4	4	4	36
51:ZIARKO	35	2	9	16	3	3	3	0	7	4	5	3	42

id	14	15	16	17	18	19	20	21	22	23	24	25	26
26:KOHONEN	6	17	7	3	18	3	27	41	10	16	27	32	605
27:LANGLEY	4	82	12	7	12	10	15	25	34	46	17	133	23
28:LAVRAC	2	8	6	0	1	3	6	10	2	15	1	21	2
29:MANGASARIAN	0	10	1	8	2	1	7	22	2	5	3	20	21
30:MANNILA	24	6	11	3	1	4	57	20	11	7	0	22	8
31:MICHALSKI	5	81	24	5	16	36	32	10	11	49	6	55	37
32:MITCHELL	10	20	1	21	14	8	18	21	11	22	12	46	18
33:MUGGLETON	3	19	12	0	6	1	5	8	13	22	2	29	9
34:NG	9	14	14	6	0	0	53	2	6	0	12	9	10
35:PAWLAK	3	7	11	8	1	189	18	6	5	9	0	32	13
36:PEARL	8	12	5	7	7	3	14	19	393	21	8	37	22
37:PIATETSKY-SHAPIRO	12	9	33	0	3	7	41	9	13	12	0	14	7
38:PROVOST	9	8	4	2	4	3	9	28	9	40	2	43	6
39:QUINLAN	13	84	49	50	39	51	87	84	70	160	21	352	132
40:RUMELHART	5	29	11	14	39	3	9	32	16	24	12	66	732
41:SALTON	24	16	9	0	13	1	17	4	13	9	18	10	120
42:SAVASERE	7	2	4	0	0	0	39	1	2	1	0	6	2
43:SCHOLKOPF	0	3	0	17	4	0	4	5	4	3	6	18	55
44:SKOWRON	0	1	1	2	0	72	3	1	0	3	0	11	5
45:SRIKANT	27	5	17	0	2	1	90	5	7	4	1	15	4
46:TOIVONEN	14	4	7	1	1	1	43	10	5	4	1	13	7
47:VAPNIK	2	3	1	65	16	6	8	37	18	14	24	83	130
48:WITTEN	3	6	5	3	4	1	18	15	2	26	5	54	20
49:ZADEH	3	16	14	5	20	24	15	18	33	8	11	14	150
50:ZAKI	11	2	7	4	4	1	31	4	0	5	3	14	4
51:ZIARKO	1	6	12	0	1	112	12	2	2	6	0	11	2

id	27	28	29	30	31	32	33	34	35	36	37	38	39
26:KOHONEN													
27:LANGLEY	240												
28:LAVRAC	22	168											
29:MANGASARIAN	10	1	124										
30:MANNILA	16	19	2	311									
31:MICHALSKI	109	66	2	21	334								
32:MITCHELL	42	19	4	14	45	142							
33:MUGGLETON	58	137	4	29	112	32	277						
34:NG	3	0	2	60	16	11	6	157					
35:PAWLAK	26	10	3	12	77	21	13	4	369				
36:PEARL	55	5	2	22	33	19	32	26	24	356			
37:PIATETSKY-SHAPIRO	22	11	2	49	42	9	18	24	14	15	147		
38:PROVOST	23	19	1	12	27	15	24	7	8	11	24	115	
39:QUINLAN	238	127	24	61	425	134	265	24	145	127	84	100	871
40:RUMELHART	52	20	15	5	77	31	39	3	18	57	8	7	243
41:SALTON	17	0	0	9	15	51	7	17	12	33	8	1	65
42:SAVASERE	3	1	0	66	7	2	3	35	0	3	22	5	14
43:SCHOLKOPF	12	0	46	3	2	20	3	4	0	5	4	4	30
44:SKOWRON	5	2	0	6	29	9	3	0	285	4	4	3	42
45:SRIKANT	9	4	1	129	18	6	8	59	7	7	50	13	44
46:TOIVONEN	9	10	1	188	14	7	22	38	5	4	28	11	30
47:VAPNIK	32	6	92	7	13	57	14	1	15	35	5	18	159
48:WITTEN	25	15	3	9	25	24	21	3	5	13	6	15	146
49:ZADEH	20	7	8	13	68	17	15	16	205	192	14	4	136
50:ZAKI	5	3	0	67	8	4	5	31	4	1	10	8	21
51:ZIARKO	11	3	0	9	42	8	6	5	247	1	19	7	65

id	40	41	42	43	44	45	46	47	48	49	50	51
26:KOHONEN	732	120	2	55	5	4	7	130	20	150	4	2
27:LANGLEY	52	17	3	12	5	9	9	32	25	20	5	11
28:LAVRAC	20	0	1	0	2	4	10	6	15	7	3	3
29:MANGASARIAN	15	0	0	46	0	1	2	120	3	12	0	2
30:MANNILA	5	9	66	3	6	129	188	7	9	13	67	9
31:MICHALSKI	77	15	7	2	29	18	14	13	25	68	8	42
32:MITCHELL	31	51	2	20	9	6	7	57	24	17	4	8
33:MUGGLETON	39	7	3	3	3	8	22	14	21	15	5	6
34:NG	3	17	35	4	0	59	38	1	3	16	31	5
35:PAWLAK	18	12	0	0	285	7	5	15	5	205	4	247
36:PEARL	57	33	3	5	4	7	4	35	13	192	1	1
37:PIATETSKY-SHAPIRO	8	8	22	4	4	50	28	5	6	14	10	19
38:PROVOST	7	1	5	4	3	13	11	18	15	4	8	7
39:QUINLAN	243	65	14	30	42	44	30	159	146	136	21	65
40:RUMELHART	626											
41:SALTON	24	131										
42:SAVASERE	2	4	154									
43:SCHOLKOPF	27	13	0	353								
44:SKOWRON	3	2	0	0	229							
45:SRIKANT	6	7	86	1	4	254						
46:TOIVONEN	3	2	75	3	3	98	241					
47:VAPNIK	165	50	0	552	7	2	5	566				
48:WITTEN	13	69	1	12	2	3	5	31	151			
49:ZADEH	230	42	3	3	62	18	9	25	8	314		
50:ZAKI	4	3	51	1	3	74	64	0	2	2	162	
51:ZIARKO	3	12	4	0	100	7	5	3	3	42	1	230

