Published in final edited form as:
*Health Serv Manage Res*. 2010 February ; 23(1): 42–46. doi:10.1258/hsmr.2009.009029.

# Data mining for health executive decision support: an imperative with a daunting future!

**Saundra Glover, PhD MBA [Director]**,
Institute for Partnerships to Eliminate Health Disparities, University of South Carolina, Columbia, SC, USA

**Patrick A Rivers, PhD MBA [Professor and Director]**,
Health Care Management, College of Applied Arts and Sciences, Southern Illinois University Carbondale, Carbondale, IL, USA

**Derek A Asoh, PhD [Assistant Professor]**,
School of Information Systems and Applied Technologies, College of Applied Arts and Sciences, Southern Illinois University Carbondale, Carbondale, IL, USA

**Crystal N Piper, PhD MHA MPH [Research Associate]**, and
Institute for Partnerships to Eliminate Health Disparities, University of South Carolina, Columbia, SC, USA

**Keva Murph [Research Assistant]**
Institute for Partnerships to Eliminate Health Disparities, University of South Carolina, Columbia, SC, USA

## Summary

Data mining is highly profiled. It has the potential to enhance executive information systems. Such enhancement would mean better decision-making by management, which in turn would mean better services for customers. While the future of data mining as technology should be exciting, some are worried about privacy concerns, which make the future of data mining daunting. This paper examines why data mining is highly profiled – the imperative toward data mining, data mining models and processes. Additionally, the paper examines some of the benefits and challenges of using data mining processes within the health-care arena. We cast the future of data mining by highlighting two of the many data mining tools available – one commercial and one freely available. Subsequently, we discuss a number of social and technical factors that may thwart the extensive deployment of data mining, especially when the intent is to know more about the people that organizations have to serve and cast a view of what the future holds for data mining. This component is especially important when attempting to determine the longevity of data mining within health-care organizations. It is hoped that our discussions would be useful to organizations as they engage data mining, strategies for executive information systems and information policy issues.

## Introduction

Executive information systems (EIS) are specialized information decision support systems (DSS) that provide top management with actionable information for decision-making. Once exclusive to the private sector, EIS began to gain ground in the public sector more than a decade ago.[1] These systems are likely to become every-day management tools in both

privers@siu.edu.

sectors because of increase demand on management to make faster and better decisions in turbulent business environments.

From the perspective of strategic and tactical decision-making, EIS may be considered the major consumer in the Corporate Information Factory,[2] where data, its basic raw material, are traditionally found. Data and models constitute the core of EIS. Some experts point out that EIS need 'access to a broader sweep of data, often at an increased level of detail, with the ability to access and analyse these data on an as-needed basis'.[1] Evidently, the power and usefulness of an EIS is affected in part by the breadth and depth of its data, and the type of models used in analysing the data. The breadth and depth of data determines the level to which executives can 'drill down' to understand business problems before making decisions. The idea of automatically discovering knowledge from databases is a 'very attractive and challenging task for executives'.[3] The models determine whether analysis on a as-need basis is possible or not, and whether executives can explore and extrapolate trends.

Despite the long existence of historic and transactional data in organizations, the tools and techniques for effectively acquiring data for EIS support have remained rudimentary until recently.[4] Years of effort in data mining (DM) have produced a variety of efficient techniques.[5] Advances in media storage technology have made it possible to accumulate huge banks of data in data warehouses. In effect, creating data warehouses through data warehousing (DW) entails the collections of current and historical operational data stored for use in executive support systems and DSS.[6] Typically, a data warehouse is structured to include three components – an operational data store, a data warehouse and a data mart.[4] The sense of a data warehouse component refers to the portion of the data warehouse that has multiple layers of operational data stores that has been cleaned and subjected to quality metrics. The sense of the operational data store refers to the traditional databases with data pertaining to daily business transactions, and a data mart refers to the collection of operational databases, often organized on a functional basis or otherwise, and used to provide decision support for a small group of users.

EIS requires both tactical and strategic decision-making capabilities. The former is derived from operational data stores, while the latter is from data warehouses and data marts. Sifting and mining chunks of data warehouses (i.e. the component) is the process referred to as DM or knowledge discovery in databases (KDD). The power of an EIS is often not fully leveraged because of the lack of data. The emergence of DW and hence DM is timely as it provides structured management data for decision support.[7] Given its timely arrival and potential to provide the much-needed data for decision support, DM was recently highly profiled and rated by researchers at MIT as one of the top 10 emerging technologies that will change the world.[8] With the pace and capabilities of current technology, it has also been speculated that in the millennium ahead, more attention will be focused on DM,[9] which eventually will emerge as an organizational imperative.

