Preface

How can a data-flooded manager get out of the "mire"? How can a confused decision maker pass through a "maze"? How can an over-burdened problem solver clean up a "mess"? How can an exhausted scientist decipher a "myth"?

The answer is an interdisciplinary subject and a powerful tool known as data mining (DM). DM can turn data into dollars; transform information into intelligence; change pattern into profit; and convert relationship into resources.

As the *third* branch of operations research and management science (OR/MS) and the *third* milestone of data management, DM can help attack the *third* category of decision making by elevating our raw data into the *third* stage of knowledge creation.

The term "third" has been mentioned four times above. Let's go backward and look at the three stages of knowledge creation. Managers are often drowning in data (the first stage) but starving for knowledge. A collection of data is not information (the second stage); and a collection of information is not knowledge. Data begets information which begets knowledge. The whole subject of DM has a synergy of its own and represents more than the sum of its parts.

There are three categories of decision making: structured, semi-structured and unstructured. Decision making processes fall along a continuum that ranges from highly structured decisions (sometimes called programmed) to highly unstructured (non-programmed) decisions (Turban et al., 2005, p. 12).

At one end of the spectrum, structured processes are routine and typically repetitive problems for which standard solutions exist. Unfortunately, rather than being static, deterministic and simple, the majority of real world problems are dynamic, probabilistic, and complex. Many professional and personal problems are classified as unstructured, or marginally as semi-structured, or even in between, since the boundaries between them may not be crystal-clear.

In addition to developing normative models (such as linear programming, economic order quantity) for solving *structured* (or *programmed*) problems, operation researchers and management scientists have created many descriptive models, such as simulation and goal programming, to deal with semi-structured alternatives. Unstructured problems, however, fall in a gray areas for which there are no cut-and-dry solution methods. The current two branches of OR/MS hit a dead end with unstructured problems.

To gain knowledge, one must understand the patterns that emerge from information. Patterns are not just simple relationships among data; they exist separately from information, as archetypes or standards to which emerging information can be compared so that one may draw inferences and take action. Over the last 40 years, the tools and techniques used to process data and information have continued to evolve from databases (DBs) to data warehousing (DW) and further to DM. DW applications, the middle of these three stages, have become business-critical. However, DM can help deliver even more value from these huge repositories of information.

Certainly, there are many statistical models that have emerged over time. Machine learning has marked a milestone in the evolution of computer science (Fayyad, Piatetsky-Shapiro, Smyth & Uthurusamy, 1996). Although DM is still in its infancy, it is now being used in a wide range of industries and for a range of tasks in a variety of contexts (Wang, 2003). DM is synonymous with knowledge discovery in databases, knowledge extraction, data/pattern analysis, data archeology, data dredging, data snooping, data fishing, information harvesting, and business intelligence (Giudici, 2003; Hand et al., 2001; Han & Kamber, 2000). There are unprecedented opportunities in the future to utilize DM.

Data warehousing and mining (DWM) is the science of managing and analyzing large datasets and discovering novel patterns. In recent years, DWM has emerged as a particularly exciting and industrially relevant area of research. Prodigious amounts of data are now being generated in domains as diverse and elusive as market research, functional genomics and pharmaceuticals. Intelligently analyzing data to discover knowledge with the aim of answering crucial questions and helping make informed decisions is the challenge that lies ahead.

The *Encyclopedia of Data Warehousing and Mining* provides theories, methodologies, functionalities, and applications to decision makers, problem solvers, and data miners in business, academia, and government. DWM lies

at the junction of database systems, artificial intelligence, machine learning and applied statistics, which makes it a valuable area for researchers and practitioners. With a comprehensive overview, The *Encyclopedia of Data Warehousing and Mining* offers a thorough exposure to the issues of importance in this rapidly changing field. The encyclopedia also includes a rich mix of introductory and advanced topics while providing a comprehensive source of technical, functional and legal references to DWM.

After spending more than a year preparing this book, with a strictly peer-reviewed process, I am delighted to see it published. The standard for selection was very high. Each article went through at least three peer reviews; additional third-party reviews were sought in cases of controversy. There have been innumerable instances where this feedback has helped to improve the quality of the content, and even influenced authors in how they approach their topics.

The primary objective of this encyclopedia is to explore the myriad of issues regarding DWM. A broad spectrum of practitioners, managers, scientists, educators, and graduate students who teach, perform research, and/or implement these discoveries, are the envisioned readers of this encyclopedia.

