Dealing with Data Overload

Livermore researchers develop methods for keeping up with the tsunami of data.

WE may think we have problems managing our ever-increasing stream of electronic personal "data," whether that information comes in the form of e-mails, social network updates, or phone texts. However, the challenges we face are miniscule compared with those faced by scientists who must parse the growing flood of scientific data critical to their work. From sequences to simulations to sensors, modern scientific inquiry is awash in electronic information. At Lawrence Livermore, computer scientists supporting various projects and programs are developing methods and algorithms that provide new ways to tame, control, and understand these large amounts of data.

"Four key steps are involved in solving data-science problems," explains Dean Williams of Livermore's Global Security Principal Directorate. "One must organize the data—arrange the numbers and bits into meaningful concepts. One must prioritize—choose the most useful data when time and resources are limited. One must analyze—find meaning in images, graphs, data streams, and so on. Finally, it's important to make it easy for researchers to use the data; that is, we must create the methods and systems that help users query, retrieve, and visualize."

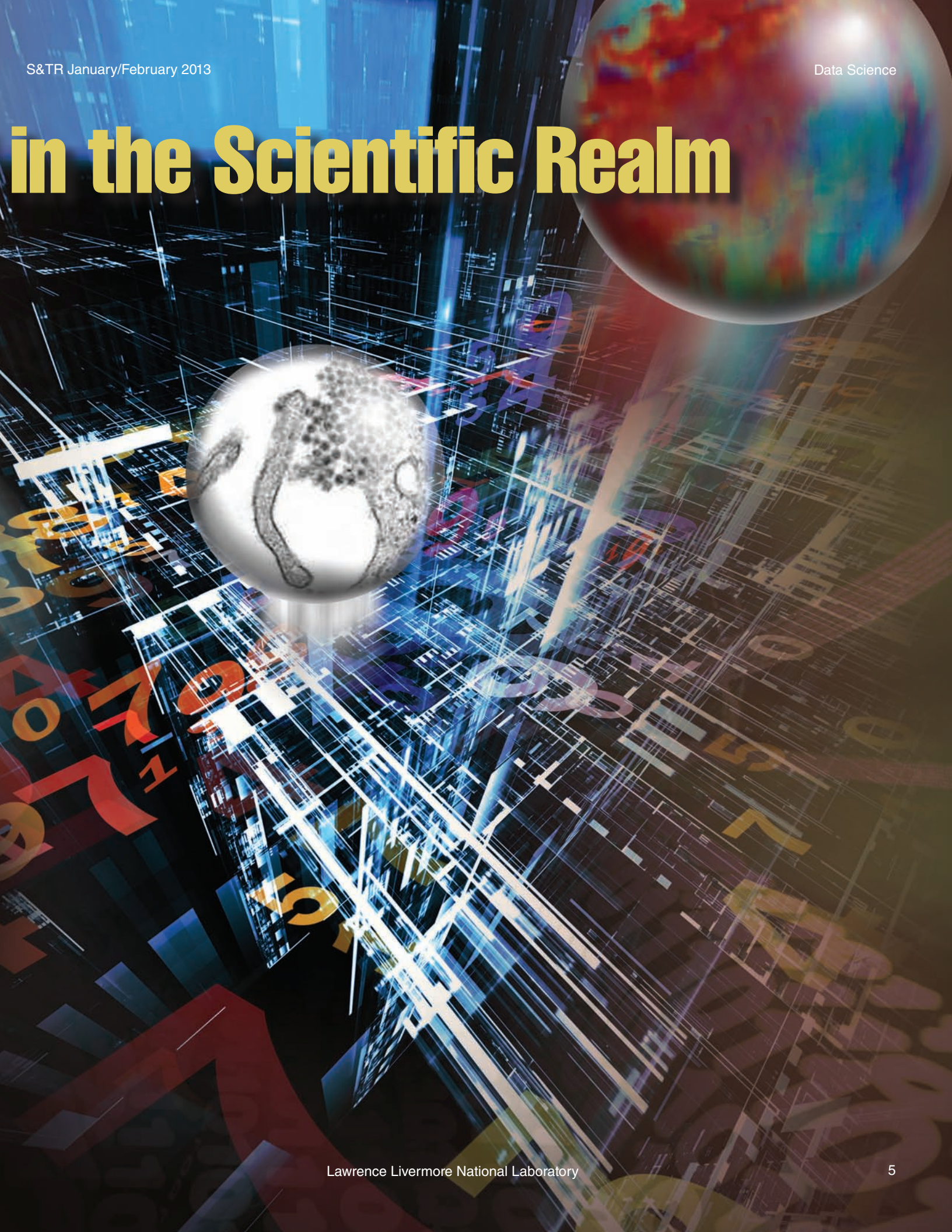
Three "V's" can sum up the type of data and the challenges involved: the variety, the velocity, and the volume. For example, in biological mission areas, the variety is high, but the velocity is low, with the volume that needs to be manipulated ranging from gigabytes to terabytes. By contrast, in the cybersecurity arena, variety and velocity are high, and the volume, which is continually changing as data streams past, can be very large. For climate research, the variety and velocity of data are also high, with an enormous volume accumulating in databases worldwide (from petabytes to exabytes).