## The DM imperative

Although the speculation on DM by Kimball[9] is technology oriented, there are other compelling reasons why DM should be highly profiled, and why speculations should be holistic. In today's turbulent business environment, private companies are increasingly being exposed to several threats, directly or indirectly associated with customers, competition and information technology (IT). Customers continue to demand more services at lower costs and are more likely to churn than ever. Competition is on the rise as a result of globalization and the emergence of new business models. The pace of IT is unsettling, providing mixed opportunities for organizations and individuals alike. Public organizations are not immune to the above threats. The citizenry is looking to government agencies to provide the same or

even better quality of services as private companies.[10] With terrorism looming and becoming a global phenomenon, there is considerable pressure for safety exerted by the citizenry on public organizations.

While many may invoke issues of privacy invasion, faced with an unpredictable environment, public and private sector executives and management must know more about the people (customers) they serve and in better ways than before. The key resource required by executives and management in making the best decisions to meet the needs of customers is knowledge. DM or KDD provides the means through which companies can better understand hidden customers' behaviours in order to adapt accordingly to beat the competition.

As DM enables the discovery of hidden knowledge in historic and transaction data about customers, such 'discovered' knowledge provides a solid foundation for the generation and application of new knowledge for making smart decisions and developing programmes to meet present and future expectations of customers. This is what constitutes the DM imperative for organizations.

DM's ability to create a logical link between the hidden knowledge of the past and the expected knowledge of the future may explain why it is profiled as a major undertaking in the millennium ahead. Furthermore, in an increasingly knowledge-intensive economy, DM is regarded as the bridge between technology and knowledge. 'DM focuses on the techniques of non-trivial extraction of implicit, previously unknown and potentially useful information from very large amounts of data'.[11] Therefore, the point is that DM is one of several methods of attaining the ends (generating knowledge) using technology as the means.[12]

As an eminent imperative, many organizations have employed DM in a number of areas including finance (detect fraudulent patterns of clients, identify correlations between financial indicators), marketing (identify buying behaviour of clients, find associations between clients' demographic characteristics, predict clients purchase patterns), transportation (analyse loading patterns, determine distribution schedules among outlets), medicine (characterize patient behaviour and predict office visits, identify successful medical therapies for different illnesses), insurance (predict behaviour patterns of risky clients, analyse claims, predict which customers are likely to buy policies), telecommunication (call tracking, churn management), manufacturing (diagnosis), web analysis (assess user browsing patterns, customer support)[7,13] and education (predict students' transferability).[14]

In summary, the DM imperative will persist because of the need for leaner customer relationship management. DM will continue to evolve because of developments in storage and computing power, improved database technology and maturing DM tools.[14] While so many issues may be identified with DM as an aid for EIS, one that we consider paramount is associated with the models and techniques.

## DM models

A model is a representation of the real world; it forms the cornerstones of DM. Without an accurate model or representation of the real world, well-intended decisions may lead to catastrophic results. From a non-technical perspective, a DM model may be considered as a black box, whose input is data to be mined and whose output is the knowledge discovered from the data. With this perspective, DM employs two types of models – predictive and descriptive, which are like two sides of the same coin. One side, predictive models operate on certain attributes or characteristics of the data (independent variables) to predict other

attributes (dependent variables). Descriptive models on the other side do the reverse – based on specific outcomes (dependent variables), descriptive models attempt to look for the contributing attributes associated with the outcomes (dependent variables). The following vivid examples have been provided to illuminate understanding of predictive and descriptive DM models: [13]

- Classification models predict an outcome given a set of input characteristics. Predicted outcomes are then sorted into classes (e.g. fraudulent or not fraudulent)
- Regression models predict a real number outcome (e.g. a customer's expenditures over the next year, given a set of input characteristics)
- Association models predict the occurrence of a second event, given the occurrence of another. For example, beer purchasers buy peanuts 75% of the time
- Sequencing models predict the sequence of events. Those who rent *Star Wars* then rent *Empire Strikes Back*, then *Return of the Jedi*, in that order
- Clustering models describe a natural group of things – such as vacation destinations by age group and income. Clustering is used extensively for determining market segments.