APPENDIX B

RAW DATA OF JOURNAL CITATION ANALYSIS

Appendix B Raw Data of Journal Citation Analysis

1997-2000

Journals	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1:TODS	101	0	8	52	1	10	1	35	1	1	0	0	0	0	0	1	17
2:AOS	0	1659	3	0	0	2	83	0	22	1	1	584	302	23	8	6	0
3:AI	15	9	1448	102	4	23	16	13	43	56	66	9	17	138	18	5	2
4:CACM	2	0	10	244	0	3	0	11	4	1	0	0	0	4	1	1	3
5:DMKD	3	5	16	18	9	3	2	23	10	2	1	16	3	49	5	17	6
6:TCOMP	18	1	6	83	0	1192	110	11	8	2	0	1	0	0	2	2	2
7:TIT	1	166	5	22	0	54	3920	0	46	6	1	50	61	21	21	9	1
8:TKDE	216	2	107	170	2	49	12	250	38	51	1	3	7	58	16	23	37
9:TPAMI	1	28	112	40	1	83	102	8	1824	145	2	70	73	36	75	483	0
10:TSMC	18	14	226	111	0	88	60	49	254	1121	7	21	23	108	81	208	1
11:JAIR	5	0	274	22	0	6	5	5	17	21	30	3	7	89	18	7	0
12:JASA	7	981	0	4	0	1	20	0	43	4	1	1894	627	5	3	8	0
13:JRSS	0	284	0	0	0	0	8	0	10	0	0	485	431	0	0	1	0
14:ML	0	38	130	46	10	4	74	8	35	39	35	22	31	522	77	19	0
15:NC	0	28	24	6	1	16	53	1	78	44	13	35	59	74	872	41	0
16:PR	0	16	78	53	0	133	108	11	1501	306	0	24	42	20	71	1363	0
17:VLDB	70	0	2	37	3	11	2	59	5	1	0	2	0	0	1	3	29

2001-2004

Journals	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1:TODS	82	0	10	23	9	8	1	51	6	3	1	0	0	4	0	1	16
2:AOS	0	1336	0	4	1	0	55	0	19	1	0	498	278	31	6	3	0
3:AI	5	4	1024	68	3	15	10	26	52	60	86	11	19	103	20	18	0
4:CACM	5	0	12	382	5	1	4	5	7	7	1	3	0	1	0	1	0
5:DMKD	3	11	8	18	29	2	12	22	7	5	6	11	12	47	7	6	2
6:TCOMP	26	0	2	84	0	903	43	33	3	3	0	1	1	0	3	2	3
7:TIT	1	177	5	17	1	31	6552	1	50	6	2	46	57	28	24	16	0
8:TKDE	227	6	139	159	22	43	23	348	76	55	15	15	17	91	25	64	55
9:TPAMI	3	40	118	71	14	61	103	7	2155	141	5	64	125	75	146	516	0
10:TSMC	13	19	235	132	14	110	61	79	579	1324	22	21	32	134	108	324	4
11:JAIR	3	9	317	21	4	10	24	8	26	26	99	6	12	107	9	6	1
12:JASA	0	710	1	2	3	3	24	0	28	2	2	1332	518	18	22	6	0
13:JRSS	0	331	1	0	1	2	9	0	11	0	0	475	445	1	9	9	0
14:ML	0	57	81	20	37	9	59	9	44	8	26	41	44	405	94	44	4
15:NC	0	72	38	6	4	7	53	4	92	39	5	47	61	89	1190	32	0
16:PR	1	35	74	52	23	132	113	12	2033	360	4	56	75	84	105	1649	3
17:VLDB	57	0	4	36	5	2	1	56	8	3	3	6	1	3	0	5	30

Note: 1:TODS represents ACM Transactions on Database Systems; 2:AOS represents Annals of Statistic; 3:AI represents Artificial Intelligence; 4:CACM represents Communications of the ACM; 5:DMKD represents Data Mining and Knowledge Discovery; 6:TCOMP represents IEEE Transactions on Computers; 7:TIT represents IEEE Transactions on Information Theory; 8:TKDE represents IEEE Transactions on Knowledge and Data Engineering; 9:TPAMI represents IEEE Transactions on Pattern Analysis and Machine Intelligence; 10:TSMC represents IEEE Transactions on Systems, Man, and Cybernetics; 11:JAIR represents Journal of Artificial Intelligence Research; 12:JASA represents Journal of the American Statistical Association; 13:JRSS-B represents Journal of the Royal Statistical Society: Series B; 14:ML represents Machine Learning; 15:NC represents Neural Computation; 16:PR represents Pattern Recognition; 17:VLDB represents VLDB Journal

APPENDIX C

INTERVIEW PROTOCOL

Appendix C Interview Protocol

General company information

1. Describe your organizational structure / background
2. Describe your customer size
3. What motivates your organization to do DM?

