

# Mining for Information Gold



Data mining offers the RIM professional an opportunity to contribute to knowledge discovery in databases in a substantial way

**Joseph M. Firestone, Ph.D.**

**D**uring the late 1980s, several trends in computing, including the emergence of client-server technology, the growing popularity of structured query language (SQL), the gravity of “the islands of information” problem, and the inaccessibility of much of the structured information “hidden away” in both legacy and SQL transactional databases led to the development of large, physically centralized, structured databases called data warehouses. These were intended for decision support. SQL-querying technology, however, was not sufficient to deliver the hoped-for information value, and the 1990s led to

the rapid growth of data warehousing and to the development and spread of new technologies for getting useful information out of those surprisingly unwieldy first-generation data warehouses.

One of these new technologies is data mining, a term based on the idea that very large databases are “mountains” of information that can be “mined” for “nuggets” of great value if the right technology is applied. During the 1980s and increasingly during the 1990s, data mining technology was becoming available in the form of statistical and artificial intelligence-based models and computing algorithms. Additionally, new software technology

## **At the Core**

### **This article**

- ▶ Defines data mining and its process context
- ▶ Discusses its value for records and information management professionals
- ▶ Shows how to get started in data mining
- ▶ Predicts the future of data mining

was being developed for integrating distributed systems based on object and web technology. The result of this confluence between need and technology has been a continually growing data

mining industry containing scores of new companies selling to large, mid-sized – and even small – organizations. There are several sectors interested in data mining: banking, medicine, insurance, retailing, and government. Data mining supports many goals, such as reducing costs, enhancing or reusing research, increasing sales, and detecting fraud.

The image suggested by the term “data mining” is an attractive one, but, unfortunately, it may not be very informative to those records and information management (RIM) professionals who need to know what data mining means for them. RIM managers need answers to these questions.

- What is the process context of data mining?
- What is its value for RIM managers?
- What is the relationship between data mining and knowledge discovery in databases (KDD)?
- How does one get started in a data mining process?
- In what direction is this fast-moving field going?

The most compelling reason for RIM managers to take an interest in data mining is simply this: the “data” in “data mining” are, for the most part, records created in the normal course of business of any organization. Records, then, become the structured data foundation to the data mining process.

### What Is Data Mining?

Definitions of data mining abound, and they vary among practitioners. (See Sidebar, “Definitions of Data Mining” on page 50.)

Selecting just one of the definitions is not as important as realizing that people will use the term *data mining* in at least the four ways described in the sidebar. It will be up to information managers to decide which meaning their organization assigns to it. Definition 3 is used in this article because it has the advantage of distinguishing “data min-

ing” from traditional analyses by emphasizing its automated character in generating patterns and relationships. It also clearly distinguishes data mining from knowledge discovery by emphasizing the much broader character of KDD as an overarching process, including steps distinct from data mining and relying more heavily on human interaction.

### What Is the Process Context of Data Mining?

The process context leads to the more comprehensive process of KDD within which data mining occurs. KDD starts with problems – seeking them in routine situations, recognizing them, and clearly articulating them. It continues with gathering information about a problem and its potential solutions. At that point, hypotheses or models are developed that are central to the solution. There are many alternative ways of developing models, including intuition, a literature review, mathematical modeling, and facilitation processes, that do not involve data mining, even when statistical and modeling techniques are used as part of the KDD process. But, at this point, one can make the choice to apply automated analysis to an organization’s very large database – that is, *data mining* – as an initial method of arriving at alternative patterns and/or relationships.

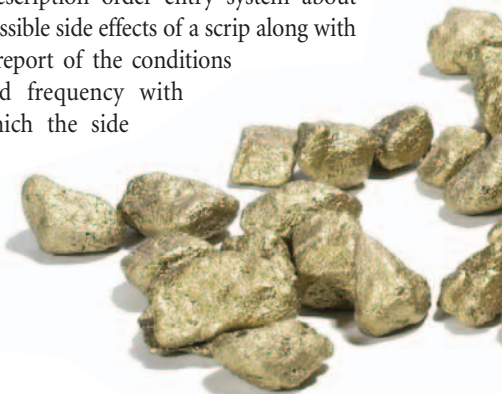
When that decision is made, then the steps of selecting, pre-processing, and transforming data must be completed, as well as the step of selecting data mining tools before data mining itself can be performed. Also, once data mining is completed, KDD is still far from done. The patterns found by data mining must still be interpreted and evaluated, and further statistical analysis and analytical modeling frequently is needed to refine, test, and evaluate the discovered patterns. In short, the process context of data mining is KDD, and KDD, in turn, is a knowledge life cycle originating in a problem, proceeding with attempts to discover patterns through a number of steps, that include – but go beyond –

**... apply automated analysis to an organization’s very large database – that is, data mining – as an initial method of arriving at alternative patterns and/or relationships.**

data mining, and ending with evaluating, interpreting, and selecting patterns that solve the original problem.

