Temple Times
ONLINE EDITION

NOVEMBER 13, 2003 VOLUME 34 NUMBER 11

For data miners, knowledge is treasure, and half the excitement is digging for it

Here at Temple, we're used to mining for diamonds in our backyard. But at the Center for Information Science and Technology (IST), director Zoran Obradovic and his team are mining for knowledge.

IT'S IN THERE—Center for Information Science and Technology director Zoran Obradovic heads a team of researchers whose expertise lies in data mining—extracting useful knowledge from reams of data.

"As it gets cheaper and cheaper to collect data, and with all of the new technologies available for us to collect data, at some point people realize that they are only getting data, when what they really want is knowledge," said Obradovic, computer and information sciences professor. "In data mining, we are trying to power this process of going from large oceans of data in an efficient way to extract useful knowledge from it."

Significant findings, according to Obradovic, are those that are not previously known, are non-trivial and can be potentially useful. Using software that it develops for Windows- and Linux-based systems, the IST team helps researchers in fields like bioinformatics, astronomy, geoinformation sciences and brain imaging to find the proverbial needles in data haystacks.

"They come to us because they believe there might be something else in this data that they have not been able to extract through usual statistical software," Obradovic said. "So there's many practical applications to what we do."

Traditional hypothesis-and-test techniques, according to Obradovic, have become increasingly infeasible as the datasets continue to grow.

For instance, in recent years astronomers, geoscientists, biochemists, high energy physicists and other scientists collect huge and high-dimension datasets, using advanced telescope technologies and multi-spectral remote sensors on satellites, integrating global positioning systems with high resolution sensors on the ground, and developing instruments like microarrays that generate gene expressions for entire organisms at once.

"Scientists can't handle that much data, so we try to be on the front end to reduce the data into something that can be analyzed in a more efficient way without losing important things," Obradovic said.

One of the studies ongoing at the center focuses on disordered proteins, which are proteins that bind in unusual ways, Obradovic said. The study was started at the National Science Foundation in 1998, and additional funding was received from the National Institutes of Health. On October 1, NIH renewed the study's funding for

another five years to the tune of $1.2 million.

“Intrinsic disordered proteins weren’t studied much in the 20th century,” he continued. “But five years back we got funding to study this process in more detail.”

The initial grant came shortly after the 1997 Nobel Prize was awarded to Stanley Prusiner for his research on prions, disordered proteins responsible for diseases such as mad-cow disease. Disordered proteins are counterexample to traditional “lock-and-key” models of protein combining that have been the favored paradigm since the late 19th century.

“There was a big surprise when, through data mining studies, we found that disordered proteins are very common in nature,” Obradovic said. “It turns out that many of these proteins have associations with cancer and other diseases.”

And, though Obradovic and his colleagues have published close to 30 papers on the subject, he feels that “we have only just begun to scratch the surface.”

Another bioinformatics study Obradovic and his team are working on involves gene microarray analysis. In microarray studies, researchers arrange and measure gene expression levels by the thousands.

“What used to require thousands of labs can now potentially be done in one lab,” he said. “Scientists who used to analyze the relationships between a few genes are suddenly in a position to get expression levels from thousands of genes.”

Vast increases in data require major changes in how the information is analyzed, so data miners facilitate the transition for many scientists.

The best part of data mining for Obradovic is its interdisciplinary nature.

“I was always an undecided major,” he said. “And that’s the beauty of this kind of work. We can choose to help where it is needed, and work on real-life problems.

“At the IST Center, we are doing data analysis of an interdisciplinary type, and we’re still learning who does what at Temple,” he added. “So, if there are people who would like to collaborate and have data they feel needs to be analyzed, we are here.”

For more information about IST research grants and data mining, visit the center’s Web site at http://ist.temple.edu/. — Helen H. Thompson