The encyclopedia contains a collection of 234 articles, written by an international team of 361 experts representing leading scientists and talented young scholars from 34 countries. They have contributed great effort to create a source of solid, practical information, informed by sound underlying theory that should become a resource for all people involved in this dynamic new field. Let's take a peek at a few articles:

The evaluation of DM methods requires a great deal of attention. A valid model evaluation and comparison can improve considerably the efficiency of a DM process. Paolo Giudici has presented several ways to perform model comparison, in which each has its advantages and disadvantages.

According to Zbigniew W. Ras, the main object of action rules is to generate special types of rules for a database that point the direction for re-classifying objects with respect to some distinguishing attributes (called decision attributes). This creates flexible attributes that form a basis for action rules construction.

With the constraints imposed by computer memory and mining algorithms, we can experience selection pressures more than ever. The main point of instance selection is *approximation*. Our task is to achieve as good mining results as possible by approximating the whole dataset with the selected instances and hope to do better in DM with instance selection as it is possible to remove noisy and irrelevant data in the process. Huan Liu and Lei Yu have presented an initial attempt to review and categorize the methods of *instance selection* in terms of sampling, classification, and clustering.

Shichao Zhang and Chengqi Zhang introduce a group of pattern discovery systems for dealing with the multiple data source (MDS) problem, mainly including a logical system for enhancing data quality; a logical system for resolving conflicts; a data cleaning system; a database clustering system; a pattern discovery system and a post-mining system.

Based on his extensive experience, Gautam Das surveys recent state-of-the-art solutions to the problem of *approximate query answering* in databases, in which "ballpark answers"(i.e., approximate answers) to queries can be provided within acceptable time limits. These techniques sacrifice accuracy to improve running time; typically through some sort of lossy data compression. Also, Han-Joon Kim (the holder of two patents on text mining applications) discusses a comprehensive text-mining solution to document indexing problems on topic hierarchies (taxonomy).

Condensed representations have been proposed as a useful concept for the optimization of typical DM tasks. It appears as a key concept within the emerging inductive DB framework where inductive query evaluation needs for effective constraint-based DM techniques. Jean-François Boulicaut introduces this research domain, its achievements in the context of frequent itemset mining from transactional data and its future trends.

Zhi-Hua Zhou discusses complexity issues in DM. Although we still have a long way to go in order to produce patterns that can be understood by most people involved with DM tasks, endeavors on improving the comprehensibility of complicated algorithms have proceeded at a promising pace.

Pattern classification poses a difficult challenge in finite settings and high dimensional spaces caused by the issue of dimensionality. Carlotta Domeniconi and Dimitrios Gunopulos discuss classification techniques, including the authors' own work, to mitigate the problem of dimensionality and reduce bias, by estimating local feature relevance and selecting features accordingly. This issue has both theoretical and practical relevance, since learning tasks abound in which data are represented as a collection of a very large numbers of features. Thus, many applications can benefit from improvements in predicting error.

Qinghua Zou proposes using *pattern decomposition algorithms* to find frequent patterns in large datasets. Pattern decomposition is a DM technology that uses known, frequent or infrequent patterns to decompose long itemsets to many short ones. It identifies frequent patterns in a dataset using a bottom-up methodology and reduces the size of the dataset in each step. The algorithm avoids the process of candidate set generation and decreases the time for counting supports due to the reduced dataset.

Perrizo, Ding, et al. review a category of DM approaches using vertical data structures. They demonstrate their applications in various DM areas, such as association rule mining and multi-relational DM. Vertical DM strategy aims at addressing scalability issues by organizing data in vertical layouts and conducting logical operations on vertically partitioned data instead of scanning the entire DB horizontally.

Integration of data sources refers to the task of developing a common schema, as well as data transformation solutions, for a number of data sources with related content. The large number and size of modern data sources makes manual approaches to integration increasingly impractical. Andreas Koeller provides a comprehensive overview over DM techniques which can help to partially or fully automate the data integration process.

DM applications often involve testing hypotheses regarding thousands or millions of objects at once. The statistical concept of multiple hypothesis testing is of great practical importance in such situations, and an appreciation of the issues involved can vastly reduce errors and associated costs. Sach Mukherjee provides an introductory look at multiple hypothesis testing in the context of DM.

Maria Vardaki illustrates the benefits of using statistical metadata by information systems, depicting also how such standardization can improve the quality of statistical results. She proposes a common, semantically rich, and object-oriented data/metadata model for metadata management that integrates the main steps of data processing and covers all aspects of DW that are essential for DM requirements. Finally, she demonstrates how a metadata model can be integrated in a web-enabled statistical information system to ensure quality of statistical results.