### What Is Its Value for RIM Professionals?

What good is data mining to RIM professionals? Of course, it depends on the person and his or her role. Using the results of KDD that rely on data mining can help with very routine decisions almost anywhere in the enterprise. Mike Ferguson gives a good example in his article “Integrating Business Intelligence into the Enterprise: Part II” regarding a bank call-center operator who receives a lending recommendation on his screen and applies a data mining-derived predictive model to a database to develop a risk score and an associated loan recommendation. Another example is the physician who receives an alert from a prescription order entry system about possible side effects of a scrip along with a report of the conditions and frequency with which the side



effects occur.

If there is a problem to solve and performing KDD to produce an explanatory causal or predictive model is being considered, then using a data mining step – in addition to traditional statistical analysis – can be very valuable. In the article

“Putting Data Mining in Its Place,” Dorian Pyle tells the story of a bank that turned to data mining when its huge direct mail marketing effort to increase loan inventory failed miserably. Data mining was able to work through 2.5 million accounts looking for those that were the most profitable. It showed that a tiny segment, one comprising only 0.1 percent of the accounts, comprised 30 percent of all people who bought ski equipment valued at \$3,000 or more in a 30-day period and then later bought travel packages valued at an additional \$3,000 or more. When the bank used this information to implement a marketing package to 8,300 others in its database who had bought \$3,000 in ski equipment in 30 days, an additional 3,300 people responded to its offer, purchased an additional \$3,000 loan, and helped the bank increase its loan inventory by \$10 million. In commenting on this case, Pyle makes the important point that this 0.1 percent segment, discovered by the brute force approach of data mining, would probably have been missed by traditional statistical analysis because its small size makes it statistically insignificant. However, from the causal, predictive, and

commercial points of view, it was highly significant, and its discovery illustrates one of the advantages of data mining over more traditional approaches in the KDD process.

Though data mining can be very useful in arriving at models, an important caution for RIM managers to keep in mind is that data mining is not magic. That is, merely taking a data set, clicking an on-screen icon, and expecting to solve a problem will not work. The steps of KDD surrounding data mining involve continuous interaction of humans and computers. How well those steps are performed depends on the skills and background knowledge of humans.

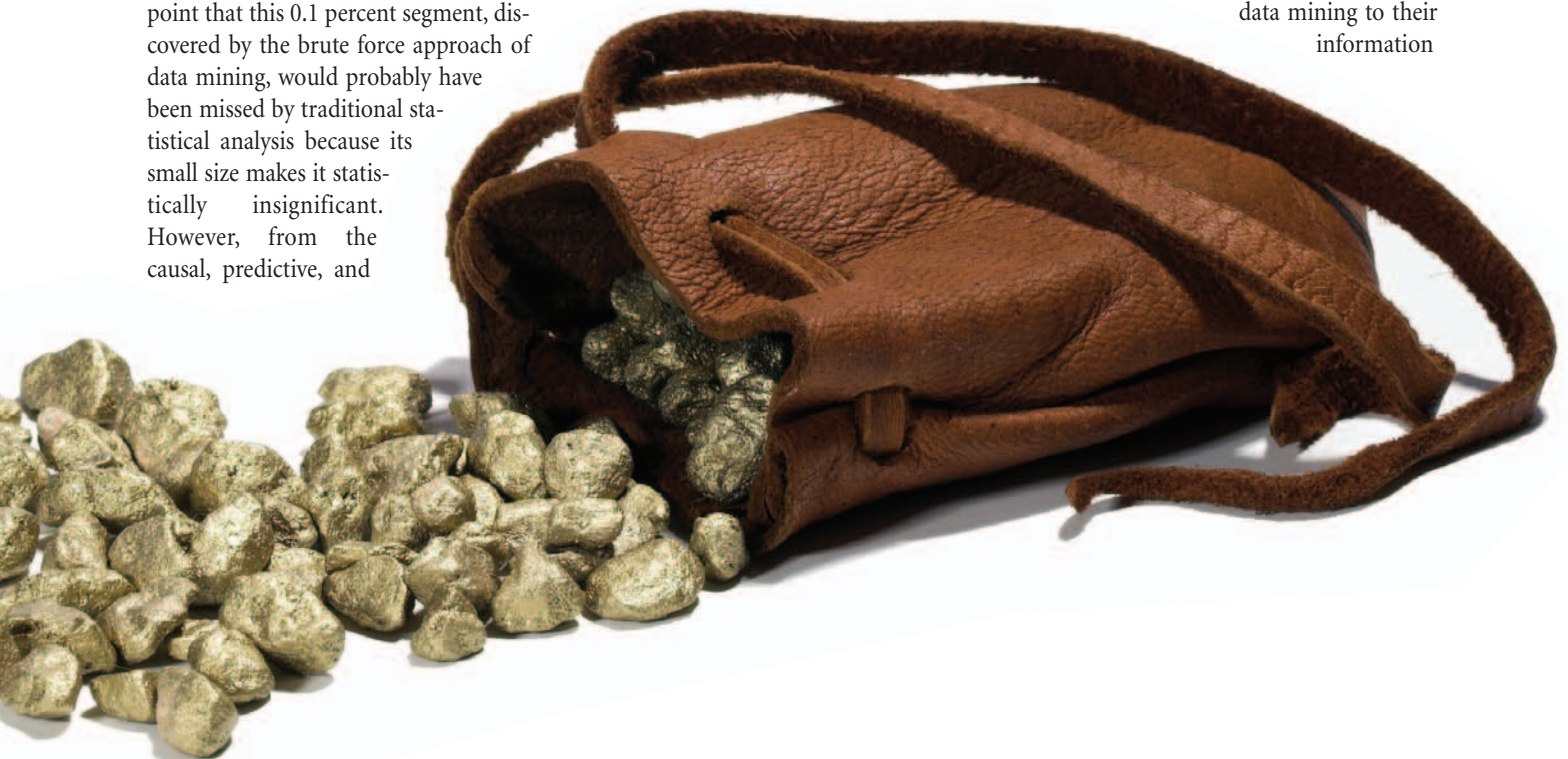
During the early days of data mining, some of its exponents claimed that data mining software would generate good results even though the data miners using it were not highly skilled or trained analysts or statisticians. But this notion has proved to be oversimplified. All the steps in KDD preceding and following data mining require good technical skills and business experience to perform effectively, and, in the end, they – not a computer – determine the success or failure of the automated data

