SFGate

# "Data mining' can locate treasure in previously obscure information

Keay Davidson, EXAMINER SCIENCE WRITER
Published Tuesday, February 18, 1997

SEATTLE - Scientists are developing ways to control and exploit an ever-swelling flood - a flood of information that threatens to overwhelm the world's research institutes and libraries.

"Data mining," they call it: the use of clever computer software to organize the maelstrom of data gushing from laboratories into seemingly bottomless databases.

Until now, much of that data has been recorded and forgotten. There simply aren't enough scientists to analyze it all.

But now, thanks to data-mining software, scientists are culling unexpected discoveries from old databases. They're uncovering fascinating facts about the Earth and sky by "mining" previously unfathomable oceans of raw scientific data, entombed on miles of magnetic tape gathering dust in archives.

As the steam engine was to the 19th century, the database is to the late 20th. Seemingly everything is recorded on computer databases, from your credit rating to the size of volcanoes on Venus.

"Databases are everywhere - and they're spreading amazingly fast," Usama Fayyad of Microsoft Corp. said Monday at the American Association for the Advancement of Science meeting.

Some databases contain as many as 1 trillion bytes of data, divided into as many as 10,000 fields. A simple database, such as one that lists a private book collection, might be divided into, say, two fields, fiction and nonfiction.

## Hard to find

At the conference, Bruce Schatz, of the University of Illinois at Urbana-Champaign's National Center for Supercomputing Applications, warned that we're heading toward a "world of billion (data) repositories. . . . Finding things is going to be a lot harder."

So scientists are developing a possible solution: data mining. Data-mining software not only analyzes databases much more precisely than previous software, it also compares data from databases that use different types of fields and terminology.

One result: Scientists have found a way to measure subtle movements of Earth's crust by analyzing space satellite photos composed of tens of thousands of pixels.

At the meeting, the "Quakefinder" software was described by Paul Stolorz of the Jet Propulsion Laboratory in Pasadena, who developed it with his colleagues Christopher Dean, Robert Crippen and Ron Blom. He showed how the Quakefinder had analyzed photos of Southern California taken from space before and after the 7.2-magnitude Landers earthquake in June 1992. Quakefinder detected extremely subtle shifts in the ground surface caused by the quake.

Other scientists have speculated that by studying long-term crust shifts, they might estimate which regions are most likely to experience quakes. Stolorz declined to speculate on this possibility, noting that other researchers' past efforts to forecast quakes had shown that it was hard to do.

## Studying a Jupiter moon

A software tool similar to Quakefinder is being used to compare photos of the surface of Europa, a moon of Jupiter, that were snapped two decades apart by the Voyager and Galileo probes. In this way, JPL scientists hope to measure how much the Europan surface ice has shifted over 20 years.

Why bother? Because ice movements on Europa may help scientists decide whether it conceals a liquid ocean - one that could be inhabited by simple life forms, such as microbes.

Similar data-mining tools are used to scan space photos for other hard-to-see objects, such as tiny, rocky satellites that orbit the much larger rocks called asteroids.

At JPL, where he used to work, Fayyad and other scientists developed a computer tool

called SKICAT, for Sky Image Cataloging and Analysis Tool. It analyzes thousands of images of the sky recorded by Caltech's Palomar Observatory in San Diego County. The images are so sensitive that they reveal extremely faint points of light near the edge of the known universe.

**Short cut**

Which of these objects are stars and which are galaxies, that is, giant clusters of stars? The images show literally billions of objects, so Caltech would normally have to hire "a few thousand graduate students" for the mind-numbingly dull work of studying each point of light and deciding whether it's a star or a galaxy or something else, such as an Earth-orbiting satellite, Fayyad said.

SKICAT can do the same task automatically, and far more cheaply.

In the 1950s, astronomer George Abell of Southern California conducted a less ambitious "sky survey." He and assistants spent a few years tediously examining Palomar photographs of the night sky until they had identified and cataloged every star, galaxy and other celestial object.

Nowadays, Fayyad said, SKICAT could have completed Abell's catalog in just two hours.

SKICAT also can catalog far dimmer objects - is it a star or a galaxy? - than human eyes can identify.

JPL scientists have also developed JARTOOL, or the "JPL Adaptive Recognition Tool," which scans close-up images of the planet Venus. Its task: to identify all the volcanoes in the Venusian surface. That's no easy task, as Venus is covered by thousands of volcanoes, some extremely small. Yet in one test, JARTOOL was able to identify volcanoes about as accurately as human experts.

**Emerging field**

Data mining is still a new area of research. The first international conference on the subject was held in Montreal less than two years ago. The third such conference is scheduled for Anaheim in August.

And next month, the first issue of a new journal will appear: "Data Mining and Knowledge Discovery."

Data mining also could benefit private enterprise. For example, Fayyad said, data mining might help businesses answer questions such as, "Which records represent fraudulent transactions?" or "Which households are likely to prefer a Ford over a Toyota?" or "Who's a good credit risk in my customer database?" <