The Laboratory, with its broad expertise in analysis, experience in storage technologies, and strong institutional computing culture, is addressing the data-science challenges of its programs and projects. These efforts range from devising methods for predicting the evolution of viruses, to creating tools for tackling streaming data in the cybersecurity arena, to fashioning accessible intuitive portals and analytics for climate scientists worldwide.

Getting Ahead of Viral Evolution

Scientists have been sequencing genomes for decades, with full genetic sequences now completed for more than 1,000 organisms as well as viruses—subcellular organisms with genomes consisting of RNA or DNA. Viruses are of particular interest because of their worldwide impact on health and welfare, the difficulty of combating them in nature, and the potential for their use as biological threat agents. (See *S&TR*, September 2012, pp. 6–13.) RNA viruses have exceptionally high mutation rates, enabling them to form mixed-variant virus populations, often referred to as "quasi-species." The high genetic variability within quasi-species helps these viruses adapt to different environments and hosts. Understanding such genetic diversity, especially in pathogenic viruses, is critical to developing accurate diagnostics and therapeutics. Although the information in existing DNA sequence databases is incomplete, the variety and volume of new data can be overwhelming.

in the Scientific Realm



To address this issue, computer scientists Adam Zemla and Tanya Kostova Vassilevska are developing a computational system called GeneSV and a stochastic simulation model to help predict viral evolution that could lead to the emergence of new strains or quasi-species clouds. Zemla's GeneSV allows characterization of possible sequence variations within a viral genome and makes predictions about the viability of a potential viral mutation. Vassilevska's model uses GeneSV results to further simulate viral evolution. Zemla and Vassilevska collaborated with researchers from the University of Texas Medical Branch (UTMB) at Galveston, who conducted experiments to test GeneSV's predictions. "This collaboration with experimentalists was like a dream project for a bioinformatician such as myself," says Zemla. "We designed and developed new algorithms, used these algorithms to create predictions and hypotheses, and worked with Galveston experimentalists to test and validate the predictions."

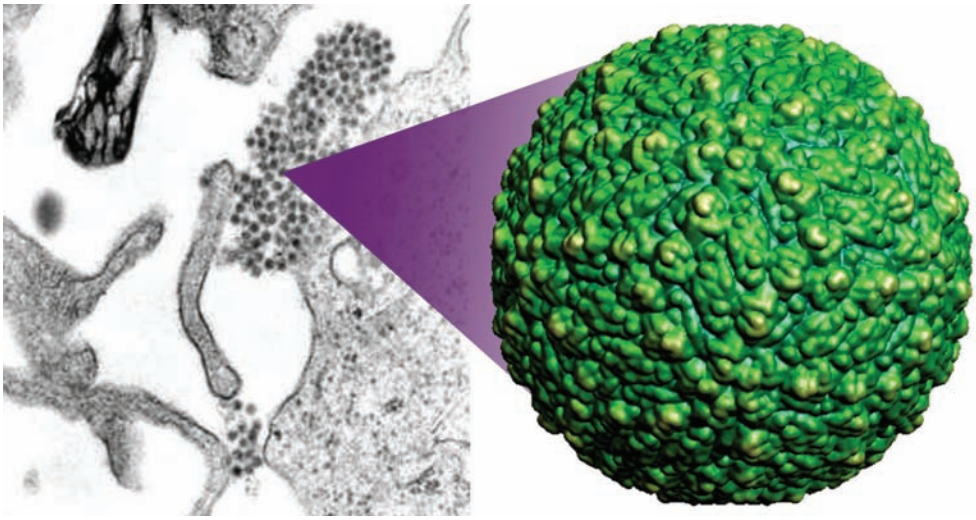
Funded by the Defense Advanced Research Projects Agency (DARPA), Livermore scientist Pejman Naraghi-Arani led a team to create these computational modeling tools that, when used with a novel microfluidics platform to grow viruses under many conditions, could evaluate and predict aspects of viral evolution. For this project, the team focused on a fast-evolving RNA flavivirus, the Dengue virus type 2 (DENV-2). Flaviviruses—a genus that includes the Dengue, West Nile, and yellow-fever viruses—mutate quickly and are of particular interest to biologists, health care specialists, and biodefense researchers.