## DM utilization in the health-care arena

KDD has become popular in health-care organizations all over the world.[15] Many health-care organizations have developed ways of storing valuable information pertaining to patients and their medical conditions to determine potentially useful patterns of information and relationships within the data. Although this practice has its perks, it has also posed some limitations for domain experts. Health-care organizations using 'manual analysis' as their primary way of inputting data have been unsuccessful in keeping up with the pace of storing empirical data as the number of cases (or patients) has rapidly increased in the past decade. Therefore, innovative discovery-based approaches to health-care data analysis warrants further attention as databases, data warehouses and data repositories are becoming ubiquitous.[15] Other challenges and issues facing health-care organizations include: underdeveloped IT infrastructures to support DM processes, its ability to view a vast amount of data and make accurate generalizations, and/or proper execution of quality assurance methods to minimize the risks of making bad decisions associated with the results of a DM analysis.

Nonetheless, various DM methods still have been used as a means of predicting health-care costs based on previous claims data reports.[16] In a case study presented to Princeton University, Bertsimas used classification trees and clustering algorithms to develop key insights on how DM methods provide accurate predictions of medical costs and represent a powerful tool for prediction of future health-care costs. In the case study, he assessed and reviewed claims data for 400,000 members over a period of three years. As a result of his review, Bertsimas discovered that the pattern of past cost data is a strong predictor of future costs and medical information provides an accurate prediction of medical costs particularly on high-risk members.[16] And most importantly from outlook perspective, he was able to make the prediction that new medical knowledge can be obtained through DM methods. So, is the future as daunting as we thought after all?

## DM behind the scene

For the technical-savvy, DM models are not really black boxes. Known rules and different algorithms from statistics, machine learning and visualization are employed in each type of model. This makes it possible to talk about DM models at a much lower and in-depth level

based on what is done during the DM. For example, classification models employ logistic, tree, naïve Bayesian and neural network algorithms. Regression models employ linear, tree and *k*-nearest-neighbour, and clustering models employ *k*-means, hierarchical and principal component algorithms.

The distinguishing characteristic of DM models is the degree of restrictions imposed on the nature and quality of data. Statistical models, for example, impose the greatest restriction. Depending on the types of data-sets processed, 'mining approaches may be classified according to the complexity of the data-set'.[5] The data-set must conform to rigid distribution criteria such as normality. On the other hand, machine-learning-based models impose fewer restrictions and are more widely used than the statistical methods;[17] with rule induction or decision trees, neural networks, case-based reasoning, genetic algorithms and inductive logic programming[17,18] being typical examples. An in-depth discussion of the DM tools and toolboxes associated with the above DM is presented in Mena.[19]

Whatever DM model is used, DM as a process entails a number of interactive and recursive activities, including several steps such as: (1) establishing DM goals; (2) selecting the data-set to be mined; (3) cleaning or preprocessing the data (to remove any noise, erroneous or incomplete data); (4) transforming data from one scale to another or from one format to another (in order to ensure conformity and to improve quality); (5) warehousing the clean data; (6) doing the DM proper (analysing the data and interpreting results); and (7) evaluating and reporting results.[7,20]

Analysing data in the actual DM process depends on the DM tool used. Regardless of the tool, the data-set is often partitioned into two or three subsets, respectively, used for training and testing, or evaluating and testing the DM model. Furthermore, the time expanded on the DM process appears to be broadly standard: steps 6 and 7 above require about 10–20% of time. All the other steps are considered as DM preparation. Based on Cross-Industry Standard Process for DM, the first five steps could be regrouped into three activities: investigating the possibility of overlaying DM algorithms directly on the data warehouse (5–15% of time); selecting a robust query tool to build the required DM files (30–75% of time); and data visualizing and validating, to appreciate what each field in the database contains (10–20% of time).[14]

## The future of DM

The scope of DM has widened in recent times, thanks to technological advances. Better tools are emerging with enticing graphical user interface(s) (GUI) and other capabilities to make DM less of a technical endeavour. The benefit of this is the possible wide use of DM. Some tools present themselves as add-ins to popular packages that are already familiar to the user community. One good example is XLMiner,[a] a DM tool developed by Professor Patel and his colleagues at MIT Sloan School of Management, USA. Once installed, XLMiner becomes an add-in to Microsoft Excel. As an Excel add-in, XLMiner widens the scope and range of data manipulation activities for the user. But while add-ins may be very practical tools to managers, they lack the flexibility and capability of being used as learning tools by students. Students need to learn and gain in-depth knowledge of the DM models and algorithms, so that better tools can be developed. To meet the needs of students, other DM tools are emerging with GUI and command line options. An example of such student-oriented DM tools is Weka,[b] developed by Professor Witten and his colleagues at the University of New Zealand. Weka[c] presents a GUI and command-line environment for

[a]See http://www.resample.com/xlminer/capabilities.shtml (last checked 6 October 2006)
[b]See http://www.cs.waikato.ac.nz/~ml/index.html (last checked 6 October 2006)

students to learn about various DM algorithms, including experiments such as concurrently running multiple DM projects. Appendices 1 and 2, respectively, highlight XLMiner, while Appendix 3 compares and contrasts these two tools.