A major obstacle in DM applications is the gap between statistic-based pattern extraction and value-based decision-making. *Profit mining* aims at reducing this gap. The concept and techniques proposed by Ke Wang and Senqiang Zhou are applicable to applications under a general notion of "utility".

Although a tremendous amount of progress has been made in DM over the last decade or so, many important challenges still remain. For instance, there are still no solid standards of practice; it is still too easy to misuse DM software; *secondary* data analysis without appropriate experimental design is still common; and it is still hard to choose right kind of analysis methods for the problem in hand. Xiao Hui Liu points out that *intelligent data analysis* (IDA) is an interdisciplinary study concerning the effective analysis of data, which may help advance the state of art in the field.

In recent years, the need to extract complex tree-like or graph-like patterns in massive data collections (e.g., in bioinformatics, semistructured or Web DBs) has become a necessity. This has led to the emergence of the research field of graph and tree mining. This field provides many promising topics for both theoretical and engineering achievements, and many expect this to be one of the key fields in DM research in the years ahead. Katsaros and Manolopoulos review the most important strategic application-domains where *frequent structure mining* (FSM) provides significant results. A survey is presented of the most important algorithms that have been proposed for mining graph-like and tree-like substructures in massive data collections.

Lawrence B. Holder and Diane J. Cook are among the pioneers in the field of graph-based DM and have developed the widely-disseminated Subdue graph-based DM system (http://ailab.uta.edu/subdue). They have directed multi-million dollar government-funded projects in the research, development and application of graph-based DM in real-world tasks ranging from bioinformatics to homeland security.

Graphical models such as *Bayesian networks* (BNs) and *decomposable Markov networks* (DMNs) have been widely applied to probabilistic reasoning in intelligent systems. Automatic discovery of such models from data is desirable, but is NP-hard in general. Common learning algorithms use single-link look-ahead searches for efficiency. However, *pseudo-independent* (PI) probabilistic domains are not learnable by such algorithms. Yang Xiang introduces fundamentals of PI domains and explains why common algorithms fail to discover them. He further offers key ideas as to how they can efficiently be discovered, and predicts advances in the near future.

Semantic DM is a novel research area that used graph-based DM techniques and ontologies to identify complex patterns in large, heterogeneous data sets. Tony Hu's research group at Drexel University is involved in the development and application of semantic DM techniques to the bioinformatics and homeland security domains.

Yu-Jin Zhang presents a novel method for image classification based on feature element through association rule mining. The feature elements can capture well the visual meanings of images according to the subjective perception of human beings, and are suitable for working with rule-based classification models. Techniques are adapted for mining the association rules which can find associations between the feature elements and class attributes of the image, and the mined rules are applied to image classifications.

Results of image DB queries are usually presented as a thumbnail list. Subsequently, each of these images can be used for refinement of the initial query. This approach is not suitable for queries by sketch. In order to receive the desired images, the user has to recognize misleading areas of the sketch and modify these images appropriately. This is a non-

trivial problem, as the retrieval often is based on complex, non-intuitive features. Therefore, Odej Kao presents a *mosaic-based technique* for sketch feedback, which combines the best sections contained in an image DB into a single query image.

Andrew Kusiak and Shital C. Shah emphasize the need for an individual-based paradigm, which may ensure the well-being of patients and the success of pharmaceutical industry. The new methodologies are illustrated with various medical informatics research projects on topics such as predictions for dialysis patients, significant gene/SNP identifications, hypoplastic left heart syndrome for infants, and epidemiological and clinical toxicology. DWM and data modeling will ultimately lead to targeted drug discovery and individualized treatments with minimum adverse effects.

The use of microarray DBs has revolutionized the way in which biomedical research and clinical investigation can be conducted in that high-density arrays of specified DNA sequences can be fabricated onto a single glass slide or "chip". However, the analysis and interpretation of the vast amount of complex data produced by this technology poses an unprecedented challenge. LinMin Fu and Richard Segall present a state-of-the-art review of microarray DM problems and solutions.

Knowledge discovery from genomic data has become an important research area for biologists. An important characteristic of genomic applications is the very large amount of data to be analyzed, and most of the time, it is not possible to apply only classical statistical methods. Therefore, Jourdan, Dhaenens and Talbi propose to model knowledge discovery tasks associated with such problems as combinatorial optimization tasks, in order to apply efficient optimization algorithms to extract knowledge from those large datasets.