Questions about how statisticians are doing DM?

1. Do you consider what you do as data mining?
2. What do you think about the relationship between statistics and DM?
3. How you collect the data for the DM projects?
4. Is the data available in the database or did you have to create the variables?
5. Do you sample? How?
6. In your applications, is the extraction of information from the database dynamical or static?
7. How do you deal with data preprocessing problems when faced with such large data sets?
8. Is model building done typically more than pattern finding?
9. Who initially proposes the problems to work on?
10. Do you check data quality?
11. Are the people in your group with primarily statistical backgrounds?
12. Please describe people involved in your project team.
13. How do you avoid overfitting your data?
14. Do you use special DM procedures or DM tools?
15. What DM/statistical techniques have been used in data mining? How?

16. What DM projects you have been personally involved as a participant? Please describe.

17. How do you evaluate the DM results?

18. How do you measure the success of the project?

19. Do you do strengths, weaknesses, opportunities, and threats (SWOT) analysis?

20. Do you do causal inference? What is your opinion about causal inference from observational study?

Questions related to others issues:

1. What factors you believe are important for DM implementation success?

2. What kind of reading do you do?

3. What can statistical techniques or statisticians further contribute to data mining?

4. Did you fail in any DM project?

5. Do you have any unsolved, methodological problems?

APPENDIX D

SURVEY QUESTIONNAIRE

Appendix D Survey Questionnaire

Dear pilot survey respondent:

This pilot survey aims to examine the appropriateness of the survey questionnaire and to initially verify the relationships between skill sets and Data Mining implementation success. It will take you about 15 minutes to complete the questionnaire. The questionnaire is anonymous, and your participation in this study is voluntary. The information you provide will not be linked to your personal identity. We will devote maximum effort to protecting the confidentiality of the information you provide. You must be 18 or older in order to participate.

If you have any question regarding this survey, please contact the researcher (awang@uta.edu) or the faculty advisors (jtteng@uta.edu and whiteside@uta.edu). You may also contact the Office of Research Compliance 817-272-3723 at The University of Texas at Arlington regarding questions related to your participation in this research project.

In this study, we characterize Data Mining (DM) as follows:

- It is often about secondary data analysis, in which the data have been collected for other purposes.
- The data may be typically stored in databases, data marts or in a data warehouse.
- The data size is relatively large.
- The purpose can be discovering hidden patterns in the data.
- It is generally conducted to support effective decision-making, to improve operational efficiency, or to create knowledge.

Part I. DM in Organization

1. Has your organization attempted DM? [ ] yes [ ] no

If NO, please complete the following questions.

a. Does your organization plan to do DM? [ ] yes [ ] no

If yes, when does your organization plan to do DM?

[ ] in 3 months [ ] in 6 months [ ] in 1 year [ ] other, please specify:__________

b. Please check any reason(s), why your organization has not attempted DM?

[ ] data unavailability [ ] lack of expertise [ ] no recognized need [ ] other, please specify: ____

Please go to part IV to complete the survey.

If YES, and you have been a project team member, please continue.
If you have never been a team member, please forward this questionnaire to someone who has been a team member.

a. How many DM projects have been conducted in your organization so far?_____ What percentage were considered successful?_____

b. How long has your organization been doing DM (in years)? ______

c. Functional areas where your organization currently applies DM (in order of frequency):

(1)____________________(2) ____________________(3)_____________________

d. For the following DM tasks, please rank the frequency of each that your organization has done, if applicable, in the past two years. (1= most frequent, 2= second most frequent, 3= third most frequent, 4= fourth most frequent, 5= fifth most frequent, 6= least frequent)

Classification	[ ]	Clustering	[ ]
Prediction	[ ]	Association rule analysis (market basket analysis etc.)	[ ]
Hypothesis testing	[ ]	Visualization	[ ]
		Other, please specify:__________________	[ ]

e. The following table lists major DM techniques. Please number the three techniques that your organization has used most often.

1= most frequent, 2= second most frequent, 3= third most frequent

Decision trees	[ ]	Association rules	[ ]
Neural networks	[ ]	Classical regression (linear, logistic etc.)	[ ]
Sequence / time series analysis	[ ]	Support vector machine	[ ]
Clustering techniques	[ ]	Bayesian networks	[ ]
Visualization	[ ]	Other, please specify:_________________	[ ]

f. Which of the following analytic tools are used in your organization? Please check all that apply.

CART/MARS/TreeNet	[ ]	SAS	[ ]	Excel	[ ]
S-Plus	[ ]	SAS Enterprise Miner	[ ]	XL Miner	[ ]
Statsoft Statistica	[ ]	SPSS Clementine	[ ]	Oracle Data Mining	[ ]
Insightful Miner	[ ]	SPSS	[ ]	IBM Intelligent Miner	[ ]
Wrote own code	[ ]	other, please specify:________________________			[ ]

g. Please indicate the extent that each of the following reasons motivated your organization to attempt DM (1= to a little extent, 3= to a moderate extent, 5= to a great extent).

