Editor: Geoffrey Fox, gcf@indiana.edu

E-Science Meets Computational Science and Information Technology

By Geoffrey Fox

OVER THE LAST DECADE OR SO, THERE HAS BEEN MUCH DISCUSSION OF AND PROGRESS IN COMPUTATIONAL SCIENCE. THE US HIGH PERFORMANCE COMPUTING AND COMMUNICATIONS INITIATIVE AND THE NATIONAL SCIENCE FOUNDATION GRAND Challenge programs largely focused on this trend. However, a subtle change has taken place recently. For example, although the NSF initiated the Information Technology Research program based on recommendations from the President's Information Technology Advisory Committee, there is no parallel Computational Science Research program nor a President's Computational Science Advisory Committee. Classic computational goals are just one part of the IT agenda—a part that's getting smaller. We see the same trend in research, with work on grids and distributed computing overshadowing that on classic parallel computing. Now the NSF has a new cyberinfrastructure report (www.cise.nsf.gov/b_ribbon) continuing the same emphasis on distributed systems. What does this mean for computational science and its associated technology and research?

A fourth paradigm of scientific study?

We used to speak of three approaches to science: experiment, theory, and the emerging computational view. Does IT and cyberinfrastructure enable a fourth paradigm? This does not seem likely, but the National Virtual Observatory (www.us-vo.org) eloquently describes a new approach to astronomy. Just a few years ago, astronomy involved individual groups designing new instruments, developing particular observational strategies, and managing the data gathered via space- or ground-based instruments. Scientists lovingly analyzed this data to discover and publish new scientific insights. This "stovepipe" approach to observational (experimental) science is typical in many fields. It ensures that data is properly interpreted with the usually difficult instrumental corrections properly applied. Note that it is usual to compare "reasonably corrected experimental data" with hypotheses (models) to which other instrumental effects are applied. Often it is not possible to remove all instrumental effects from a data set to compare it with a pristine theory (developed by a "pristine" theorist who needs no significant understanding of the instrument).

In the new approach to astronomy, investigators will have the best analysis and visualization capabilities connected to the global Internet. This analysis will integrate data from multiple instruments spanning multiple wavelengths and regions of the sky. This vision stresses IT-enabled high-performance networks, data access, management, and processing.

A similar story can be told in other fields. Bioinformatics involves the integration of gene data banks scattered around the Internet; visionaries imagine this extended to include massive amounts of data associated with individuals, enabling a new, personalized medicine. High-energy and nuclear-physics experiments will involve thousands of physicists analyzing petabytes of data coming from a new generation of detectors at high-intensity accelerators (such as CERN's Large Hadron Collider). Fields such as climate, environment, earthquake, and weather research will also benefit from what has been termed the data deluge. Of course, the gathering and computer-based analysis of data is not new; in fact, the World Wide Web originated in CERN from a tool to support transcontinental high-energy physics collaborations. Rather, the scale (the amount of data) and breadth (the integration of data sets) have changed dramatically.

The virtual lab

Computational science was built on the vision that computers would represent a virtual laboratory where we could explore new concepts through simulation and compare these with experimental data. Successful agency supercomputers (at the NSF, the US Defense and Energy Departments, NASA, the National Institutes of Health,

1521-9615/02/$17.00 © 2002 IEEE

and elsewhere) have supported a growing interest in this mode of computational science. We understood that Moore's law would inevitably drive this field with steadily increasing simulation, storage, and networking performance. Now we see that advances in device physics are driving both computer and instrument performance. Thus, the data deluge is reinvigorating and reshaping observational fields.

This phenomenon is a major force in the current British e-science (see www.escience-grid.org.uk) and NSF cyberinfrastructure initiatives. We can and should integrate these themes, as Figure 1 shows. Sensors will produce a lot of information, but so will simulations. Various techniques—visualization, statistical analyses, data mining, and so on—will extract knowledge from the information that we glean from simulations and raw data sources. These will be fed back into theoretical science, and the classic collaboration between the twin pillars of theory and experiment will advance our scientific fields. Computational science will concentrate on enhancing theoretical science, with IT having a major focus on both experimental science and the distributed infrastructure for all aspects of research.