"The RNA viruses have mutation rates between 10^{-3} and 10^{-5} per base per generation. In a single replication event of a virus, such as DENV-2 in a host cell, the majority of the produced genomes will have at least one variation," says Zemla. "Each host cell produces hundreds or thousands of progeny viruses, which can infect other host cells. This exponential growth of the number of multivariants helps the virus spread in the organism."

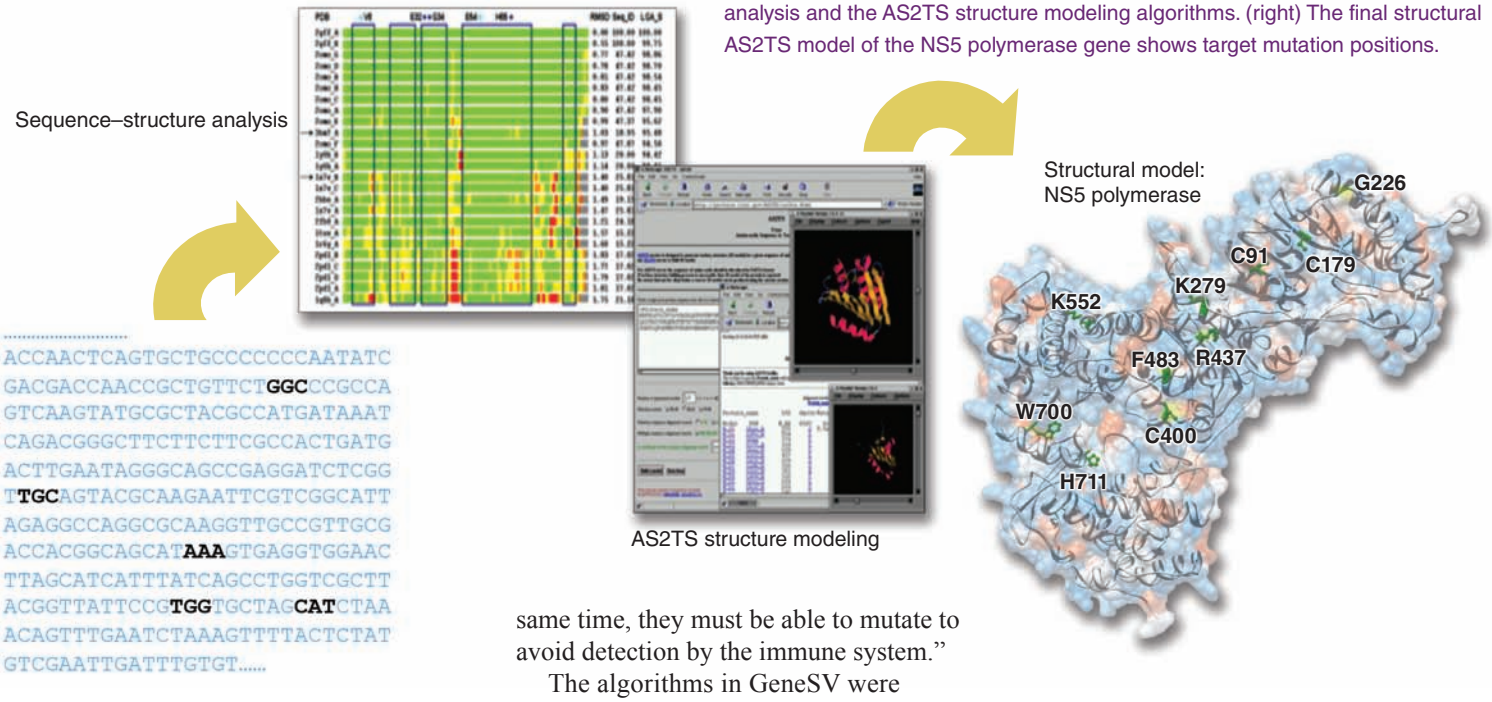
Public genomic databases used by researchers contain tens of thousands of genomic sequences as well as three-dimensional protein models of different viruses. However, this information encompasses only a part of the genetic diversity of viral species. "Furthermore," says Zemla, "the databases tend to be biased toward the dominant viral genome. It's unlikely that the full array of viable viral genotypes of a given species will ever be represented in these databases, because of sampling biases and the fact that these quasi-species clouds are constantly evolving." Because in vivo research and experiments cannot keep up with the millions of viral mutations and determine which genomic shifts lead to a viable virus, in silico predictive systems such as GeneSV are invaluable tools as they become more fully developed.