Reducing cost	1	2	3	4	5
Increasing revenue	1	2	3	4	5
Creating more business opportunities	1	2	3	4	5
Following competitors	1	2	3	4	5
More targeted customer service	1	2	3	4	5
Gaining market share	1	2	3	4	5

Part II DM Team Skills

1. Please assess your team’s level of skill in the following areas.

   1= strongly disagree, 3= moderately agree, 5= strongly agree

Our team members are knowledgeable about DM techniques (decision trees, neural networks, etc.).	1	2	3	4	5
Our team members are knowledgeable about classical statistical techniques (logistic regression etc.).	1	2	3	4	5
Our team members are familiar with DM software.	1	2	3	4	5
Our team members are familiar with statistical software.	1	2	3	4	5
Our team members are comfortable with learning new data analysis technology/tools.	1	2	3	4	5
Our team can express our ideas clearly to our clients.	1	2	3	4	5
When interacting with clients, our team members are good listeners.	1	2	3	4	5
Our team generally can collaborate well with our clients.	1	2	3	4	5
Our team generally can communicate well about the DM project in non-technical language and within a business context to non-IT specialists.	1	2	3	4	5
Our team generally can communicate effectively with people at different levels of the organization (e.g., with clients).	1	2	3	4	5
Our team members are familiar with data extraction tools such as SQL.	1	2	3	4	5
Our team members are knowledgeable about extracting data that we want.	1	2	3	4	5
Our team members are able to extract relevant data that we need.	1	2	3	4	5

2. Please assess the following conditions regarding your team's business competence (1= to a little or no extent, 3= to a moderate extent, 5= to a great extent).

Our team takes actions to stay informed about business developments not directly related to DM.	1	2	3	4	5
Our team participates in business activities that are not directly related to DM.	1	2	3	4	5
Our team is concerned about the overall performance of our business organization.	1	2	3	4	5
Our team is experienced at recognizing potential ways to exploit new business opportunities using DM.	1	2	3	4	5
Our team is experienced at analyzing business problems in order to identify DM-based solutions (understanding situations, identifying underlying root problems, etc.).	1	2	3	4	5
Our team is experienced at evaluating the organizational impacts of DM solutions?	1	2	3	4	5
Our team is knowledgeable about the alignment between business goals and DM projects goals in the organization as a whole.	1	2	3	4	5

Part III Project Characteristics

Please answer all remaining questions in reference to a recently COMPLETED DM project (referred to as **Project P**).

1. What year was Project P attempted?_____ The project took ____ months to complete.
2. Which functional area did Project P involve?

Finance	[ ]	Marketing	[ ]	Quality	[ ]
Manufacturing	[ ]	Operations	[ ]	Other, please specify:________________	[ ]

3. How many members did Project P team have?______Of these, how many have formal training in

Finance:__________	Marketing:_______	Computer Science: _________
Management:_________	Statistics:__________	Other, please specify:__________

4. Size of the data: [ ] over 1,000,000 records [ ] 100,000-999,999 records [ ] 10,000-99,999 records [ ] below 10,000 records

5. The data were stored in [ ] data warehouse [ ] multiple database tables [ ] spreadsheet [ ] text file [ ] other, please specify: _________

6. Who initially raised the question that led to Project P? [ ] Business analyst [ ] Data miner

   [ ] DBA [ ] Manager [ ] other, please specify: ______

7. What was your main role in Project P? [ ] Business analyst [ ] Data miner [ ] DBA [ ] Manager [ ] other, please specify: ______.
8. Please briefly describe Project P.

- Project title: ____
- Question(s) addressed:____

  ____
- Software and techniques used: ____

  ____

9. Please indicate to what extent Project P has been successfully accomplished (1= to a little extent, 3= to a moderate extent, 5= to a great extent).

The outcomes of Project P have fulfilled the clients' expectations.	1	2	3	4	5
The outcomes of Project P have been implemented and generated measurable benefits.	1	2	3	4	5
Project P has led to improved operational procedures.	1	2	3	4	5
Project P has spawned additional DM projects.	1	2	3	4	5
Project P has fulfilled the goals of the project.	1	2	3	4	5

10. Did your team use a formal DM process in Project P? [ ] yes [ ] no

If yes, please indicate the name of the process.

[ ] CRISP-DM [ ] SAS SEMMA [ ] My organization's own (This process was [ ] developed in-house [ ] provided by a consulting group [ ] a consultant's process modified for our company)

[ ] other, please specify: __________

11. A DM project typically has to go through the following 6 stages. For each stage of Project P, consider the strength of effort (time and resources) spent on it. Please assess this strength of effort for each stage (1= very weak, 2= weak, 3= median, 4= strong, 5= very strong). If the stage was not attempted, please leave the box blank.

[ ] Stage 1: Business Understanding • Understand the project objectives and requirements from a business perspective. • Convert this knowledge into a DM problem definition. • Develop a plan to achieve the objectives.	**[ ] Stage 4: Modeling** • Select modeling technique • Generate test design • Build model • Assess model accuracy and reliability (e.g., with lift chart and confusion matrix)
[ ] Stage 2: Data Understanding • Collect and become familiar with the data. • Verify data quality. • Explore data for initial insight and hidden information.	**[ ] Stage 5: Evaluation** • Evaluate the extent that the model meets business objectives.
[ ] Stage 3: Data Preparation • Clean, format, integrate, and transform data. • Select tables, records, and attributes. • Construct the final dataset for input to models.	**[ ] Stage 6: Deployment** • Report final results to customers • Plan deployment (e.g., integrate with legacy systems, and implement a repeatable DM process) • Plan monitoring and maintenance

12. The following factors might inhibit DM success or present problems for DM implementation. For each, please indicate the extent that the factor was a problem during Project P (1= not a problem, 2= a minor problem, 3= a significant problem, 4= a major problem, 5= an extreme problem).