Founded on the work of Indrani Chakravarty et al.'s research, handwritten signature is a behavioral biometric. There are two methods used for recognition of handwritten signatures – offline and online. While offline methods extract static features of signature instances by treating them as images, online methods extract and use temporal or dynamic features of signatures for recognition purposes. Temporal features are difficult to imitate, and hence online recognition methods offer higher accuracy rates than offline methods.

Neurons are small processing units that are able to store some information. When several neurons are connected, the result is a neural network, a model inspired by biological neural networks like the brain. Kate Smith provides useful guidelines to ensure successful learning and generalization of the neural network model. Also, a special version in the form of *probabilistic neural networks* (PNNs) is explained by Ingrid Fischer with the help of graphic transformations.

The sheer volume of multimedia data available has exploded on the Internet in the past decade in the form of webcasts, broadcast programs and streaming audio and video. Automated content analysis tools for multimedia depend on face detectors and recognizers; videotext extractors; speech and speaker identifiers; people/vehicle trackers; and event locators resulting in large sets of multimodal features that can be real-valued, discrete, ordinal, or nominal. Multimedia metadata based on such a multimodal collection of features, poses significant difficulties to subsequent tasks such as classification, clustering, visualization and dimensionality reduction – which traditionally deal only with continuous-valued data. Aradhye and Dorai discuss mechanisms that extend tasks traditionally limited to continuous-valued feature spaces to multimodal multimedia domains with symbolic and continuous-valued features, including (a) dimensionality reduction, (b) de-noising, (c) visualization, and (d) clustering.

Brian C. Lovell and Shaokang Chen review the recent advances in the application of face recognition for multimedia DM. While the technology for mining text documents in large DBs could be said to be relatively mature, the same cannot be said for mining other important data types such as speech, music, images and video. Yet these forms of multimedia data are becoming increasingly common on the Internet and intranets.

The goal of Web usage mining is to capture, model, and analyze the behavioral patterns and profiles of users interacting with a Web site. Bamshad Mobasher and Yongjian Fu provide an overview of the three primary phases of the Web mining process: data preprocessing, pattern discovery, and pattern analysis. The primary focus of their articles is on the types of DM and analysis tasks most commonly used in Web usage mining, as well as some their typical applications in areas such as Web personalization and Web analytics. Ji-Rong Wen explores the ways of enhancing Web search using *query log mining* and *Web structure mining*.

In line with Mike Thelwall's opinion, *scientific Web intelligence* (SWI) is a research field that combines techniques from DM, Web intelligence and scientometrics to extract useful information from the links and text of academic-related Web pages, using various clustering, visualization and counting techniques. SWI is a type of Web mining that combines Web structure mining and text mining. Its main uses are in addressing research questions concerning the Web, or Web-related phenomena, rather than in producing commercially useful knowledge.

Web-enabled electronic business is generating massive amount of data on customer purchases, browsing patterns, usage times and preferences at an increasing rate. DM techniques can be applied to all the data being collected. Richi Nayak presents issues associated with DM for Web-enabled electronic-business.

Tobias Scheffer gives an overview of common email mining tasks including email filing, spam filtering and mining communication networks. The main section of his work focuses on recent developments in mining email data for support of the message creation process. Approaches to mining question-answer pairs and sentences are also reviewed.

Stanley Loh describes a computer-supported approach to mine discussions that occurred in chat rooms. Dennis Mcleod explores incremental mining from news streams. JungHwan Oh summarizes the current status of video DM. J. Ben Schafer addresses the technology used to generate recommendations.

In the abstract, a DW can be seen as a set of materialized views defined over source relations. During the initial design of a DW, the designer faces the problem of deciding which views to materialize in the DW. This problem has been addressed in the literature for different classes of queries and views, and with different design goals. Theodoratos and Simitsis identify the different design goals used to formulate alternative versions of the problem and highlight the techniques used to solve it.

Michel Schneider addresses the problem of designing a DW schema. He suggested a general model for this purpose that integrates a majority of existing models: the notion of a well-formed structure is proposed to help design the process; a graphic representation is suggested for drawing well-formed structures; and the classical star-snowflake structure is represented.

Anthony Scime presents a methodology for adding external information from the World Wide Web to a DW, in addition to the DW's domain information. The methodology assures decision makers that the added Web based data are relevant to the purpose and current data of the DW.

Privacy and confidentiality of individuals are important issues in the information technology age. Advances in DM technology have increased privacy concerns even more. Jack Cook and Yücel Sayg1n highlight the privacy and confidentiality issues in DM, and survey state of the art solutions and approaches for achieving privacy preserving DM.

