From Data to Knowledge to Action:
Enabling 21st Century Discovery in Science and Engineering

Randal E. Bryant
Carnegie Mellon University

Erwin P. Gianchandani
Computing Research Assoc.

Edward D. Lazowska
University of Washington

Computing Community Consortium
Version 13: June 18, 2010[1]

A new revolution in discovery and learning

Recent rapid advances in information and communication technologies – both hardware and software – are creating *a new revolution in discovery and learning*.

Over the past several decades, computational science – the large-scale simulation of phenomena – has joined theory and experiment as a fundamental tool in many branches of science and engineering.

Today we are at the dawn of a second revolution in discovery – a revolution that will have *far more pervasive impact*. The focus of this new approach to science – called *eScience* – is *data*, specifically:

- the ability to collect and manage orders of magnitude more data than ever before possible;
- the ability to provide this data directly and immediately to a global community;
- the ability to use algorithmic approaches to extract meaning from large-scale data sets; and
- the ability – and, in fact, the *need* – to use computers rather than humans to guide the hypothesis/measurement/evaluation loop of scientific discovery.

Enormous numbers of tiny but powerful sensors are being deployed to gather data – deployed on the sea floor, in the forest canopy, on the sides of volcanoes, in buildings and bridges, in living organisms (including ourselves!). Modern scientific instruments, from gene sequencers to telescopes to particle accelerators, generate unprecedented amounts of data. Other contributors to the data tsunami include point-of-sale terminals, social networks, the World Wide Web, mobile phones (equipped with cameras, accelerometers, and GPS technology), and electronic health records. These sensors, instruments, and other information sources – and, indeed, simulations too – produce enormous volumes of data that must be captured, transported, stored, organized, accessed, mined, visualized, and interpreted in order to extract knowledge and determine action.

This "computational knowledge extraction" lies at the heart of 21st century discovery.

[1] Contact: Erwin Gianchandani, Director, Computing Community Consortium (202-266-2936; erwin@cra.org). For the most recent version of this essay, as well as related essays, visit http://www.cra.org/ccc/initiatives

A national imperative

The fundamental tools and techniques of eScience include sensors and sensor networks, backbone networks, databases, data mining, machine learning, data visualization, and cluster computing at enormous scale. eScience, even more than computational science, illustrates the extent to which *advances in all fields of science and engineering are married to advances in computer science and the mathematical sciences*.

Traditional simulation-oriented computational science was transformative, but it was a niche. Simulations can predict the behaviors of systems given underlying models, but this is only one aspect of the scientific process. In contrast, eScience – 21st century computational science – will be pervasive, affecting a broad spectrum of investigators and a broad spectrum of fields, since *virtually all of science requires extracting useful insights from the data arising from measurements and simulations*.

Thus, to ensure our nation's future competitiveness, investments of two forms are required:

- We must be the world leader in inventing successive generations of eScience tools and techniques.
- We must be the world leader in the application of these tools and techniques to move from data to knowledge to action – including real-time, mission-critical decision making.

Widespread recognition of the need

The Sloan Digital Sky Survey was the previous decade's most visible eScience project. It has been estimated that the widespread dissemination of SDSS data by means of a Web front-end to a commercial relational database system *increased by an order of magnitude the amount of science that was accomplished* – as well as driving a dramatic "democratization" of access.

The next-generation astronomical survey project – the Large Synoptic Survey Telescope – will generate as much data *every two days* as SDSS generated in seven years! The Large Hadron Collider will generate *two SDSS's worth of data each day*! A single lab equipped with several dozen modern desktop gene sequencers generates an SDSS worth of data each day. The NSF Ocean Observatories Initiative will transform oceanography from an expeditionary science to an observatory-based science, with many thousands of chemical, physical, and biological sensors streaming data in real time to scientists and schoolchildren alike. Social scientists who once studied the creation, evolution, and dissolution of cliques by paying a dozen undergraduates to participate in a focus group can now mine the behavior of 400 *million* Facebook users.

