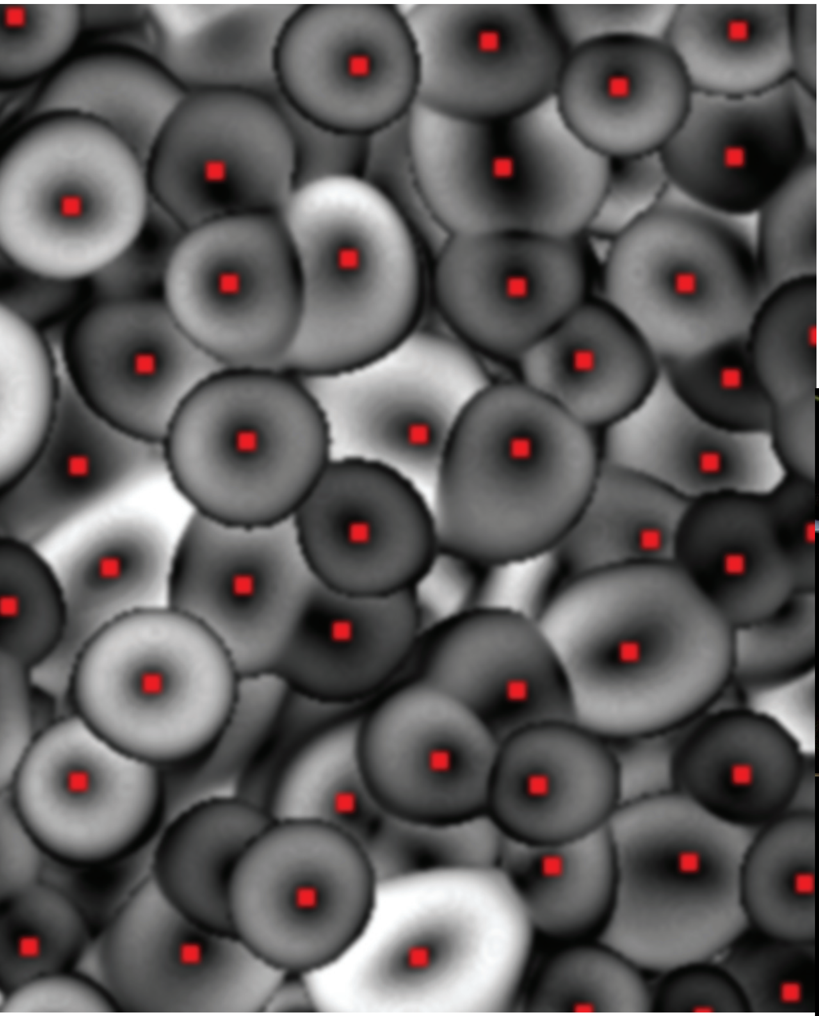
# A GEM OF A SOFTWARE TOOL

WITH the explosive growth of more powerful computers that process large amounts of data, researchers often find themselves confounded by data sets with trillions, even quadrillions of bytes of data. These sets are so large and complex that useful information is easily overlooked, and the potential benefits of increased data-gathering capabilities are only partially realized.

To address the problem of data overload, a team of scientists from Livermore's Computation Directorate developed a scientific data-mining tool called Sapphire. The team, led by computer scientist Chandrika Kamath, and funded in part by two Laboratory Directed Research and Development projects, received a 2006 R&D 100 Award for this novel software. Sapphire provides an end-to-end approach that finds useful information in enormous data sets and offers a breadth of functionality. It is being used for a range of research applications, including astronomy, experimental physics, remote sensing, and computer simulations. According to Kamath, no other data-mining tool meets the analysis needs for such a diverse set of disciplines.



Sapphire software was used to characterize and track bubbles and spikes in an 80-terabyte data set from a three-dimensional, high-fidelity simulation of Rayleigh–Taylor instability. This image shows the bubble counts (in red) highlighted on the magnitude of the *x*–*y* velocity at the bubble boundary.

## Sapphire Tackles Data, Start to Finish

Sapphire starts by preprocessing raw data—images, videos, points in space, or mesh data—and finds the user-defined objects of interest. It then extracts characteristics that represent these objects and uses them to identify patterns. Scientists look at the results to validate that the recognized patterns are the ones they need. “Most data-mining tools do not address the tedious and time-consuming task of preprocessing the raw data,” says Kamath. “They only perform pattern recognition. Sapphire can handle the entire spectrum of a problem.”

Sapphire's architecture also distinguishes it from other data-mining tools. Each task in the pipeline is considered a separate module and is packaged as one or more libraries. The Sapphire

Sapphire development team (from left to right): Cyrus Harrison, Chandrika Kamath, and Abel Gezahegne. Not pictured: Erick Cantú-Paz, Samson Cheung, and Nu Ai Tang.

software, written in C++, addresses computationally intensive tasks such as identifying objects and extracting features. Public domain software is used extensively in the modules that deal with more routine tasks such as reading, writing, and displaying data. Intermediate data extracted from raw images or meshes are stored in a database or other storage scheme that is either commercially available or in the public domain. Sapphire's modules can be connected directly with software code or through a scripting language such as Python. This packaging allows users to extract the information they need and combine tasks in the order that is appropriate for a particular application. Other modules can be easily added to enhance the software's functionality.

"From the start, we designed Sapphire to meet the diverse needs of scientific applications," says Kamath. "We used various technologies from the fields of data mining, machine learning, image and video processing, statistics, and pattern recognition to provide this end-to-end approach. Designing a flexible system was a key focus of our efforts."

In addition to its modular, extensible architecture, Sapphire can support several algorithms—the detailed sequence of operations for a given task—so users can experiment to find the best algorithms for each problem. The data-mining package has user-friendly interfaces, allowing users to fine-tune each algorithm to specific data sets. In addition, the software can process incrementally growing data sets. That way, users can explore a sample of data to determine the best approach for a problem. Once they select a method, they can apply Sapphire to the entire data set.

## Bringing Data Overload under Control

Researchers have used Sapphire to analyze scientific data sets ranging in size from a few megabytes to tens of terabytes. "For example," says Kamath, "Sapphire was used to identify galaxies with a 'bent-double' shape. The researchers were originally faced with examining a million galaxy images by hand to visually identify these particular galaxies—a task that would be tedious, subjective, and error-prone." (See *S&TR*, September 2000, pp. 20–22.)

Another study used Sapphire to characterize bubbles and spikes in a high-fidelity simulation of Rayleigh–Taylor instability in a fluid mix problem. The massive amount of data (80 terabytes) was a major challenge, as was the difficulty of defining the bubbles and spikes that form as the fluids mixed. A quantitative task, such as counting the thousands of bubbles and spikes over time, did not lend itself well to manual or visual analyses. "The Sapphire team provided advanced bubble-tracking algorithms and a framework to analyze our simulations," says Livermore physicist Paul Miller. "The results were striking—revealing distinct regimes of fluid behavior. This perspective is helping us understand the development process of a complicated physical system."

To date, Sapphire has mainly been used in scientific research, but it could also be applied to commercial problems, for example, to detect credit card fraud or analyze documents and video images. "Data overload is also of great concern to national security organizations," says Kamath. "The common issue in all of these disciplines is how to make sense of large amounts of data—effectively and efficiently—without overlooking something of importance. We see Sapphire as unique among data-mining software packages in its ability to provide a complete solution to the analysis needs of a diverse set of applications."

*—Ann Parker*

**Key Words:** data-mining software, R&D 100 Award, Sapphire.

***For further information contact Chandrika Kamath (925) 423-3768 (kamath2@llnl.gov).***