

4

The Rise of Data-Intensive Research

And generally let this be a rule, that all partitions of knowledge be accepted rather for lines and veins than for sections and separations; and that the continuance and entireness of knowledge be preserved.
—Francis Bacon, *The Advancement of Learning* (1605)

Key Findings

- ◆ The United States government, in particular the National Science Foundation, has been instrumental in the development of the infrastructure required for data-intensive research.
- ◆ These governmental initiatives have influenced the directions of science as it is practiced today—for example, the emphasis on “grand challenges” impacting national issues, collaborative partnerships across institutions, and interdisciplinarity.
- ◆ The 2003 report of the NSF Blue Ribbon Advisory Panel on Cyberinfrastructure provides a current manifestation of a vision for how data-intensive research will transform both science itself and the social organization of how that science is conducted.
- ◆ The contemporary research environment is characterized by several key features:
 - ❖ Computation is a “third branch” of science.
 - ❖ Research is more likely to cross disciplinary boundaries.
 - ❖ Multiple sources of data add a new dimension of complexity to research.
 - ❖ Research is more likely to be framed around specific problems and applications.
 - ❖ Every discipline is becoming computationally intense.

This chapter provides a context for this ECAR study by reviewing recent political, economic, and intellectual developments that have shaped the contemporary research environment. The past half century has seen remarkable innovations in how IT has both enabled research and generated new research approaches (National Research Board, 2005). Computational science¹ has come into its own as an approach distinct from computer science, transforming those who work with IT from providers of services into full collab-

orative peers with researchers (Yood, 2005). The social sciences and humanities (American Council of Learned Societies, 2005) have joined the physical and life sciences in the development of computationally intense research methods.

As happened with science itself a few decades earlier (de Solla Price, 1963), the value of computational sciences to national security, economic strength, and U.S. competitiveness in science and technology has been established. The government has therefore invested

in waves of support for academic computing, moving from the creation of campus-level computing centers to national supercomputing centers and now to grid computing (Colwell, 2003). Comparable developments have occurred in networking. Through funding, laws, and regulations, both federal and state governments continue to play a significant role in the conduct of research (Bohlin, 2004; Rogers, 1998). These multiple forces have molded research ideas, practices, and institutions into forms that are quite different from those that dominated the history of modern science (Gibbons et al., 1994).

The National Agenda for Research and IT

In the United States, the national agenda for research and IT developed out of a long appreciation for the value to society of scientific and technological innovation. The inclusion of intellectual property rights in the U.S. Constitution was one manifestation of this insight, and by the early 19th century government units such as the Post Office were engaged in scientific activity. The role of colleges and universities in the creation of knowledge was institutionalized by Congress between 1862 and 1914 through the Morrill, Hatch, and Smith-Lever Acts. University faculty were heavily involved with governmental activity during and after World War I (Rudy, 1991), but it was primarily after World War II and the launch of the National Science Foundation (NSF) that the groundwork for today's developments was laid.

Following its creation in 1950, the NSF initially supported the development of campus computing centers, thereby bringing this new form of technology to the attention of a wider academic audience. But when the NSF retreated from this support in the early 1970s, only those researchers who worked with agencies such as the Department of Energy and NASA had access to supercomputers. This limited the scope of

research for most faculty to what they could do on departmental minicomputers. Researchers in experimental computer science and computation voiced their concerns through the Feldman Report (Feldman & Sutherland, 1979) about the scholarly—and national—consequences of these restrictions. This was followed shortly by the Press Report (Press, 1981) from the physics community and then by the Lax Report (Lax, 1982), which extended the argument across the physical and biological sciences. Taken together, this suite of reports argued for the revitalization of experimental computer science to generate new types of computer systems that could solve hitherto intractable problems; highlighted the importance of computation for the intellectual, economic, and military strength of the United States; and pointed to the need for more supercomputers and more access to them by academic-based researchers.

The government was equally instrumental in the development of networks. The Department of Defense's ARPANET and the first Interface Messaging Protocol were developed at the University of California, Los Angeles (UCLA), in 1969. ARPANET's first four nodes were at universities: UCLA; the University of California, Santa Barbara; Stanford University; and the University of Utah. Less than five years later, the Transmission Control Protocol/Internet Protocol (TCP/IP) was developed to interconnect devices on the network, and by 1983 it was connecting every device on the ARPANET. By 1988, the NSF had created a backbone network of T1 (1.544 Mbps) speeds that connected 50,000 host computers. By 1990, the NSF was installing the first T3 (45 Mbps) networks, and the number of hosts had mushroomed to more than 300,000.