Beyond current efforts, the next generation of activities is going to focus on video mining and web usage mining (WUM). In video mining, speech recognition is blended with image understanding and natural language processing to obtain even more actionable information for top management. Pioneer work in this area has begun at Carnegie Mellon University. As reported in MIT's Technology review, given a video clip archive, Carnegie's Informedia II system can produce computer-readable index.[8] Clearly, video mining depicts an exciting future of DM boom. How far DM goes will depend on a number of factors, which we now examine. Tao and his colleagues have explored a category of online browsing known as WUM by exploring a new data source called intentional browsing data.[11] This allows businesses to use the web in order to extract knowledge needed to enhance the long-term viability of the organization.

First, Pratte[13] points to technical difficulties mostly associated with optimum use of DM. DM tools have different capabilities and the user must be knowledgeable about these capabilities and differences. For example, the XLMiner package mentioned above cannot handle missing data. Weka on its part requires excessive manual manipulation of data files to prepare the data into the appropriate forms.

Second, the inability to choose the right tools for the appropriate tasks means misuse and possible failure: obtaining no answer, obtaining the wrong answer (even with right tools used) and misinterpreting the answer.[13] The consequences of these difficulties could be very detrimental to organizations to the degree that use of DM becomes limited or even avoided.

Third, another facet of technical difficulties is associated with creating and mining a data warehouse with data from different organizations. The major problem here is how to integrate different data formats and create common measures and languages for different attributes so that the knowledge discovered means the same thing to the different organizations. Lin et al.[21] have attempted to resolve some of the incompatibility difficulties by designing semantic association rules, transforming quantitative data into semantic data and using regular DM algorithms on retail data from different organizations. This task is relatively simple because they are integrating data from the same rather than different areas. While technical constraints may be temporal, others may not, which brings us to the fourth factor.

The fourth factor is social concerns about the privacy of individuals, especially within health-care organizations – as the majority of data assessed through these organizations will contain pertinent health information. Personnel within these organizations will have to input security measures and implement ongoing quality assurance tests to refrain from accidentally leaking medical information and violating governmental compliance standards. Private companies may soon meet unexpected limitations to their attempts to collect and use data about customers. Existing legislation is highly being contested and/or modified. In the Netherlands, for example, recent legislation stipulates that subjects must be informed when data about them are to be used in DM and subjects have the right to be removed from the data at no cost. Furthermore, insurance companies are not allowed to differentiate premiums on the basis of general risk analyses with DM. Unequal treatment of customers is only possible if reasonable arguments are given, rather than using DM results.[20] So far, the USA

[c]Weka can actually be used as a learning module in professional DM application (java implementation)

legislation is much relaxed compared with that of the Netherlands. What will happen to DM if the Netherlands type of legislation becomes universal or becomes effective in the USA?

Already in the USA, many think government is over-stepping the line in using DM and related systems. This is ironical, given increased terrorists' threats. One would cherish and applaud government's current effort and even reprimand government for not being more involved in DM earlier than now. The reason noted was, for example, that if CopLink (a text-mining-like application) had been used by the USA Government early enough, the DC area snipers would have been caught sooner.[22] Under the same expectations, more sophisticated systems ought to be deployed to track terrorists based on historic and current data.

Some current systems being used by the USA government, such as CopLink, are not viewed favourably by the public because of privacy concerns. Carnivore[d] was one such system that has been discontinued. But other systems such as the HITIQ system being developed by Professor Strzalkowski and his colleagues at the University at Albany may be highly supported. The HITIQ system uses publicly available foreign affairs information from various sources, and acts as an intelligent analyst to brief officials of the latest events.

However, if 'appropriate' legislation is put in place, it will become difficult to explore DM to the fullest, when personal data are included in the data warehouse. Consider the Netherlands example again: companies are not allowed to enrich customer data from their transactional systems with demographic data bought from list-brokers without permission of the customers. Recall that for the most part, the essence of DM is to gain better knowledge about the customers who organizations serve.[e] How can organizations best serve if they do not know the customer well?