GeneSV uses information from existing genomic sequences, related protein sequences, and constructed protein-structure homology models to classify new variants and assess their viability. The system starts with a mutated gene sequence and searches the databases for similar sequences in closely related organisms. For example, if the input is a sequence from an unknown variant of DENV-2, GeneSV first looks in databases at sequences of other Dengue serotypes. (A serotype is a group of closely related microorganisms distinguished by a characteristic set of antigens.) It also examines sequences from more distantly related viruses, such as West Nile, and compares those sequences to the unknown mutant.

GeneSV identifies positions in the genomic sequences where the compared virus types—unknown and known—are different or similar. Then GeneSV



Using the Livermore-developed GeneSV computational system, Laboratory scientists analyzed never-before-seen designer mutants of an RNA polymerase gene (gene NS5) from the infectious Dengue virus type 2 (DENV-2) clone. Experimental tests conducted at the University of Texas Medical Branch at Galveston showed that the algorithm correctly predicted viability of evaluated mutants more than 80 percent of the time. In this transmission electron micrograph of Dengue virus virions (a virion is a single virus particle), the virions show up as a dark cluster of dots in the center. (inset) A capsid model simulates one virion.



(bottom left) A fragment of the DENV-2 genomic sequence shows highlighted mutation position candidates for GeneSV system analysis and experimental tests. (center) Two components of the GeneSV system are StralSV sequence–structure analysis and the AS2TS structure modeling algorithms. (right) The final structural AS2TS model of the NS5 polymerase gene shows target mutation positions.

converts the nucleic-acid sequences to protein sequences and continues similarity searches for closely related organisms on the protein level, using the same methodology as before. Finally, GeneSV builds protein structural models and conducts similarity searches against protein-structure databases.

For each level of evaluation—genomic sequence, protein sequence, and protein structure—GeneSV creates predictions and identifies characteristics that can be assigned to particular positions within the unknown, mutated sequence. One such characteristic is the location of the region in question on the protein structure, for instance, whether that region is buried inside or exposed on the surface. Another characteristic is sequence regions conserved within closely related organisms. "For example," says Zemla, "to invade a host cell, the virus binds to a specific receptor on the cell. The regions and residues responsible for binding need to be conserved in some ways, but at the same time, they must be able to mutate to avoid detection by the immune system."

The algorithms in GeneSV were designed to allow detection of such characteristics, with each algorithm focusing on a different aspect. For analysis on the protein-structure level, the system leverages the StralSV (structure alignment–based sequence variability) and AS2TS (amino-acid sequence to tertiary structure) algorithms developed under previously funded Laboratory Directed Research and Development (LDRD) projects to identify regions of sequence or structure variability. This diverse combination of sequence- and structure-based approaches results in a detailed analysis of the proteins of interest within a given genome.

For example, GeneSV was used to estimate the frequencies of mutations observed in codon (nucleotide triplet) positions in different genes within the set of all currently available DENV-2 genomic sequences. Results showed that the most mutable regions are in the envelope protein within segments characterized as helical, exposed, and predicted as antigenic determinants.

For the DARPA project, the team used GeneSV to evaluate two kinds of possible mutations for the NS5 polymerase gene from the DENV-2 virus: single-point mutations and compensatory mutations. (Compensatory mutations involve a process in which two or more residue positions, sometimes closely located in three-dimensional space, mutate simultaneously to preserve protein function.) GeneSV identified a set of proposed 32 single-mutation types, 31 of which had never been observed in any publicly available DENV-2 sequence. The UTMB scientists engineered each of the mutated viruses and experimentally confirmed that in 26 of the 32 mutations, GeneSV correctly predicted the viability of the mutants generated. GeneSV also generated five possible compensatory mutations in which double amino-acid substitutions occurred in two different parts of the genomic structure. Four of the five GeneSV predictions were experimentally proven to be correct.

The success of Phase I of the DARPA project has Zemla and colleagues hopeful for the future. With information such as that generated by GeneSV, medical researchers would have an advantage against possible attacks from all kinds of viruses. "Our approach can be applied to genomic sequences from any organism," says Zemla. "The current version of GeneSV has shown potential to help characterize possible variations in genomic sequences and sequence annotation efforts."

Finding Relevant Data in the Flood

The world is riddled with viruses of all kinds and not just of the biological variety. In the cybersecurity and surveillance arenas, threats akin to viruses abound, but they may go undetected within vast amounts of data. Livermore's Ana Paula de Oliveira Sales is leading an LDRD project focused on improving the ability to analyze streams of data that arrive at very high rates.