Insufficient training	1	2	3	4	5
Insufficient tool support	1	2	3	4	5
Data unavailability	1	2	3	4	5
Too many potential leads	1	2	3	4	5
Complexity of data	1	2	3	4	5
The noise in the data	1	2	3	4	5
Data not current enough or missing critical data	1	2	3	4	5
Insufficient analysis of the client' requirements	1	2	3	4	5
The DM / analytic tools difficult to use	1	2	3	4	5
Lack of appropriate DM process	1	2	3	4	5
The chosen technology/tools not a match with the task	1	2	3	4	5
Scope of the DM project defined inappropriately	1	2	3	4	5
Difficulty in interpreting the DM result	1	2	3	4	5
Unrealistic expectations from the clients	1	2	3	4	5
Poor communication between DM team members and the clients	1	2	3	4	5
Poor management of the DM endeavor	1	2	3	4	5
Poor problem definition	1	2	3	4	5
Lack of expertise in statistics	1	2	3	4	5
Lack of top management support in DM efforts	1	2	3	4	5
Insufficient business knowledge	1	2	3	4	5
Insufficient knowledge in functional areas (e.g., marketing, finance, etc.)	1	2	3	4	5
Difficulty in finding DM team members who have the required skills and	1	2	3	4	5
Top management's insufficient understanding about DM effort	1	2	3	4	5
Lack of DM project champion	1	2	3	4	5
Lack of user participation and assistance in the DM effort	1	2	3	4	5
Insufficient knowledge in using database / data warehouse	1	2	3	4	5

Part IV General Information

1. Please indicate the primary business activity of your organization (e.g., Manufacturing, service). __________________

2. What is your organization's annual sales volume?_______(million $), and total number of employees?________

3. Where is your corporate headquarters?_______________

4. Does your organization have an enterprise data warehouse/data mart (note: a data warehouse stores historical data and is different from an operational database)? [ ] yes [ ] no

5. Your job title: ___________

6. Your highest education level: _______________, and field:______________________.

7. Your DM experience (in years): _______

8. How long have you worked in this organization (in years)?________

9. Your gender: [ ] Female [ ] Male

10. Your age (in years): ________

11. The following is a list of statistical / OR techniques. Please check one or both if the item has been used within DM, outside the DM activities, or both in your organization.

	Have used with DM	Have also Used outside DM
Descriptive statistics (mean, median, variance etc.)	[ ]	[ ]
Statistical visualization (cross-tabulation, histogram, scatter plot etc.)	[ ]	[ ]
Sampling and resampling (random, convenient sampling, bootstrap etc.)	[ ]	[ ]
Modeling with distribution (normal, Poisson etc.)	[ ]	[ ]
Inferential statistics (estimation, hypothesis testing)	[ ]	[ ]
Regression (simple, multiple)	[ ]	[ ]
Logistic regression	[ ]	[ ]
Clustering (k-means, etc.)	[ ]	[ ]
Data reduction techniques (factor analysis etc.)	[ ]	[ ]
Probability theory	[ ]	[ ]
Nonparametric techniques	[ ]	[ ]
Bayesian analysis	[ ]	[ ]
Analysis of Variance	[ ]	[ ]
Experimental design	[ ]	[ ]
Other multivariate techniques (MANOVA, MDS)	[ ]	[ ]
Forecasting & Time Series	[ ]	[ ]
Decision Analysis	[ ]	[ ]
CART, CHAID, MARS etc	[ ]	[ ]
Simulation	[ ]	[ ]
Linear/non-linear programming	[ ]	[ ]
Data Envelopment Analysis	[ ]	[ ]

You have reached the end of the survey. Thank you.

REFERENCES

Abajo, N. d., A. B. Diez, et al. (2004). "ANN Quality Diagnostic Models for Packaging Manufacturing: An Industrial Data Mining Case Study." Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington: 799-804.

Adderley, R. and P. B. Musgrove (2001). "Data Mining Case Study: Modeling the Behavior of Offenders Who Commit Serious Sexual Assaults." Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California: 215 - 220.

Agrawal, R., H. Mannila, et al. (1996). "Fast Discovery of Association Rules." In Fayyad, Usama, et al., Eds. Advance in Knowledge Discovery and Data Mining. MIT Press, Cambridge, MA.

Apte, C., B. Liu, et al. (2002). "Business Applications of Data Mining." Communications of the ACM 45(8): 49-53.

Bashein, B. and L. M. Markus (1997). "Credibility Equation for IT Specialists." Sloan Management Review 38(4): 35-44.

Bassellier, G. and I. Benbasat (2004). "Business Competence of Information Technology Professionals: Conceptual Development and Influence on IT-Business Partnerships." MIS Quarterly 28(4): 673-694.

Benbasat, I., D. K. Goldstein, et al. (1987). "The Case Research Strategy in Studies of Information Systems." MIS Quarterly 11(3): 369-386.

Berry, M. J. A. and G. S. Linoff (2000). "Mastering Data Mining." John Wiley & Sons, Inc. New York.

Berry, M. J. A. and G. S. Linoff (2004). "Data Mining Techniques: for Marketing, Sales, and Customer Relationship Management." John Wiley & Sons, Inc. New York.