mining step. When it comes to KDD, then, there is no free lunch or “magic.” There are only careful and smart humans working through difficult problems with, admittedly, more leverage than they used to have over their databases.

### Enhancing the Quality of Information

The job of the RIM professional is to enhance the quality of the information in the enterprise by enhancing the quality of the processes used in producing, storing, using, and integrating information within it. Much, not all, of this information is in the form of structured information – i.e., records – and is found in enterprise databases. Data-mining capability enables staff to enhance the quality of this information by facilitating knowledge discovery in these databases. Other capabilities, however, such as SQL and online analytical processing (OLAP), do not discover new patterns as much as they confirm patterns already thought and articulated by the investigator. As noted above, data mining can discover patterns and relationships that conventional statistical approaches can easily miss.

So, RIM managers who add data mining to their information



technology arsenal are enhancing the KDD process by adding an additional and powerful capability to create information that through interpretation and evaluation can lead to new knowledge – information of even greater quality and value.

During business or intelligence crises, the contribution of data mining to KDD can be very valuable for RIM managers. Pyle's example of the failure of the bank's marketing campaign may be resolved successfully with the aid of a RIM manager, who brings new value-added strategies. On the other hand, such situations must be approached carefully since data mining results can be misinterpreted by unskilled analysts or information workers who uncritically use data mining software without an awareness of the business problems being investigated – or of the characteristics and pitfalls of the particular data mining technologies they were using. There are many data mining software applications available, but all do not work equally well on the same type of analysis.

Part of the job of RIM managers is to specify and implement changes in formal rules and procedures for handling information in IT environments. Data mining can help RIM managers model the likely impact of these changes and measure their impact after implementation.

### Getting Started

When getting started in a new field, familiarity with some of its leading works is helpful. Pyle's books, such as *Business Modeling and Data Mining* and *Data Preparation for Data Mining*, along with Michael Berry and Gordon Linoff's *Data Mining Techniques* and *Mastering Data Mining* are widely read. A good comprehensive treatment is given by Jiawei Han and Micheline Kamber in *Data Mining: Concepts and Techniques*.

The state of software in data mining is more complex. During the early days in the field, there were only small companies in the market, as large vendors would not take the risk of moving

## Definitions of Data Mining

Here are four representative definitions reflecting somewhat different ideas about its nature.

1. Data mining is traditional data analysis methodology updated with the most current advanced analysis techniques applied to discovering previously unknown patterns in large data bases.
2. Data mining is the activity of automatically extracting hidden information (patterns and relationships) from large databases without benefit of human intervention or initiative in the knowledge discovery process.
3. Data mining is the step in the process of knowledge discovery in large databases that (a) inputs predominantly cleaned, transformed data, (b) searches the data using purposeful algorithms, and (c) outputs patterns and relationships to the interpretation/evaluation step of the KDD process, which is "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data," according to the authors of *Advances in Knowledge Discovery and Data Mining*.
4. Data mining is "the application of database technology and techniques (such as statistical analysis and modeling) to uncover hidden patterns and subtle relationships in data and to infer rules that allow for the prediction of future results," says a U.S. Government Accounting Office [GAO] report entitled "Data Mining: Federal Efforts Cover a Wide Range of Uses."

