Better Data Just Can't Wait

Imagine this: You're a scientist with time scheduled on a high-end user facility, such as a powerful x-ray source, particle accelerator, or observatory. You wrote a proposal to use the facility, were one of the fortunate 10 or 20 percent to receive approval, obtained funding, and showed up for your 72-hour time slot. You successfully crammed in all your experiments—working night and day with brief naps and take-out meals—and returned home with a massive amount of new data.

Now imagine this: Half a year later, you're still chugging through the terabytes of experimental data, and you make an unexpected discovery. Maybe it was a faulty measurement or some incorrect setting when you conducted the experiment. Or maybe it's something groundbreaking, something you hadn't planned to look for that just barely showed up in a small subset of the data. Either way, here you are, your time on the fancy user facility long gone, and only *now* do you see what you really *should have* measured while you were there. Only now do you see that the tsunami of data you did collect has errors or is in some other way less valuable than it could have been.

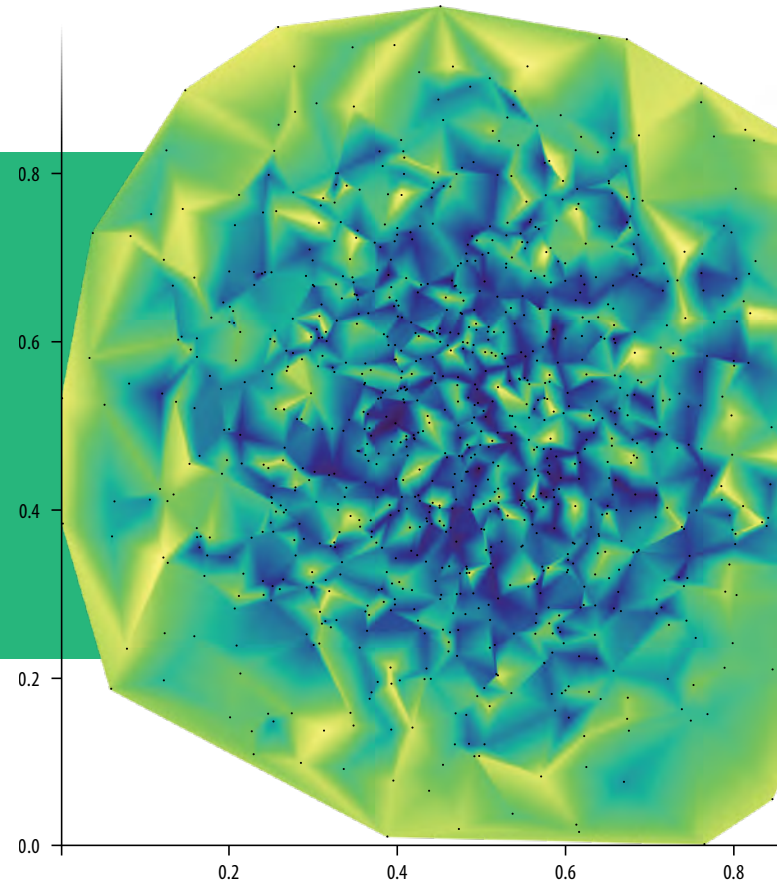
It's not such a rare occurrence. Much of modern science is so complex that it requires time-consuming supercomputer simulations to interpret the effects of adjusting experimental parameters and equally time-consuming data reduction to interpret the results. Furthermore, as experimental facilities have grown more sophisticated in their instrumentation and data collection, the volume of data they produce has grown to the point where it takes months or even years to pore through it and see what it really contains. And by then, it may be too late.

That's why Laboratory computer scientist James Ahrens, experimental materials scientists Cindy Bolme and Richard Sandberg, statistician Earl Lawrence, and a team of other scientists of various disciplines started the ASSIST project. Short for Advanced Simulation, experiments, Statistics, and Information Science and Technology, ASSIST defines a software workflow that allows scientists to compare simulation results with experimental data as the data is generated and presents the results in a ready-to-interpret visual format. This enables scientists to make key decisions on the fly about what to measure next—that is, what experiments are most worth doing in the limited time available.

ASSIST accomplishes this with an emulator, a fast statistical model that replicates a more computationally expensive simulator. Using previously obtained simulation results for different combinations of input parameters, the emulator is trained to predict variations in what the full simulation would produce with different inputs. ASSIST quickly extracts only the most relevant measurement data and compares that with the emulator's predicted outcomes. It then employs powerful visualization tools to show the comparison graphically, in various formats, to scientists who might want to change tack as a result. And for experiments intended to seek a desirable set of output values, it can estimate the proper conditions to help the experimenter get there. Critically, the whole process is fast enough to help scientists make decisions during their brief experimental window.

But perhaps ASSIST's greatest strength is its ability to identify exactly where things get interesting and update that assessment with each new data point it receives. It can analyze a large multidimensional input and output space and suggest combinations of parameters to test in order to gain the best information about the least-understood aspects of the system under study. It identifies which combinations of input settings change the outputs the most and which settings are so sensitive that dialing them a hair in either direction means the difference between night and day—right there, right then, for the scientist racing to collect data all night and all day. LDRD

—*Craig Tyler*



ASSIST data visualization comparing two input parameter values, one on each axis (for example, the temperature at which an experiment is being conducted and the intensity of a laser being used), with the color scale representing uncertainty in current knowledge. The *x* and *y* coordinates of the most uncertain regions (yellow) indicate the experimental-parameter settings that would reveal the greatest amount of new information.