Biehl, M. and e. a. H. Kim (2005). "Relationships among the Academic Business Disciplines: A Multi-method Citation Analysis." Omega 34(4): 123-148.

Bonoma, T. V. (1985). "Case Research in Marketing: Opportunities, Problems, and a Process." Journal of Marketing Research 22(XXII): 199-208.

Brachman, R. and T. Anand (1996). "The process of Knowledge Discovery in Databases: A Human-Centered Approach." In Advances in Knowledge Discovery and Data Mining: 37-58.

Brachman, R. J., T. Khabaza, et al. (1996). "Mining Business Databases." Communications of the ACM 39(11): 42-48.

Breiman, L., J. H. Friedman, et al. (1984). "Classification and Regression Trees." Wadsworth, Belmont, CA.

Cabena, P., P. Hadjinian, et al. (1998). "Discovery Data Mining: From Concept to Implementation." Prentice Hall, Englewood Cliffs, NJ.

Chang, S.-C., H.-C. Chang, et al. (2003). "The Effect of Organizational Attributes on the Adoption of Data Mining Techniques in the Financial Service Industry: An Empirical Study in Taiwan." International Journal of Management 20(4): 497-503.

Chapman, P., J. Clinton, et al. (2000). "CRISP-DM 1.0 Step-by-step Data Mining Guide." CRISP-DM Consortium, available at http;//www.crisp-dm.org.

Chatfield, C. (1995). "Model Uncertainty, Data Mining, and Statistical Inference." Journal of Royal Statistical Society: Series A 158: 419-466.

Chenoweth, T., K. Corral, et al. (2006). “Seven Key Interventions for Data Warehouse Success. Communications of the ACM 49(1): 114-119.

Chou, D. C. and A. Y. Chou (1999). "A Manager's Guide to Data Mining." Information Systems Management 16(4): 33-41.

Chung, H. M. and P. Gray (1999). "Special Section: Data Mining." Journal of Management Information Systems 16(1): 11-16.

Clogg, C. C. and E. S. Shihadeh (1994). "Statistical Models for Ordinal Variables." Sage Publications, Thousand Oaks, CA.

Cooper, B. L., H. J. Watson, et al. (2000). "Data Warehousing Supports Corporate Strategy at First American Corporation." MIS Quarterly 24(4): 547-567.

Cottrill, C. A., A. M. Rogers, et al. (1989). "Co-citation Analysis of the Scientific Literature of Innovation Research Traditions." Knowledge: Creation, Diffusion, Utilization 11: 181-208.

Crane, D. (1972). "Invisible Colleges: Diffusion of Knowledge in Scientific Communities." University of Chicago Press, Chicago, IL.

Culnan, M. J. (1986). "The Intellectual Development of Management Information Systems, 1972-1982: A Co-Citation Analysis." Management Science 32(2): 156-172.

Culnan, M. J. (1987). "Mapping the Intellectual Structure of MIS, 1980-1985: A Co-Citation Analysis." MIS Quarterly 11(3): 342-350.

Davenport, T. H., J. G. Harris, et al. (2001). "Data to Knowledge to Results." California Management Review 43(2): 117-138.

DeLone, W. D. and E. R. McLean (1992). "Information Systems Success: The Quest for the Dependent Variable." Information Systems Research 3(1): 60-95.

Eisenhardt, K. M. (1989). "Building Theories from Case Study Research." Academy of Management Review 14(4): 532-550.

Fayyad, U., G. Piatetsky-Shapiro, et al. (1996). "From Data Mining to Knowledge Discovery in Databases." AI Magazine 17(3): 37-54.

Fayyad, U., G. Piatetsky-Shapiro, et al. (1999). "The KDD Process for Extracting Useful Knowledge from Volumes of Data." Communications of the ACM 39(11): 27-34.

Fayyad, U., G. Piatetsky-Shapiro, et al. (2003). "Summary from the KDD-03 Panel---Data Mining: The Next 10 Years." ACM SIGKDD Explorations 5(2): 191-196.

Fayyad, U. and P. Stolorz (1997). "Data Mining and KDD: Promise and Challenges." Future Generation Computer Systems 13: 99-115.

Fayyad, U. and R. Uthurusamy (2002). "Evolving Data Mining into Solutions for Insights." Communications of the ACM 45(8): 28-31.

Fayyad, U. M., G. Piatetsky-Shapiro, et al. (1996). "Advances in Knowledge Discovery and Data Mining." AAAI Press, CA.

Friedman, J. H. (1997). "Data Mining and Statistics: What's the Connection?" In Computing Science and Statistics. Proceedings of the 29th Symposium on the Interface: 195-204.

Friedman, J. H. (1997). "The Role of Statistics in the Data Revolution?" International Statistical Review 69: 5-10.

Garver, M. S. (2002). "Using Data Mining for Customer Satisfaction Research." Marketing Research 14(1): 8-12.

Gehrke, J. (2002). "Report on the SIGKDD 2001 Conference Panel "New Research Directions in KDD'." ACM SIGKDD Explorations 3(2): 76-77.

Glaser, B. G. and A. L. Strauss (1967). "The Discovery of Grounded Theory, Strategies for Qualitative Research." Aldine de Gruyter, Hawthorne, New York.

Glymour, C., D. Madigan, et al. (1996). "Statistical Inference and Data Mining." Communications of the ACM 39(11): 35-41.

Glymour, C., D. Madigan, et al. (1997). "Statistical Themes and Lessons for Data Mining." Data Mining and Knowledge Discovery 1(1): 11-28.

