# e-Science

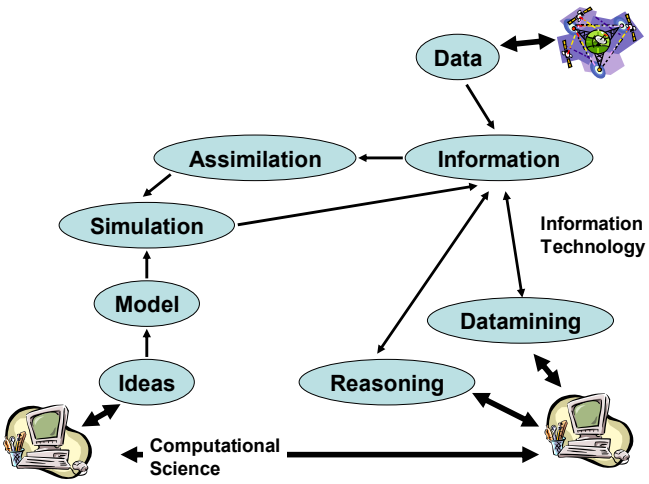
Geoffrey Fox
*Indiana University*
*Computer Science, Informatics and Physics*
*Community Grid Computing Laboratory,*
*501 N Morton Suite 224, Bloomington IN 47404*
gcf@indiana.edu

**e-Science meets Computational Science and Information Technology**

Over the last decade or so, there has been much discussion of and progress in computational science. There was the HPCC (High Performance Computing and Communications) Initiative and the NSF Grand Challenge programs which largely focused on this. However recently there has been a subtle change – for example NSF initiated the ITR (Information Technology Research) program based on the recommendations of PITAC – the President's Information Technology Advisory Committee. We did not have the CSR (Computational Science Research), CSITR or PCSAC. One will find classic computational goals as one part of the Information Technology agenda but they are just one part and perhaps a part that is getting smaller. One sees the same trend in research with work on grids and distributed computing overshadowing that on classic parallel computing. Now NSF has a new cyberinfrastructure report ((http://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?

Once we spoke of the three approaches to science: experiment, theory and the emerging computational view. Does Information Technology and cyberinfrastructure enable a fourth paradigm? Such a proliferation of scientific methods does not seem too reasonable. However the National Virtual Observatory (http://www.us-vo.org/) eloquently describes the new approach to astronomy. Just a few years ago, astronomy involved individual groups designing new instruments, developing particular observing strategies and managing the data taking on some instrument either in the space or on the ground – this data was lovingly analyzed to discover and publish new scientific insights. This "stove-pipe" approach to observational (experimental) science is typical in many fields. It ensures that those who build the equipment can ensure their 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 dataset so that it can be compared with a pristine theory (developed by a pristine theorist who needs no significant understanding of the instrument). Nevertheless we imagine a new astronomy where our investigator has the best analysis and visualization capabilities connected to the global internet. This analysis will integrate the data from multiple instruments spanning multiple wavelengths and multiple regions of the sky. This vision stresses the high performance networks, data access, management and processing enabled by "information technology". A similar story can be told in other fields. Bio-informatics involves the integration of gene data

banks scattered around the Internet; visionaries imagine this extended to include massive data associated with individuals and enabling a new personalized medicine. High energy and nuclear physics experiments will involve thousands of physicists analyzing petabytes of data each year coming from a new generation of detectors at high intensity accelerators such as the LHC at CERN. Fields such as climate, environment, earthquake and weather will also benefit from what has been termed the data deluge. Note the gathering of data and its computer-based analysis 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 (amount of data) and breadth (integration of datasets) has changed dramatically.




We can simplistically summarize the situation this way. Computational Science was built on the vision that computers would represent a virtual laboratory where one could explore new concepts from simulations and comparison of these with experimental data. The very successful agency supercomputers (NSF DoD DoE NASA NIH and others) 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 re-invigorating and re-shaping observational fields. This phenomenon is a major force in the current e-Science (as pioneered in the UK at http://www.escience-grid.org.uk/) and NSF cyberinfrastructure initiatives. In fact as shown in the figure one can and should integrate these themes. Sensors will produce a lot of information but so also will simulations. Various techniques: visualization, statistical analyses, "datamining" will extract knowledge from the information gleaned from simulations and raw data sources These will be fed back into theoretical science and so the classic collaboration between the twin pillars – theory and experiment – will advances our scientific fields. So we find computational science is largely focused on enhancing theoretical science with information technology having a major focus on both experimental science and the distributed infrastructure for all aspects of research. Of courses these labels are rather arbitrary and one can use computational science in either a narrow fashion as above or to describe the entire process represented in the figure. Alternatively one could call the whole process CSIT – computational science and information technologist. The confusion in the meaning of the term has perhaps handicapped the wide spread emergence of computational science as a separate academic field. Safely we can assert that computing is of growing importance in all aspects of science even while theory and experiment (observation) remain the dominant methodologies.

In earlier articles in this series, we described key technologies for e-Science: Peer-to-peer networks, Grids and XML. We will explore these in more detail in later articles.

One broad area to be discussed is the step from information to knowledge. Here we could cover techniques like the Semantic Web (http://www.w3.org/2001/sw/) stressing the use of XML based meta-data to express a large number of linked "information nuggets" from which knowledge comes as an emergent phenomenon. The vision of DoE's ASCI (Accelerated Strategic Computing Initiative) program is important here; this activity asserts that high fidelity simulations can produce knowledge with modest observational support. More generally we suggest a hallmark of the nest 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 can be expected to increase in importance.

This article has tried to set up the motivation for a following sequence on the detailed technologies needed by e-Science. Stay tuned ..,..