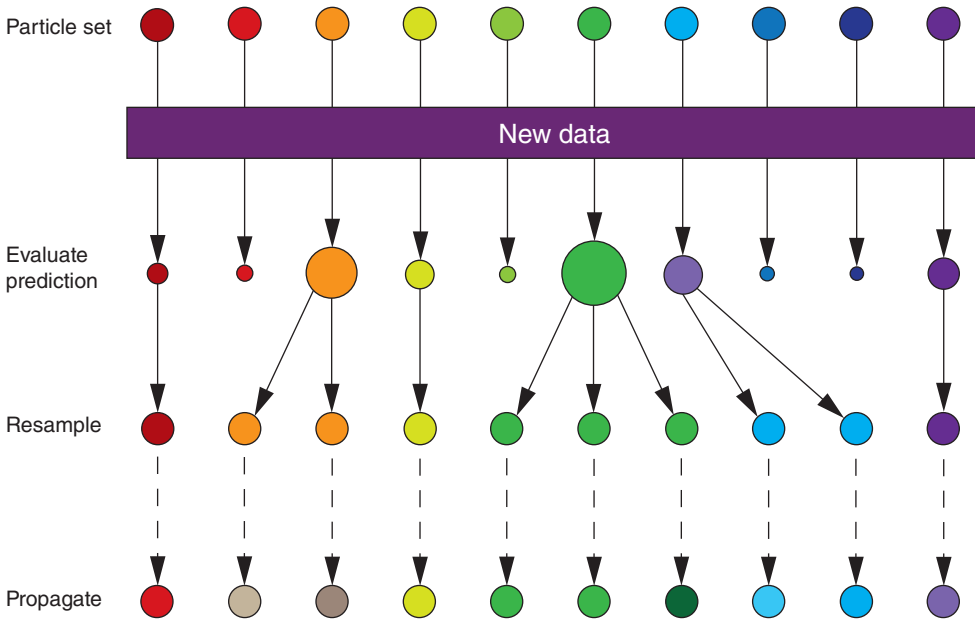
Her work could not only help fighters of cyber-crime but also benefit biosurveillance and real-time energy-distribution efforts. The challenges include the variety of data and the velocity at which it streams.

"In many domains of science, our ability to collect data continues to outpace our ability to analyze data," Sales explains. The deficit is particularly apparent in areas that involve streaming data. "For example, consider the common electronic communication that surrounds us every day, such as Twitter, texts, and Internet searches," says Sales. "For cyber-surveillance experts, monitoring the constant, data-rich activity is difficult. Storing all the information is impractical. Our project focuses on narrowing this gap between data collection and analysis rates by analyzing data 'on the fly' as the information streams past."

Sales notes that some of the key statistical problems underlying many of today's national security applications share a common need for continuously deployable, self-adapting learning systems. The goal of Sales's research is to create innovative computational learning algorithms that enable using sophisticated predictive-modeling techniques on modern streaming-data sources. This robust platform must quickly and accurately classify behavior and detect anomalies in large data streams, while effectively interacting with an ever-changing stream of data. "We need alternative predictive models for accomplishing standard learning tasks at a fraction of the computational cost," says Sales.

Sales and her colleagues designed and tested a highly customizable system to classify behavior and detect anomalies in large data streams. The code is implemented in high-performance C++, and the approach is flexible and applicable to a vast array of domains, such as real-time threat detection, video surveillance, energy distribution, and biosurveillance.

The system combines Bayesian techniques of "particle filtering" and time-evolving composite mixture models. "Particle filters are a kind of stochastic 'survival of the fittest' algorithm, evolving over time and adapting to their environment—in this case, incoming data," says Sales. The system begins with a data set for developing initial models, or particles. For example, if the system were used for spam detection, each particle would represent a possible model of spam. A particle would look at every new e-mail and answer the question: How probable is it that this e-mail is spam? "The question would be answered by considering variables such as sender, e-mail subject line, and



Particle filters are key to analyzing streaming data for anomalies. A particle set (a set of possible models) makes predictions about a new piece of data. The predictions are evaluated, and those models that are "more correct" are given more weight, while those that prove "less correct" receive less weight. In this way, the system evolves and changes character along with new, incoming data.

e-mail body," says Sales. "The beauty of using the composite mixture model is that it enables us to deal with a variety of data, numerical or categorical, and combinations thereof. Our model was intentionally built to be flexible and modular so it can be easily deployed to different domains."

When new data arrive to be analyzed, such as an e-mail message, the algorithm makes a prediction. In this case, the algorithm might determine that the e-mail message has a certain probability of being spam. The initial particles are resampled using weights that are proportional to how well each particle fits each new piece of information. As such, particles that are "more correct" receive more weight than those that are more incorrect, and they have a greater chance of surviving the resampling process. But even those that have less weight may continue to exist because of the stochastic nature of the system. "This feature is important," says Sales, "because it allows for greater diversity among the models. By having a variety of particles, we can be more confident about how well our ensemble represents the data."