It is not just the volume of data that is driving change; it is also the rate at which data are being generated, and the dimensionality. At every level of the "discovery pyramid," from "small science" to "big science," *the nature of discovery is changing rapidly and dramatically*. Furthermore, these advances in eScience are beginning to drive new fields of research, such as:

- "Astroinformatics" – large-scale exploration of the sky from space and from the ground, requiring data mining, analysis, and visualization;

- Chemistry and materials science, including “matinformatics” – real-time chemical analysis of complex sample mixtures, requiring data unification, uncertainty quantification, clustering, and classification; and
- Systems biology – systems analysis of underlying biochemical interactions that give rise to biological functions and behaviors, requiring data unification, clustering, classification, feature detection, information extraction, uncertainty analysis, anomaly detection, and optimization.

These eScience-driven fields, in turn, are changing our lives. For instance, systems biology is giving rise to personalized medicine approaches, enabling us to lead longer, healthier lives.

Importantly, eScience can assist in both the front end and back end of traditional simulation-oriented computational science. On the front end, it can help in model development, e.g., in trying to hypothesize a model of bird migration based on observational data. On the back end, it is a key tool for analyzing the data generated by big simulations, e.g., in trying to determine whether the model of bird migration validates (or refutes) the underlying model hypotheses. The system can then automatically iterate the hypothesis generation/evaluation loop by refining the model parameters and re-running the simulation and evaluation, a process known as *active learning.*

Traditionally, the term “visualization” refers to creating graphic displays so that humans can analyze and understand the data generated by simulations. But now this term is starting to refer to a much more powerful capability: *having computer programs be the visualizers*. Dramatic algorithmic improvements are yielding computer programs capable of detecting subtle phenomena. In the next decade, they will take on an even more active role, guiding the core process of scientific discovery.

For background in the nature and magnitude of the change, see the 2005 compendium *2020 Science*[2], the associated special issue of *Nature* titled “2020 Vision”[3], the 2009 compendium *The Fourth Paradigm*[4], and the 2010 workshop from the NSF MPS Advisory Committee *Data-Enabled Science in the Mathematical and Physical Sciences*[5].

Intellectual infrastructure: The essential investment

The National Science Foundation’s investment in cyberinfrastructure is often thought of as being primarily about physical rather than intellectual infrastructure. In truth, the NSF Supercomputer Centers program and the NSFNET program each built *communities of expertise* – intellectual infrastructure whose importance greatly exceeded that of the related physical infrastructure.

The same is true for eScience. A broad set of intellectual investments are required. These are detailed, for example, in the 2008 white paper *Big-Data Computing: Creating revolutionary*

[2] http://research.microsoft.com/en-us/um/cambridge/projects/towards2020science/downloads/T2020S_ReportA4.pdf
[3] http://www.nature.com/nature/journal/v440/n7083/edsumm/e060323-01.html
[4] http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf
[5] http://www.cra.org/ccc/docs/reports/DES-report_final.pdf

breakthroughs in commerce, science, and society[6] and cover all of the tools and techniques noted above: sensors and sensor networks, backbone networks, databases, data mining, machine learning, data visualization, and cluster computing at enormous scale. They are also detailed in the Executive Summary of the MPS Advisory Committee workshop referenced earlier[5].

The focus of this current paper, as captured in the title, is the subset of the above targeted at "data analytics" or "scientific inference." These tools and techniques – such as data mining, machine learning, and data visualization – come from computer science and the mathematical sciences. They have the potential to dramatically advance our ability to assimilate, handle, and exploit massive, complex data by facilitating uncertainty quantification, clustering and classification, feature detection and information extraction, anomaly detection, optimization, etc. – all in real time for mission-critical decision-making. They will allow us to standardize data, enabling the same data set to be used for multiple purposes, and they will dramatically improve our ability to visualize large, multi-dimensional data.

Dramatic investments in this intellectual infrastructure – in advancing the tools and techniques of eScience – are essential.

