# Data Mining: A Whole New World

Save to myBoK

*by Jane Blumenthal*

We are drowning in data and starving for knowledge," said J. Michael Hardin, PhD, during his "Data Mining" presentation during AHIMA's National Convention. Hardin, of the Departments of Health Services Administration, Biostatistics, and Computer Science at the University of Alabama at Birmingham, faces the challenge of extracting meaningful information from vast databases every day.

In an ever-shrinking world, we are continually gaining access to greater amounts of information. The fact that we don't always know what to do with it makes the data mining process necessary—a way to cope with the magnitude of the data and apply it in a useful way to a variety of industries, including healthcare.

Hardin summed up data mining in one simple phrase: "Using data in large databases to try and help us make better decisions." However, this task isn't easy. With large amounts of data, getting the computer to automatically "push" the information to the programmer's fingertips is difficult, because it would require the computer to think—a task that machines cannot perform. Though data miners are working on a "learning process" for computers, they are currently using more of a "pull" technique to obtain data. This process demands human expertise on how to best extract the information in a usable form.

That's where knowledge discovery in databases (KDD) comes into play. Defined by Hardin as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in the data," KDD is essentially the process of teaching a computer to "learn" how to incorporate its data into constructive information. Since the intelligence presented from data mining requires human analysis, accomplishing KDD would greatly change the face of data mining. Not only would it offer valid information at a glance, it would reduce the amount of time these experts spend creating a cognitive pattern of information.

Currently, the KDD process consists of six steps, all requiring human input:

- Creating a target set of data: The user must develop an understanding of the application domain, the goals of the end user, and relevant variables and data
- Data cleaning and pre-processing: The user must remove noise, as well as define strategies to handle missing data
- Data reduction and projection: The user must find useful features to represent the data and transformation methods to reduce the number of variables
- Data mining: The user must decide which method is best suited to the project and search for patterns of interest
- Data evaluation and interpretation: Domain experts must interact with a large output set
- Refinement of earlier steps based on evaluation and interpretation

Data mining's popularity has increased, Hardin said, especially in the past three years. Two international conferences on the KDD process have taken place, and a new KDD journal recently has been created in response to industrial and academic interest. This interest has grown due to the enormous amounts of data being collected in various industries—from mass-retail chain stores to the National Aeronautics and Space Administration. He even cited one company that earned a 900 percent return on investment by basing its marketing plan on information gleaned from data mining.

In healthcare, data mining has many uses. For example, the human genome project has looked to data mining to help sort out the tremendous amounts of sequencing data that are being discovered every day. In the area of disease control, data mining narrowed an infection down to a certain region of a city by coordinating the data of a group of patients with a particular infection. By defining a geographic location or another specific variable of the disease carriers, an infection may be traced to its source.

While it seems that data mining is well on its evolutionary way, challenges still remain. Few systems exist that fully illustrate the KDD process, Hardin said, and there are few large-scale implementations of data mining. Also, there remains a lingering

need for an evaluation of the impact data mining has on the end-user community. This evaluation of how often the process will be used, as well as its true effect on the healthcare industry, will only come with time and widespread use of the process.

**Article Citation:**
Blumenthal, Jane. "Data Mining— A Whole New World." *Journal of AHIMA* 69, no. 1 (1998): 67-68.