The resampled particles then update their probability densities using relevant information from the data point (in this example, the e-mail message). Over time, the ensemble of particles becomes a more accurate representation of the data. This evolution allows the model to more effectively counter the adversary, while the adversary—in this example, the spammer—is also evolving to bypass the filter system.

To date, the Livermore-developed algorithm has primarily been used to analyze data sets that evolve gradually. A current research direction for the project, now in the second year of its three-year LDRD funding, involves teaching the algorithm to accommodate sudden-switch situations. Examples of such situations include an abrupt change in power distribution as a result of a local blackout or the attack of a new computer virus. In these cases, the system needs to be able to "switch on a dime" when vital information changes suddenly and without warning.

Going forward, the project's primary focus is increasing computational performance, so that the system can be deployed for data streams with higher and higher rates of data arrival. One technique Sales and colleagues are pursuing is to make the approach more parallelizable by creating specialized small filters, or "ensembles of ensembles." Replacing a single, large particle filter with many smaller, parallel filters should significantly improve computational speed at little or no cost to prediction performance.

Sales expects that the most gains in computational speed will come from using adaptive sampling theory. "Updating the probability densities with each new data point is the most computationally expensive aspect of the model," says Sales. "We'll use adaptive sampling to intelligently choose data points with high information value for updating the model." The Laboratory has significant expertise in adaptive design of computer experiments in cases where data are scarce and data points sparse. "Using this technique, we could restrict expensive model updates to those observations that improve our understanding of the system dynamics," Sales explains.

Coding for Collaboration

Laboratory scientists are also devising methods that make it easier to deal meaningfully with enormous quantities of stored data. While the physical storage of exabytes (10^{18} bytes) of data is achievable today, the challenge is to make the data meaningful and widely used. Nowhere is this issue more apparent than in the realm of climate research, where the amount of information at researchers' virtual fingertips continues to grow at an enormous rate. The types of information resources include observational data from National Aeronautics and Space Administration satellites and instruments; in situ, ground-based data such as albedo measurements and surface and air temperatures; simulations data; and reanalysis data (a mix of observational and model data). In climate research, all three of the basic challenges are present: volume, velocity, and variety of data.

The gathering and sharing of climate data is a key effort of the Coupled Model Intercomparison Project (CMIP), which is the worldwide standard experimental protocol for studying the output of coupled atmosphere–ocean general circulation models. Established in 1995 by the World Climate Research Programme's Working Group on Coupled Modelling (WGCM), CMIP provides a community-based infrastructure in support of climate model diagnosis, validation, intercomparison, documentation, and data access. Through CMIP, scientists are able to analyze general circulation models in a systematic fashion, a process that serves to facilitate model improvement. Virtually the entire international climate modeling community has participated in this project since its inception.

Williams, who leads the Earth System Grid Federation (ESGF) project at Livermore and abroad, remembers that in 2003, to gather the data used for the third CMIP (CMIP3), large-size "bricks" that could hold a single terabyte of data were shipped around the globe. Climate researchers loaded their data to send back to Livermore, where the entire 35 terabytes of data were then stored in a single centralized location. "We also created a Web portal so the user community could access the database," says Williams.