Goldenstein, D. K. and J. F. Rockart (1984). "An Examination of Work-Related Correlates of Job Satisfaction in Programmer/Analysts." MIS Quarterly 8(2):103-115.

Goodman, A. (1999). "Report from Interface'98: Knowledge Discovery and the Interface of Computing and Statistics." ACM SIGKDD Explorations 1(1): 12-13.

Goodman, A. (2000). "Interface'99: A Data Mining Overview." ACM SIGKDD Explorations 1(2): 97-103.

Goodman, L. (1991). "New Methods for Analyzing the Intrinsic Character of Qualitative Variables Using Cross-classified Data." American Journal of Sociology 93: 529-583.

Green, G. I. (1989). "Perceived Importance of Systems Analysts' Job Skills, Roles, and Non-Salary Incentives." MIS Quarterly 13(2): 114-133.

Groth, R. (1998). "Data Mining - A Hands-On Approach for Business Professionals." Prentice Hall, Upper Saddle River, NJ.

Grover, V., S. R. Jeong, et al. (1995). "The Implementation of Business Process Reengineering." Journal of Management Information Systems 12(1): 109-144.

Han, J., R. B. Altman, et al. (2002). "Emerging Scientific Applications in Data Mining." Communications of the ACM 45(8): 54-58.

Han, J. and M. Kamber (2001). "Data Mining: Concepts and Techniques." Academic Press, San Diego, CA.

Hand, D. J. (1998). "Data Mining: Statistics and More?" The American Statisticians 52(2): 112-118.

Hand, D. J. (1999). "Statistics and Data Mining: Intersecting Disciplines." ACM SIGKDD Explorations 1(1): 16-19.

Hand, D. J. (2000). "Data Mining: New Challenges for Statisticians." Social Science Computing Review 18(4): 442-449.

Hand, D. J., G. Blunt, et al. (2000). "Data Mining for Fun and Profit." Statistical Science 15(2): 111-131.

Hirji, K. K. (2001). "Exploring Data Mining Implementation." Communications of the ACM 44(7): 87-93.

Hofmann, M. and B. Tierney (2003). "The Involvement of Human Resources in Large Scale Data Mining Projects." International Symposium on Information and Communication Technologies, Dublin, Ireland: 24-26.

Hotz, E., U. Grimmer, et al. (2001). "REVI-MINER, a KDD-Environment for Deviation Detection and Analysis of Warranty and Goodwill Cost Statements in Automotive Industry." Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA: 432-437.

Hymes, D. H. (1971). "On Communicative Competence." Sociolinguistics, J. B. Pride and J. Holmes, London.

ICDM (2004). "Data Mining - Where to go?" ICDM 2004 Panel Discussion.

Inmon, W. H. (1996). "The Data Warehouse and Data Mining." Communications of the ACM 39(11): 49-50.

Kass, G. V. (1980). "An Exploratory Technique for Investigating Large Quantities of Categorical Data." Applied Statistics 29(2): 119-127.

Kettenring, J. R. (1997). "Shaping Statistics for Success in the 21st Century." Journal of the American Statistical Association 92(440): 1229-1234.

Kleinberg, J., C. Papadimitriou, et al. (1998). "A Microeconomic View of Data Mining." Data Mining and Knowledge Discovery 2: 311-324.

Ko, D.-G., L. J. Kirsch, et al. (2005). "Antecedents of Knowledge Transfer from Consultants to Clients in Enterprise System Implementations." MIS Quarterly 29(1): 59-85.

Kohavi, R. and F. Provost (2001). "Applications of Data Mining to Electronic Commerce." Data Mining and Knowledge Discovery 5: 5-10.

Kohavi, R., N. J. Rothleder, et al. (2002). "Emerging Trends in Business Analytics." Communications of the ACM 45(8): 45-48.

Koo, S. (1998). "Interview with Knowledge Discovery Nuggets^{TM} owner: Gregory Piatetsky-Shapiro." I.T. Times - a supplement of Hong Kong Economic Times.

Lee, D. M. S., E. M. Trauth, et al. (1995). "Critical Skills and Knowledge Requirements of IS Professionals: A Joint Academic/Industry Investigation." MIS Quarterly: 19(3): 313-340.

Loveman, G. (2003). "Diamonds in the Data Mine." Harvard Business Review: 109-113.

Luan, J. (2002). "Data Mining and Knowledge Management in Higher Education." Cabrillo College, Presentation at the Annual Forum for the Association for Institutional Research, Toronto, Ontario, Canada.

Mackinnon, M. J. and N. Glick (1999). "Data Mining and Knowledge Discovery in Databases - An Overview." Australia & New Zealand Journal of Statistics 41(3): 255-275.

Mannila, H. (2000). "Theoretical Frameworks for Data Mining." ACM SIGKDD Explorations 1(2): 30-32.

McCain, K. W. (1990). "Mapping Authors in Intellectual Space: A Technical Overview." Journal of the American Society for Information Science 41(6): 433-443.

META Group (2004). "Data Mining Tools Market Summary." META Group, Inc.

Muggleton, S. H. and L. D. Raedt (1994). "Inductive Logic Programming: Theory and Methods." Journal of Logic Programming 19(20): 629-679.

Nemati, H. R. and C. D. Barko (2001). "Issues in Organizational Data Mining: A Survey of Current Practices." Journal of Data Warehousing 6(1): 25-36.

Nemati, H. R. and C. D. Barko (2003). "Key Factors for Achieving Organizational Data-mining Success." Industrial Management & Data Systems 103(4): 282-292.