Facilitating intellectual growth through Federal investment

As suggested in part by participants of the NSF MPS Advisory Committee's recent workshop[5], the Federal government has a critical role in fostering the growth of this intellectual infrastructure. Federal agencies can support:

- Fundamental research that advances the core tools and techniques underlying eScience, as well as efforts that apply these tools and techniques in support of scientific discovery;
- Highly collaborative, multi-disciplinary groups of researchers, spanning both core scientists as well as computational and eScience researchers;
- Data sharing protocols through incentive-based funding opportunities;
- Professional communication through workshops and conferences focused on data-driven science; and
- Stronger programs for education and outreach, highlighting the value of data-driven science.

Much of the underlying computing technology to support eScience comes from other application domains, and hence the scientific world can benefit from the tremendous commercial and research activity underway to deal with the massive amounts of data being generated, stored, and analyzed in all facets of society. For example, Internet companies such as Google, Amazon.com, Yahoo!, and Microsoft have built upon the results of two decades of distributed systems research to construct reliable and cost-effective computer systems that dwarf the size of the world's largest supercomputers. These systems are no match to supercomputers in terms of computing power for traditional computational science applications, but they are far more effective at collecting, storing, and analyzing large-scale data. New parallel processing techniques and new database technologies promise to enable data-driven science to achieve the scale and processing capacity required to make scientific breakthroughs across many disciplines.

[6] http://www.cra.org/ccc/docs/init/Big_Data.pdf

However, eScience has some important differences compared to the types of applications implemented by Internet companies. For example:

- eScience has a much greater need for careful stewardship of data, so that the provenance can be tracked in validating any resulting conclusions, and results can be reproduced by other scientists. Imagine, for example, using an eScience system to analyze data gathered during trials of a new drug to determine whether it is safe for general use. Internet services such as search engines and book recommendations have much weaker correctness requirements than is the case for eScience.
- Whereas search engines can pre-compute much of the information required to respond to search queries, eScience applications will tend to rerun computations starting from raw data more frequently. This will place higher demands on the performance of the machines and the communication networks connecting them.
- Whereas search engines support relatively simple computations by millions of users, eScience systems must support much more demanding computations, but by fewer users. This affects many aspects of system operation, such as how resources are allocated.

NSF (including the Directorate for Computing and Information Science and Engineering (CISE) and the Office of Cyberinfrastructure (OCI), in conjunction with other directorates and offices), the National Institutes of Health (e.g., the National Cancer Institute and the National Institute of Biomedical Imaging and Bioengineering), the Department of Energy's Office of Science and Advanced Research Projects Agency-Energy (ARPA-E), and the Department of Defense particularly stand to benefit from going beyond "big iron" to support large-scale data-intensive computing; they must think about eScience in their key application areas, such as astronomy, chemistry, physics, biology, medicine and healthcare, energy, intelligence, etc.

Indeed, in FY 2008, NSF established a crosscutting program, Cyber-enabled Discovery and Innovation (CDI), to support the very kinds of eScience efforts that we have outlined above – radically new concepts, approaches, and tools at the intersection of the computational and physical or biological worlds. The program seeks ambitious, transformative, multidisciplinary research proposals that bridge the data-to-knowledge chasm. As a Foundation-wide initiative, CDI is able to attract people, resources, and knowledge across institutional, geographical, and cultural boundaries to cultivate discovery, innovation, and learning.

To date, CDI has supported small- and medium-sized research teams that have generated new knowledge from a wealth of digital data. Funded projects have spanned embedded monitoring, data analysis, and eventual control of smart buildings, through creation of data-driven complex models followed by development of associated sensing-communication-computation-control systems; the assimilation of complex geospatial, observational, and experimental data for furthering our understanding of volcanic processes – at a time when volcanoes have repeatedly disrupted the global economy by bringing worldwide air traffic to a halt; and real-time, adaptive imaging for atomic force microscopy, to study the dynamics of oxidation, crystallization, and assembly of inorganic and macromolecular systems; etc.

