ITBUSINESSEDGE | An IT Business Edge Site

Datamation®

# Drilling Down With A Data Mining Pioneer

**Posted November 6, 2002**
**By Nathan Segal**

Dr. Usama Fayyad, CEO and co-founder of DigiMine, has worked on data mining technology since 1989. Here he shares his thoughts about how to get the most out of your data. Sidebar:

Dr. Usama Fayyad is a data mining pioneer who began working in the field in 1989. He got his start at NASA's Jet Propulsion Laboratory, compiling data on astronomical phenomena such as volcanoes, star systems, etc. From there, he went on to work for Microsoft research and then, frustrated by problems he was seeing in the data mining industry, he left Microsoft and started digiMine to deal with the issues of data mining and data warehousing. In this article, he shares his thoughts about the industry and how to get the most out of your data.

"There are two sides to data mining, descriptive and predictive," says Dr. Fayyad. "Descriptive data mining reorganizes the data, digging deeper into it and pulling out patterns, such as customer similarity, which allows you to create a short description about that group of customers.

"Predictive data mining looks for the best prediction, such as the best product to pitch to a customer. You won't get much insight, but it increases the performance, the ROI. Using both techniques will give you the best results.

"An important issue today is SQL, the standard interface for databases, which has proven to be the wrong interface," Fayyad says. "As an example, let's say you worked for a telecommunications company, and you want to find records about cell phone fraud. Well, guess what? These naturally asked questions cannot be answered by today's databases, because the interface was designed to address problems where you know the target and you want the database to quickly retrieve the result. If you don't have an exact description of the target, you're lost with a database today. This is why data mining is seeing a lot of demand.

"When I started in this field back in 1989, there were many people in large corporations struggling with large data sets. And even though there's a lot of data out there, it's not necessarily the right kind. Also, there's big difference in the ability to store data and the ability to access it in a useful way.

"In response, companies began building data warehouses, which convert transactional database data into a format that allows for more analytically oriented queries to go against it. In theory, it contained all the details for data mining. But in reality, it was a huge challenge. Today, industry analysts are recording an 85% failure rate on data warehouses.

**Data Warehousing Woes**

"The big question today is: Where's the data? If there was a data warehouse, most likely it failed or it's not working. I saw this so consistently that I started digging into it and discovered three problems:

- Invariably, data warehouses were built by a company such a IBM or NCR, but once that company left, the data warehouse started dying;
- In some cases companies tried to build it themselves, but it became so complicated that they couldn't maintain it anymore;
- The data in the warehouse became stale or dead, after the employee who built it left the company."

At this point, Fayyad left Microsoft to create digiMine. He says he realized that "you cannot mine if you can't have access to the data. And you can't have the right data in the right format unless you ensure that there's a successful data warehouse. At digiMine, we build and host data warehouses for companies; then we embed data mining on top as a solution."

Here are some guidelines for collecting data:

- The data needs to be in the right format, which not only collects the data but also collects it with enough detail.
- The data should not be aggregated in any way, because if the detail is lost, you cannot do data mining on top of it.
- There is no specific format for collecting data, as each data mining tool has its own format. At a base level, if the data involves customers, you record all the data and bring it into one data record, but that's not how databases represent data today.

Fayyad says digiMine begins from the other end, asking the client what data needs to be mined and how to apply the algorithms. From there, digiMine sets up the data warehouse and the technology to grab the data from a variety of formats. The customer installs their software, which they maintain and run from their data center.

Fees charged depend on the subscription, ranging from $7,000-$10,000/month for a customer who wants pure data mining and not much warehousing, to $30,000-$40,000/month if the customer wants data warehouse hosting, maintenance and enterprise solutions.

Then there is the issue of purchasing software. According to Fayyad, "There are many tools available from companies such as SAS or IBM, but in order to use them properly, you had better be an expert, preferably a Ph.D. in the area of data mining or statistics. If you're not, you just bought a bunch of shelfware.

"For most users, data mining tools offer the wrong interface. You need data mining solutions. If you have a large staff of experts who know data mining very well, data mining tools will do the job," he says. "However, this department of experts is now acting as the interface between the tools and the ultimate user."

If you're considering purchasing data mining software, Fayyad recommends you look at applications that package the data mining inside the software, or that you purchase a service option.