Piatetsky-Shapiro, G. (1991). "Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop." AI Magazine 11(5): 68-70.

Piatetsky-Shapiro, G. (1999). "The Data Mining Industry Coming of Age." IEEE Intelligent Systems: 32-34.

Piatetsky-Shapiro, G. (2001). "KDnuggets Interview with Usama Fayyad." ACM SIGKDD Explorations 3(1): 45-48.

Piatetsky-Shapiro, G. (2002). "Knowledge Discovery in Databases: 10 years after." ACM SIGKDD Explorations 1(2): 59-61.

Piatetsky-Shapiro, G., R. Brachman, et al. (1996). "An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications." In E. Simoudis, J. Han and U. Fayyad (Eds.), KDD-96: the Second International Conference on Knowledge Discovery and Data Mining. AAAI Press, Menlo Park, California: 89-95.

Piatetsky-Shapiro, G., T. Khabaza, et al. (2003). "Capturing Best Practice for Microarray Gene Expression Data Analysis." Proceedings of the Ninth International Conference on Knowledge Discovery and Data Mining, August 24-27, Washington, DC.

Pieters, R. and H. Baumgartner (2002). "Who Talks to Whom? Intra- and Interdisciplinary Communication of Economics Journals." Journal of Economic Literature 40(2): 483-508.

Pieters, R. and e. a. H. Baumgartner (1999). "Importance and Similarity in the Evolving Citation Network of the International Journal of Research in Marketing." International Journal of Research in Marketing 16(2): 113-127.

Ponzi, L. J. (2002). "The Intellectual Structure and Interdisciplinary Breadth of Knowledge Management: A Bibliometric Study of its Early Stage of Development." Scientometrics 55(2): 259-272.

Quinlan, R. (2000). "KDD-99 Panel on Last 10 and Next 10 Years." ACM SIGKDD Explorations 1(2): 62.

Quinlan, R. J. (1986). "Induction of Decision Trees." Machine Learning 1(1): 81-106.

Quinlan, R. J. (1993). "C4.5: Programs for Machine Learning." Morgan Kaufmann Publishers, Inc. San Mateo, CA.

Robey, D., J. W. Ross, et al. (2002). "Learning to Implement Enterprise Systems: An Exploratory Study of the Dialectics of Change." Journal of Management Information Systems 19(1): 17-46.

Rockart, J., M. Earl, et al. (1996). "Eight Imperatives for the New IT Organization." Sloan Management Review 38(1): 43-55.

Sanders, G. L. and J. F. Courtney (1985). "A Field Study of Organizational Factors Influencing DSS Success." MIS Quarterly 9(1): 77-93.

Schaefer, D. R. and D. A. Dillman (1998). "Development of a Standard E-mail Methodology: Results of an Experiment." Public Opinion Quarterly 62(3): 378-393.

Senn, J. A. (1978). "A Management View of Systems Analysts: Failures and Shortcomings." MIS Quarterly 2(3): 25-32.

Sim, J. (2003). "Critical Success Factors in Data Mining Projects." unpublished dissertation, University of North Texas.

Sircar, S., S. P. Nerur, et al. (2001). "Revolution or Evolution? A Comparison of Object-Oriented and Structured Systems Development Methods." MIS Quarterly 25(4): 457-464.

Strout, E. (1971). "The Activities and Education of Systems Analysts." Journal of Systems Management 22(1): 37-40.

Teng, J. T. C., S. R. Jeong, et al. (1998). "Profiling Successful Reengineering Projects." Communications of the ACM 41(6): 96-102.

Todd, P. A., J. D. McKeen, et al. (1995). "The Evolution of IS Job Skills: A Content Analysis of IS Job Advertisements from 1970 to 1990." MIS Quarterly 19(1): 1-27.

White, H. D. and B. C. Griffith (1981). "Author Cocitation: A Literature Measure of Intellectual Structure." Journal of the American Society for Information Science 32:163-171.

White, H. D. and K. W. McCain (1998). "Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972-1995." Journal of the American Society for Information Science 49(4): 327-355.

White, k. B. (1984). "MIS Project Teams: An Investigation of Cognitive Style Implications." MIS Quarterly 8(2): 95-101.

White, K. B. and R. Leifer (1986). "Information Systems Development Success: Perspectives from Project Team Participants." MIS Quarterly 10(3): 215-223.

Wixom, B. H. and H. J. Watson (2001). "An Empirical Investigation of the Factors Affecting Data Warehousing Success." MIS Quarterly 25(1): 17-41.

Yin, R. K. (1994). "Case Study Research: Design and Methods." Sage Publications, Newbury Park, CA.

Zinkhan, G. M. and T. Leigh (1999). "Accessing the Quality Ranking of the Journal of Advertising, 1986-1997." Journal of Advertising 28(2): 51-70.

Zinkhan, G. M. and M. S. Roth (1992). "Knowledge Development and Scientific Status in Consumer-Behavior Research: A Social Exchange Perspective." Journal of Consumer Research 19(2): 282-291.

Zmud, R. W. (1979). "Individual Differences in MIS Success: A Review of the Empirical Literature." Management Science 25(10): 966-979.

BIOGRAPHICAL INFORMATION

Ailing Wang received her Doctoral degree in Business Administration from The University of Texas at Arlington. Her research focuses on Statistics and Data Mining, Knowledge and Data Management, Performance Metrics, and Information Security in various areas of business.