## Conclusion

We have examined some prospects of DM to support EIS. While technology attempts to bring better tools, it also poses and/or faces problems because of the complexity of the tools and/or the desire to do more by organizations. While technological obstacles are temporal in the sense that they eventually get resolved with time, social issues seem to present the greatest obstacles. Social issues become more complex with time, as people become more knowledgeable about their privacy and confidentiality rights. This complicates matters for DM because recent research points to the necessity of incorporating 'human data' in data warehouses in order to better understand customers.[23]

Human data are considered to be whatever customers say, don't say and what they do as they walk on the shop floors. Combining transactional data with human data in DM provides better actionable information for decision makers. Companies that are attempting to incorporate human data into their warehouses testify to the advantage of having both transactional and human data. In addition to the current DM practice, we estimate that extended video surveillance and subsequent intensive video mining is the best way forward. This is so because it is not possible to tag salespersons to every customer on the shop floors. However, with privacy and confidentiality issues looming high, striking a balance between how and what organizations must collect about people and what the same people can allow organizations to collect about them will be one of the toughest challenges to beat in the practice of DM. This is what makes the future of DM daunting!

[d]Carnivore is a tool that the FBI previously used to obtain lawful information from network traffic while maintaining privacy. The FBI is now using Coplink (See http://www.coplink.com/) with oversight by the US Department of Justice

[e]DM can also be used in other areas not involving data and information about people, for example, in weather forecasting. This paper emphasized the case of using DM on data warehouses with data and information about people

## References


1. Mohan L, Holstein WK, Adams RB. EIS: it can work in the public sector. MIS Q. 1990; 14:435–48.
2. Imhoff C. The corporate information factory. DM Rev. Dec.1999
3. Alcala-Fdez, J.; Sanchez, L.; Garcia, S., et al. KEEL: A Software Tool to Assess Evolutionary Algorithms for Data Mining Problems. Springer-Verlag;
4. Imhoff C, Geiger JG. My, how times change! Ten years of data ware-housing. DM Rev. Feb.2001
5. Hong T, Horng C, Wu C, Wang S. An improved data mining approach using predictive itemsets. Expert Syst Appl. 2009; 36:72–80.
6. Quass A. Data warehousing principles for the strategic management of information: a synthesis of contemporary practices, theories and principles. Southern Afr Bus Rev. Special Issue on Information Technology. 2000; 4:41–9.
7. Hui SC, Jha G. Data mining for customer service support. Inform Manag. 2000; 38:1–13.
8. Waldrop M. Data Mining. Ten emerging technologies that will change the world. MIT Technol Rev. Jan-Feb;2001
9. Kimball, R. Millennium ahead. Data Webhouse; 2000. see www.intelligent-enterprise.informationweek.com/000101/
10. Asoh, D.; Belardo, S.; Neilson, R. Knowledge Management: Issues, Challenges and Opportunities for Governments in the New Economy; Paper Presented at the 35th Hawaii International Conference on System Sciences (HICSS35); Hawaii. 2002.
11. Tao Y, Hong T, Su Y. Web usage mining with intentional browsing data. Expert Syst Appl. 2008; 34:1893–904.
12. Spiegler I. Technology and knowledge: bridging a “Generating” Gap. Inform Manag. 2003; 40:1–7.
13. Pratte D. Can data mining predict the future of your enterprise? TechRepublic. 2001
14. Luan J. Data mining and its applications in higher education. New Directions for Institutional Research. 2002; 113:17–36.
15. Heiat A. Knowledge Discovery and Data Mining in Healthcare: Challenges and Issues. Journal of AHIMA. 2005; 11:35–55.
16. Bertsimas D, Bjarnadottir M, Kane M, Kryder C, Pandey R, Vempala S, Wang G. Algorithmic Prediction of Health Care Costs. Operations Research. 2008; 56:1382–92.
17. Bose I, Mahapatra RK. Business data mining - a machine learning perspective. Inform Manag. 2001; 39:221–5.
18. Witten, IH.; Frank, E. Data Mining: Practical Machine Learning Tools and Techniques With Java Implementations. Morgan Kaufmann Publishers; San Franscisco: 2000.
19. Mena J. Data mining FAQ’s. DM Rev. 1998
20. Feelders A, Daniels H, Holsheimer M. Methodological and practical aspects of data mining. Inform Manag. 2000; 37:271–81.
21. Lin Q-Y, Chen Y-L, Chen Y-C. Mining inter-organizational retailing knowledge for an alliance formed by competitive firms. Inform Manag. 2003; 40:431–42.
22. Mnookin S. A Google for cops. Newsweek. Mar 3.2003
23. Davenport TH, Harris JG, Kohli AK. How do they know their customers so well? MIT Sloan Manag Rev. 2001; 42:63–73.