into an unproven field. Today, however, the primary vendors of statistical packages now all have comprehensive offerings in data mining, and one of them could be the foundation of a new data mining initiative, especially if there is interest in integrating data mining infrastructure with the rest of the organization's IT architecture. A big issue is whether the data being mined will remain in one's own data warehouses or be migrated to a specialized data mining application server or client machine. The primary vendors can often offer better facilities for integrating either home-grown or third-party algorithms into their platforms. Having said that, data mining is one field where it is important to keep an eye out for novel

techniques or off-the-shelf software providing convenient integration of older techniques; these are often introduced by small companies.

A new trend in data mining is its incorporation in offerings of business intelligence (BI), database, and ERP vendors.

### Concerns About Data Mining Use

While the technological uses of data mining are attractive, there are several issues that must be addressed. Ensuring the quality of the data mined – completeness, currency, and accuracy, for example – require considerable staff time. Privacy is a concern within the commercial and governmental sectors. For example, will data used for one pur-

## A Word on Text Mining

**W**hile beyond the scope of this article, text mining may be as important as data mining. But it is different in many important ways because it works with unstructured information and often involves models and technologies that are very different from those involved in data mining using structured data.

pose be reused for purposes to which the donor never agreed? Interoperability among platforms and databases used by different organizations and agencies remains a concern, said Jeffrey W. Seifert in his CRS Report for Congress, "Data Mining: An Overview."

### The Direction of Data Mining

Data mining is still a rapidly changing field. The basic structure of the vendor marketplace is unlikely to change over the next few years. The four primary vendors, however, are likely to remain very important. Small companies will continue to drive algorithmic innovation in the field, and some will even drive platform innovation. Within this framework there will, however, be four significant changes.

First, the veins of innovation currently being developed at university laboratories will continue and will find their way first to smaller vendors and eventually to the major vendors. One reason innovation will continue is that government funding for both data mining and text mining technology will continue to support it in our post-9/11 world.

Second, the open-analytical workbench platforms of vendors will become increasingly popular, in part because

other vendors will want to use them to customize data mining capabilities from third parties but also because open architectures will facilitate the transfer of new technology from university laboratories and from companies developing new technologies on government contracts and grants.

Third, integration of BI and data mining technologies will occur. This trend is being driven by business performance management (BPM) and will continue. BPM is oriented toward metrics development and data gathering for performance measurement. The

boost to data mining from this trend is obvious.

Finally, a new wave in data mining is related to the continued development of intelligent agent and distributed knowledge processing technology and the growing glut of information on the web. This wave, of course, is that of distributed multi-agent-based data mining. When the wave will hit in earnest is uncertain, but with research money again available for solving the problems of distributed agent-based computing, new companies are likely to begin marketing such systems before long. ■

*Joseph M. Firestone, Ph.D., is CEO of the Knowledge Management Consortium International and is the author of several books on knowledge management. He may be reached at eisai@comcast.net.*

## References

- Berry, Michael J. A. and Gordon Linoff. *Data Mining Techniques*, 2nd ed. New York: John Wiley & Sons, 2004.
- Berry, Michael J. A. and Gordon Linoff. *Mastering Data Mining*. New York: John Wiley & Sons, 1999.
- Dalkilic, Mehment. "Introduction and Motivation to Data Mining." Available at [www.cs.indiana.edu/~dalkilic](http://www.cs.indiana.edu/~dalkilic) (accessed on 22 July 2005).
- Fayyad, Usama M., Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy (eds.). *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA: MIT Press, 1996.
- Ferguson, Mike. "Integrating Business Intelligence into the Enterprise: Part II." *Business Integration Journal* 7, no. 3, April 2005, p. 28 - 31.
- General Accounting Office. *Data Mining: Federal Efforts Cover a Wide Range of Uses*. Washington, DC: U.S. Government Printing Office, 2004.
- Han, Jiawei and Micheline Kamber. *Data Mining: Concepts and Techniques*. San Francisco, CA: Morgan Kaufmann, 2000.
- Pyle, Dorian. "Putting Data Mining in Its Place." Available at [www.modelandmine.com/dm\\_ie.htm](http://www.modelandmine.com/dm_ie.htm) (accessed on 22 April 2005).
- \_\_\_\_\_. *Business Modeling and Data Mining*. San Francisco, CA: Morgan Kaufmann, 2003.
- \_\_\_\_\_. *Data Preparation for Data Mining*. San Francisco, CA: Morgan Kaufmann, 1999.
- Seifert, Jeffrey W. *Data Mining: An Overview*. CRS Report for Congress, Congressional Research Service of the Library of Congress, 2004.