Political support for these goals received a tremendous stimulus when the Japanese government launched its \$500-million "Fifth Generation Computer Project" in 1981. Intended to leapfrog existing technologies, this move catalyzed both the U.S. government and the

U.S. scientific community. Federal agencies such as the NSF, the National Security Agency (NSA), and the Department of Energy began to believe that supporting academic access to supercomputing would serve three goals: it would help protect the domestic supercomputing industry in the face of Japanese challenges; it would provide increased access to these machines for computer and computational science research; and it would provide experimental facilities needed to attract the best computer scientists to academic work, where they would also train the next generation of computer scientists. Momentum continued to build. A growing number of scientists from across the disciplinary spectrum began to refer to themselves as “computational” scientists—that is, as biologists or chemists or physicists who primarily used computational techniques rather than experiment, theory, or observation in their research—because they realized that computer simulations and other new types of data interactions could provide new theoretical insights into biological and physical phenomena.

In 1987, the U.S. Office of Science and Technology Policy (OSTP) issued a report that advocated the use of networked computing to support research. After additional studies and congressional hearings, Congress established a \$4.7-billion High Performance Computing and Communications Program (HPCC) in 1993. Five elements of the HPCC were instrumental in shaping our current thinking about computation-intensive research. The first was recognition of the importance of high-performance computing and communications to national security and to economic competitiveness. Second, funding projects were framed within the context of scientific and engineering “grand challenges.” Third, HPCC mandated that research using high-performance computers be collaborative partnerships among entities from academia, the private sector, and government. Fourth,

research projects within the sciences were encouraged to be interdisciplinary. And finally, computational scientists were recognized as full collaborators with researchers rather than as just service providers.

The NSF operationalized the concept of grand challenges by establishing funding programs that joined disciplinary researchers, computer scientists, and emerging information technologies to solve fundamental science and engineering problems with broad economic and scientific impact. The NSF Blue Ribbon Panel on High-Performance Computing (1993) differentiated these projects from others funded by the NSF through the goal of accelerating progress in virtually every branch of science and engineering concurrently rather than by discipline, and by highlighting the intention to stimulate the U.S. economy as a whole. Admitting that this augured a transformation of science and engineering, the panel made specific recommendations that sought to remove technological and implementation barriers to the rapid evolution of high-performance computing, to build scalable access to a pyramid of computing resources, to democratize the base of participation in high-performance computing, and to develop future intellectual and management leadership for high-performance computing. Accompanying the panel’s scientific and technical conclusions was a discussion of political implications. For policymakers, these emerging “critical technologies” required changes in legal and regulatory support, such as exemptions from antitrust law and changes in the structure of taxes (Branscomb, 1993).

When the Blue Ribbon Panel issued its report in 1993, it was assumed that the United States would continue its dominance in science and technology. Only a decade later, however, this was no longer clear. The September 11 terrorist attacks had caused some rethinking of the technologies critical for survival, and successes in analysis of the genome

had dramatized the nature and utility of high-performance computation. Accordingly, the NSF reconsidered its position on supercomputing centers and, in 2003, refocused its strategy in the physical and biological sciences around the concept of “cyberinfrastructure.” As we will see shortly, this concept incorporated a number of features, including grid computing, the harvesting of computing cycles, massive storage, and the long-term management of data in both raw and analyzed forms. Such matters as middleware, top-level administrative leadership, and the development of new analytical tools and research methods were all identified as key to the knowledge production infrastructure. The agency called for increasing the speed and capacity of computers another 100 to 1,000 times to serve current research needs.

A second NSF report (Berman & Brady, 2005) focused on the particular cyberinfrastructure needs of the social sciences, arguing that progress in this area is important both for research on society and for what social scientists can teach us about human and computer interactions. Recommendations included an emphasis on funding and institutional support for training social scientists to apply advanced computational techniques and to engage in collaborations that take advantage of cyberinfrastructure capacities, since both of these require innovations in research practices and institutional structures. Attention was also drawn to policy issues such as security, confidentiality, and privacy.