Recognizing the importance of this new approach to discovery, the scientific community has responded by dramatically over-subscribing the CDI program. But the CDI program has not expanded nearly as rapidly as projected: when CDI began in FY 2008 at a budget level of $53.18 million, the NSF request suggested a "growth of about $50 million per year for a full five years"[7]; however, the President's budget request to Congress for FY 2011 – four years into the program – includes only $105.5 million assigned to CDI, of which only $50 million would be shared by CISE.

It is imperative that CDI be implemented as the major NSF-wide initiative that it was intended to be, that funding be provided at the originally envisioned $250 million level or greater, and that large-scale, longer-duration projects with multidisciplinary teams comprised of multiple senior investigators, graduate students, and senior personnel be added to the portfolio. CDI is the key to advancing data-driven discovery and the tools and techniques that enable it.

Moreover, just as CDI is catalyzing advances in basic science and engineering areas within NSF's portfolio, the Federal funding agencies mentioned above stand to significantly advance their interests by implementing similar programs that are directly aligned to their missions.

For example, there is a wealth of experimental data being generated in biomedicine, and, as described previously, data analytics of digital libraries in this space will help us better understand mechanisms of fundamental chronic diseases like cancer, arthritis, and heart disease. Already, systems biology approaches are facilitating analysis of large-scale protein expression data sets (spanning thousands of proteins by thousands of patients), enabling us to better target drug therapies to specific patients or patient subpopulations and increase patient survival times even for the deadliest of illnesses. Consequently, **NIH – particularly NCI and NIBIB – stands to gain tremendously by investing in a CDI-like program of its own that is specifically targeted to teams of pathologists (and other biomedical researchers) *and* computer scientists**. Ideally, the program would fund projects of similar size and duration as NSF's CDI initiative.

Similarly, machine learning and data mining can readily detect household and commercial energy usage patterns and preferences – and thereby optimize energy delivery and reduce overall energy consumption and costs. For example, researchers have shown that instrumenting a home with just three sensors – for electricity, power, and water – makes it possible to determine the resource usage of *every individual appliance* in the home – simply by analyzing patterns in the data collected by the three sensors. Until now, ARPA-E has invested in materials technology supporting energy storage/transportation. For instance, in October 2009, the agency awarded $151 million for direct solar fuel, novel batteries, etc. However, **ARPA-E must invest a similarly sized chunk of its $400 million budget into groundbreaking eScience insights – forged by teams of energy technologists and information technologists – that are directly targeted toward ARPA-E's mission of developing a smart energy infrastructure for the U.S.** This investment would run parallel to NSF's CDI effort, but it would be specific to the smart grid challenges with which ARPA-E is tasked.

While the CDI program at NSF has funded outstanding research in eScience, this work has naturally been in basic science and engineering areas within the Foundation's portfolio, such as

[7] http://www.nsf.gov/about/budget/fy2008/pdf/39_fy2008.pdf

mathematics, astronomy, physics, chemistry, and so on. However, key questions in areas in which the NIH, DoE, and DoD usually invest remain unanswered. These agencies can certainly look to the commercial world for some of the technology required to support eScience applications corresponding to their missions, but they must also target research efforts that adapt this technology to the unique properties of eScience – and to the unique properties of the work driving their missions. **Ultimately, NSF's CDI program must continue – and it serves as an ideal model for new initiatives at the NIH, DoE, and DoD.**

The road ahead

Data-driven science is reshaping discovery and learning in the 21st century. The combination of rich data sources, new computing technology, and advanced data mining and machine learning algorithms allows scientists to gain much deeper insights into many scientific phenomena than ever before possible. We are just beginning to revolutionize how we study the workings of the Universe; how we diagnose and treat medical ailments; how we generate, price, and deliver energy to homes and businesses; how we gather intelligence and "connect the dots" between multiple sources of information; etc. Fundamental research in computer science has played a critical role in creating the computing technology and the analysis techniques underlying the transformation in discovery and learning embodied by eScience. Continued forward progress is essential to our nation's leadership – and essential to a broad spectrum of federal agencies fulfilling their missions. This progress requires increased investment in the tools and techniques of eScience, and close partnership between disciplinary scientists and computer scientists. America's competitiveness depends upon a vigorous eScience effort.