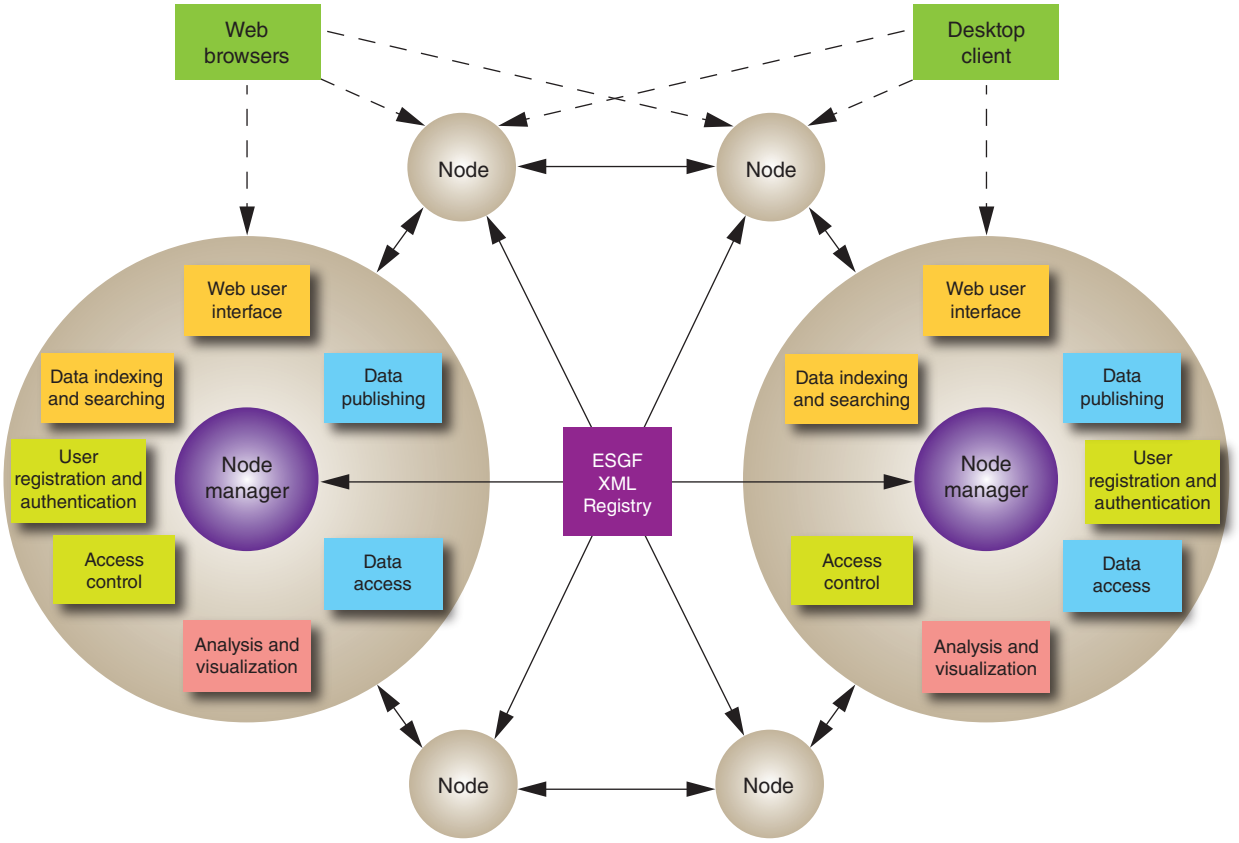
Fast forward to 2006, when WGCM agreed to promote a new set of coordinated climate model experiments. These experiments comprise the fifth phase of CMIP. CMIP5 is providing a multimodel context to assess the mechanisms responsible for model differences associated with the carbon cycle and with clouds, examine the ability of models to predict climate on decadal timescales, and more generally, determine why similarly forced models produce a range of responses.

For CMIP5, the total data are estimated to reach 3.1 petabytes (10^{15} bytes)—two orders of magnitude more than the total of CMIP3. To increase the accessibility and usefulness of the mountains of CMIP5 climate data, the international ESGF was formed (www.esgf.org). The federation grew out of the larger Global Organization for Earth System Science Portals community. Collaborating partners in the federation include Livermore and other national laboratories as well as a host of other organizations such as the German Climate Computing Centre and the University of Tokyo Center for Climate System Research.

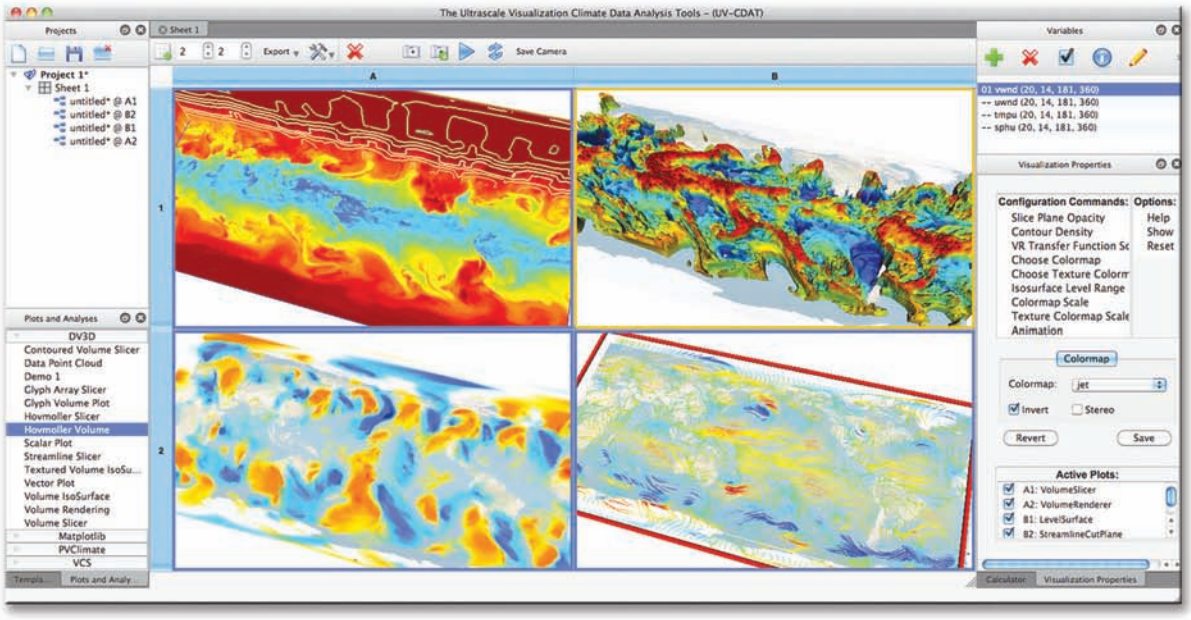
The ESGF portals are gateways to scientific data collections hosted at sites around the globe. Gateways are Web portals that allow the user to register and potentially access the entire ESGF network of data and services. Currently more than 24 portals are in use, including Livermore's Program for Climate Model Diagnosis and Intercomparison.

