is that not a day passes without some government agency, academic consortium or private company developing a clever web interface or piece of data analysis software that makes sifting through the petabytes that much easier.

Any number of indexes and 'master directories' of environmental and global-change data have opened online in the past two years — perhaps too many, says Thomas Karl, director of the US National Climatic Data Center. "I know this is heresy for the head of a data centre," he says. But advertising a one-stop shop for climate data is "some data manager's pipedream". Savvy scientists will never buy the idea of a single place that has it all.

Most researchers are accustomed to studying a relatively small data set for a long time, using statistical models to tease out patterns. "At some fundamental level that paradigm has broken down," says Grossman. "You can't be afraid of data today."

Soon the question for scientists will be "how do you manage a terabyte in front of you?" — an even more difficult challenge considering that a computer may need to generate 1,000 times that amount in the course of manipulating the data.

Mining for data

Various schemes have been proposed to extract knowledge from such large stores of information, including supercomputer-powered scientific 'visualization' and what's been called 'data mining,' or the semi-automated discovery of patterns, associations and statistically significant structures. Data mining borrows tricks from other fields such as artificial intelligence and neural networks to produce software that can plough through large datasets, looking for nuggets that humans would take forever to find.

Astronomers at the California Institute of Technology used a program called Skicat to automatically sort through three terabytes of image data from the Palomar Observatory Sky Survey. Using decision trees and classification rules, the system was able quickly and accurately to classify very faint objects, effectively tripling the number of objects in the catalogue and accelerating the pace at which high-redshift quasars were discovered. The program was developed by a Jet Propulsion Laboratory software engineer (who later moved to Microsoft).

Scientists at the EBI have mined large stores of yeast gene expression data. Astronomers at Australia's Mount Stromlo and Siding Springs Observatories used mining programs to hunt through data on 20 million stars taken nightly for four years, making more efficient the search for Compact Halo Objects (MACHOS).

Data mining is sometimes overhyped, admits Grossman. Some call it "data analysis with better marketing," while others worry that machines searching blindly for subtle associations in large datasets could do "very stupid things" such as correlate heart attacks with the stars. Human judgement will always be necessary, he says. Data mining is just a tool, albeit a powerful one.

Richard Gibbs, a gene sequencer at Baylor College of Medicine in Texas, believes it is imperative in the era of high-throughput genomics for biologists not to turn over all the action to computer scientists, who might miss biologically important information.

Francis Collins, head of the US National Human Genome Research Institute (NHGRI), agrees. "When I give talks to young scientists seeking advice about areas of future intense scientific excitement, computational biology is my number one recommendation," he says. Money alone will not effect change, although training grants in bioinformatics have been stepped up by federal agencies such as the National Institutes of Health (NIH) and the Department of Energy, and private groups such as the Pharmaceutical Research and Manufacturers of America Foundation.

Lisa Brooks, programme director for genome informatics at the NHGRI, says there are two pressing needs when it comes to training. One is for biologists to learn to use publicly available genome analysis software (developed at NIH and elsewhere) through courses, online tutorials, and mentoring with other scientists.

That, she says, is the relatively easy part. Much harder will be building up a corps of scientists who can come up with new statistical approaches to analysing large volumes of genomic data. Today, according to Brooks, "only a small proportion of biologists are capable of developing the tools".

Bioinformatics is still fairly new, and many training programmes are taking only a few students, says Collins. The discipline has had trouble establishing itself.

"Computational biologists in academia often find themselves without a clear career track or an academic home," he says. "Their efforts are seen as too applied to earn respect in departments of computer science, and too ethereal to be accepted in biochemistry or physiology. We need a new mindset."

Gibbs agrees, and says that scientists have no choice but to adapt to a data-rich world. "We've all been doing it slowly, but we've all got to do it faster. The accelerated pace of this is just leaving everybody breathless." □

Reaching for the digital sky

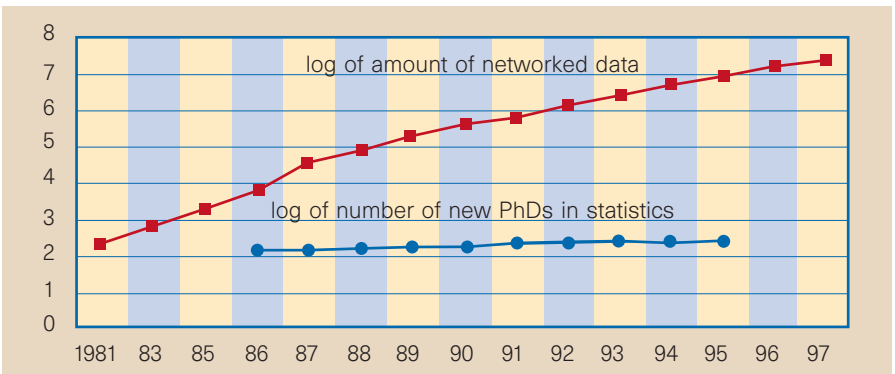
'All-sky' surveys are a popular undertaking in astronomy these days. The 2MASS project is using telescopes in Arizona and Chile to photograph northern and southern skies in the 2-micrometre infrared range. The Digital Palomar Observatory Sky Survey is expected to catalogue some two billion sources in the north, and the Very Large Array in New Mexico has two sky surveys in progress.

These and other searches will generate tens of terabytes of high-quality astronomical images and data. The goal of the Digital Sky project is to show how all this information might be merged into a "multi-wavelength digital library covering a significant fraction of the real sky".

Astrophysicist Tom Prince of the California Institute of Technology is the principal investigator, with funds coming from the US National Science Foundation's National Partnerships for Advanced Computational Infrastructure programme. Digital Sky will focus on the "switchyard" problem, says Prince — how to develop standards and techniques for merging data with different formats into a virtual, distributed database. The Sloan Digital Sky Survey, the most ambitious of the planned surveys, will not be incorporated in the Digital Sky database, but will be among the groups cooperating in setting the standards.

The payoff for future operational databases could be enormous. Scientists might, for example, be able to search easily a billion astronomical objects across a wide range of wavelengths, looking for unusual patterns of energy output.

Digital Sky is at www.cacr.caltech.edu/SDA/digital_sky.html



SOURCE: NCIDM

Skills gap: networked data rises, while the number of those trained to handle it remains constant.

© 1999 Macmillan Magazines Ltd