Ken Goodman provides one of the first overviews of ethical issues that arise in DM. He shows that while privacy and confidentiality often are paramount in discussions of DM, other issues – including the characterization of appropriate uses and users, and data miners' intentions and goals – must be considered. Machine learning in genomics and in security surveillance are set aside as special issues requiring attention.

Increased concern about privacy and information security has led to the development of privacy preserving DM techniques. Yehuda Lindell focuses on the paradigms for defining security in this setting, and the need for a rigorous approach. Shamik Sural et al. present some of important approaches to privacy protection in *association rule mining*.

Human-computer interaction is crucial in the knowledge discovery process in order to accomplish a variety of novel goals of DM. In Shou Hong Wang's opinion, *interactive visual* DM is human-centered DM, implemented through knowledge discovery loops coupled with human-computer interaction and visual representations.

Symbiotic DM is an evolutionary approach that shows how organizations analyze, interpret, and create new knowledge from large pools of data. Symbiotic data miners are trained business and technical professionals skilled in applying complex DM techniques and business intelligence tools to challenges in a dynamic business environment. Athappilly and Rea opened the discussion on how businesses and academia can work to help professionals learn, and fuse the skills of business, IT, statistics, and logic to create the next generation of data miners.

Yiyu Yao and Yan Zhao first make an immediate comparison between scientific research and DM and add an explanation construction and evaluation task to the existing DM framework. Explanation-oriented DM offers a new perspective, which has a significant impact on the understanding of the complete process of DM and effective applications of DM results.

Traditional DM views the output from any DM initiative as a homogeneous knowledge product. Knowledge however, always is a multifaceted construct, exhibiting many manifestations and forms. It is the thesis of Nilmini Wickramasinghe's discussion that a more complete and macro perspective, and a more balanced approach to knowledge creation, can best be provided by taking a broader perspective of the knowledge product resulting from the KDD process: namely, by incorporating a people-based perspective into the traditional KDD process, and viewing knowledge as the multifaceted construct it is. This in turn will serve to enhance the knowledge base of an organization, and facilitate the realization of effective knowledge.

Fabrice Muhlenbach and Ricco Rakotomalala are the authors of an original *supervised multivariate discretization* method called *HyperCluster Finder*. Their major contributions to the research community are present in a DM software called TANAGRA, which is freely available on Internet.

Recently there have been many efforts to apply DM techniques to security problems, including homeland security and cyber security. Bhavani Thuraisingham (the inventor of three patents for MITRE) examines some of these developments in DM in general and link analysis in particular, and shows how DM and link analysis techniques may be applied for homeland security applications. Some emerging trends are also discussed.

In order to reduce financial statement errors and fraud, Garrity, O'Donnell and Sanders proposed an architecture that provides auditors with a framework for an effective continuous auditing environment that utilizes DM.

The applications of DWM are everywhere: from *Kernel Methods in Chemoinformatics* to *Data Mining for Damage Detection in Engineering Structures*; from *Predicting Resource Usage for Capital Efficient Marketing* to *Mining for Profitable Patterns in the Stock Market*; from *Financial Ratio Selection for Distress Classification* to *Material Acquisitions Using Discovery Informatics Approach*; from *Resource Allocation in Wireless Networks* to *Reinforcing CRM with Data Mining*; from *Data Mining Medical Digital Libraries* to *Immersive Image Mining in Cardiology*; from *Data Mining in Diabetes Diagnosis and Detection* to *Distributed Data Management of Daily Car Pooling Problems*; and from *Mining Images for Structure* to *Automatic Musical Instrument Sound Classification*…The list of DWM applications is endless and the future of DWM is promising.

Knowledge explosion pushes DWM, a multidisciplinary subject, to ever-expanding regions. Inclusion, omission, emphasis, evolution and even revolution are part of our professional life. In spite of our efforts to be careful, should you find any ambiguities or perceived inaccuracies, please contact me at *wangj@mail.montclair.edu.*

REFERENCES

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in knowledge discovery and data mining.* AAAI/MIT Press.

Giudici, P. (2003*). Applied data mining: Statistical methods for business and industry.* John Wiley.

Han, J., & Kamber, M. (2000). *Data mining: Concepts and techniques.* Morgan Kaufmann Publishers.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. MIT Press.

Turban, E., Aronson, J. E., & Liang, T. P. (2005). *Decision support systems and intelligent systems.* Upper Saddle River, NJ: Pearson Prentice Hall.

Wang, J. (2003). *Data mining: Opportunities and challenges.* Hershey, PA: Idea Group Publishing.