Of course, these labels are rather arbitrary; we can use *computational science* in either a narrow fashion as I've just done or to describe the entire process represented in the figure. Alternatively, we could call the whole process *computational science and information technology*. The confusion in the term's meaning has perhaps handicapped the widespread emergence of computational science as a separate academic field. It is safe to assert that computing is of growing importance in all aspects of science even while theory and experiment (observation) remain the dominant methodologies.

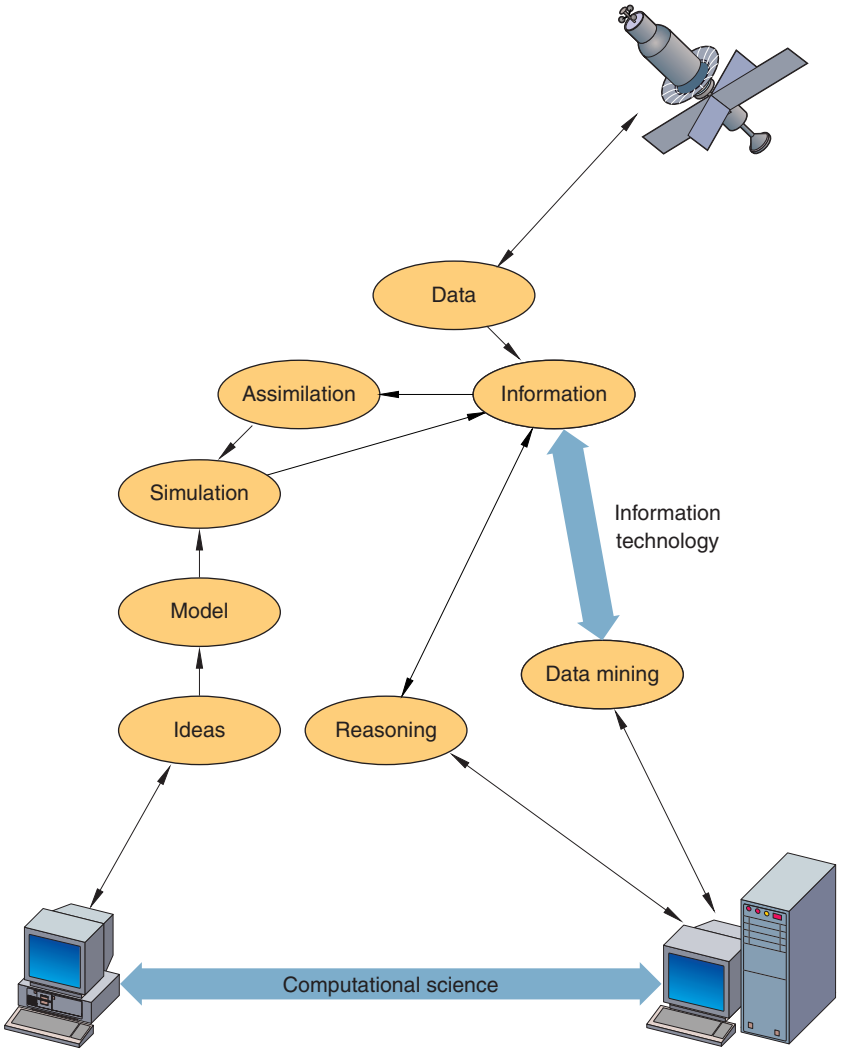

Figure 1. Computational science and information technology in the service of science.

In earlier issues, I described key technologies for e-science: peer-to-peer networks, grids, and XML. One broad area still to be covered is the step from information to knowledge. I could cover techniques such as the Semantic Web (www.w3.org/2001/sw), stressing the use of XML-based metadata to express linked "information nuggets" from which knowledge comes as an emergent phenomenon. The vision of the DoE's Accelerated Strategic Computing Initiative, that high-fidelity simulations can produce knowledge with modest observational support, is important here. A hallmark of the next decade will be the integration of such simulations with the data deluge. Here, data assimilation (common already in weather and other fields), closely integrating time-dependent simulation and observation, will become increasingly important.

So what specific technologies will e-science need? Stay tuned…. CiSE

Geoffrey Fox is director of the Community Grids Lab at Indiana University. He has a PhD in theoretical physics from Cambridge University. Contact him at gcf@indiana.edu; www.communitygrids.iu.edu/IC2.html.