Instead of consolidating all the data from numerous locations, as was done for CMIP3, ESGF uses distributed storage in conjunction with software that harvests and integrates the data. "Far too much data exists to ship it about physically, so we took a decentralized, cloud-type approach," says Williams. Modeling research centers download the ESGF software stack from the Livermore servers. Centers then use this software to publish their data to the federation for harvesting. Users access all data as if they are on one centralized archive system.

To date, ESGF has made 60 CMIP5 simulation model runs (more than 1.8 petabytes) from 25 climate research centers, which are available to users worldwide. ESGF provides access to 18 highly visible national and international climate data products, with more on the way. As a result, ESGF offers a promising option for



The Earth System Grid Federation (ESGF) peer-to-peer architecture is based on the principles of modularity and equality. Each system node (blue, orange, pink, chartreuse) can offer different services depending on how it is configured. Some nodes can be scaled differently, for example, to provide increased computational resources. All nodes interact as equals, so no single points of failure arise.



The Ultrascale Visualization Climate Data Analysis Tools display visualization results from different geographical data sets around the world via the ESGF distributed archive. (Shown here are atmospheric simulations.) Users can select from a variety of analytical and visualization tools such as CDAT, R, DV3D, VisIt, ParaView, and Matplotlib. The framework is flexible enough that other tools can be added at any time. Workflows can also be stored for later use and sharing with others.

building a collaborative knowledge system in the climate community.

The system makes all of this data-handling transparent to the user, while still allowing for local ownership. "If a modeling center improves a computer model and produces new output, the system handles archiving and notifications to those people who use the data," explains Williams.

The ESGF peer-to-peer architecture is based on the concept of a dynamic system of nodes that interact on an equal basis and offer a broad range of user and data services, depending on how each node is set up. This extensible and scalable system supports geospatial and temporal searches and includes a dashboard that shows system metrics, a user interface for notifications, and a rich set of analysis tools to help manipulate the data. For example, the Ultrascale Visualization Climate Data Analysis Tools have workflow scripts that automate scientific analysis and visualization, making it easy for users to re-run analyses and to work together, which encourages collaboration and openness in scientific enquiry.

Williams envisions systems that will make it even easier for scientists to collaborate in the future. "I'd like to see all of the visualization tools put on the back end, with the results easily available on a laptop," he says. "We'd like to pull in more elements, such as Twitter and Whiteboard, and have people be able to use their tablets, smartphones, and whatever else may be coming in the years ahead." Williams is also looking toward the creation of an ESGF "Lite" version that would work similarly to today's social networks. By 2020, ESGF will embrace an estimated hundreds of exabytes; thus, tools for the future will be welcome.

Looking toward a Data-Rich Future

In nearly every arena of scientific inquiry imaginable, the amount of data needed to be organized, prioritized, analyzed, and utilized will continue to increase in velocity, variety, and volume. The data sources in many cases are huge and growing dynamically. Whether it's forecasting ecological tipping points, predicting and mitigating energy grid instability, or identifying new computer viruses, the tools must evolve to meet and match the needs of researchers. "Prioritization will become ever-more important as data volumes and velocities grow," says Williams. "Organization will also gain importance as data sets become increasingly complex." The present challenges of analysis and utilization will remain as well. Livermore computer scientists and researchers are already addressing many of the challenges and have their sights set on the data-rich future that is just around the corner.

—Ann Parker

Key Words: algorithm, biosurveillance, climate research, Coupled Model Intercomparison Project (CMIP), cybersecurity, cyber-surveillance, data science, Earth System Grid Federation (ESGF), energy distribution, gene sequencing, GeneSV, genome, particle filtering, streaming data.

For further information contact Dean N. Williams (925) 423-0145 (williams13@llnl.gov).