An early humanities project to use digital infrastructure—Project Gutenberg, which offers free, full-text, digital versions of important books—was launched in 1971. Since then, there has been a growing awareness of the great effort required to develop computational environments, software, and instrumentation. In 2005, the American Council on Learned Societies joined with other groups to identify computational needs in the humani-

ties. The ACLS emphasized the humanities’ important contributions to the cultural record and noted that data sources, artifacts, and cultural objects were currently fragmented across institutional boundaries, owned by institutions and individuals, and subject to cultural and other barriers. The need was seen as especially acute as researchers in the humanities increasingly collaborated with researchers in the physical and life sciences in such areas as document and image analysis. Further, humanities databases can equal or even exceed those of the physical and biological sciences. As an example, the Sloan Digital Sky Survey of a quarter of the entire sky, using X-ray, infrared, and visible-light images of more than 100 million celestial objects, produced a data set of more than 40 terabytes of information. In addition, the compilation of video interviews of Holocaust survivors by the Survivors of the Shoah Project produced 180 terabytes of data.

The National Research Board (2005), which governs the National Science Foundation, produced a report on long-lived data needs that built on work by the Library of Congress and other federal agencies. The board noted several features of today’s data environment that have radically altered the ancient problem of information storage: the ephemeral nature of digital storage media, the constancy and rate of technological innovation, the desire to revisit old data through new analytical lenses, and the growing complexity and time-sensitivity of data curation. The data collection universe now includes data collections, related software, and hardware and communication links, all shared by a universe of data authors, managers, users, data scientists, research centers, and other institutions. The board distinguished three types of data storage entities: *focused research collections*, which support a limited group of researchers on a single research project;

intermediate-level resource collections, which derive from a specific facility or center that spans research projects and teams; and *reference collections*, which have global users and impact, and which incorporate data of multiple types across disciplinary, institutional, geopolitical, research project, and data type boundaries.

The President's Information Technology Advisory Committee (PITAC, 2005) issued a report in June 2005 reemphasizing the importance of computational science for U.S. competitiveness generally and for such specific issues as nuclear fusion, the folding of proteins, and the global spread of disease. In addition to strengthening computational and long-term data storage capabilities and supporting interinstitutional relationships needed to carry out large-scale distributed research, PITAC recommended the creation of national software sustainability centers to ensure that archived data would remain accessible regardless of innovations in hardware and software. The committee warned that without new initiatives, U.S. leadership in computing and computationally intense science would continue to deteriorate. This long-standing White House committee was dissolved by President Bush shortly after issuing this report.

The Mathematical Association of America issued a report in 2005 that exemplifies how disciplinary associations are thinking about the changes taking place. For most of the 20th century, mathematics was linked most closely with physics and engineering, but today biology is seen as the stimulus for innovation. This report noted that evidence is now as often mathematical as observational, and remarked on the shift in the relative prestige of various disciplines as they become more computationally intense.

National policies have remained critically important to the sustenance of academic research (Croissant, Rhoads, & Slaughter,

2001), although unfortunately appreciation of the knowledge economy does not always translate into financial support to colleges and universities as the sites of knowledge production and distribution (Bleiklie & Burkjeftot, 2002). The number of institutional players involved, however, continues to grow; the National Institutes of Health (NIH), the largest single source of funds for university research today, is currently "turbocharging" its efforts in biomedical computing (Brainard, 2003).

A Vision of Cyberinfrastructure

As a frame for our study of the engagement of institutional IT in academic research, we will highlight one formulation of cyberinfrastructure: the 2003 report of the NSF Blue Ribbon Advisory Panel on Cyberinfrastructure, *Revolutionizing Science and Engineering Through Cyberinfrastructure*, also known more prosaically as the Atkins Report. There are other conceptualizations, of course, but the Atkins Report nicely summarizes the range and complexity of opportunities facing scientists. The fact that the NSF named Daniel E. Atkins to head its newly created Office of Cyberinfrastructure in February 2006 also gives the report added prominence.

So what is cyberinfrastructure? The report offers multiple iterations of a definition. Positing it to lie between the "electro-optical components of computation, storage, and communication" and the programs, information, and practices of disciplines and research communities, the cyberinfrastructure layer consists of "enabling hardware, algorithms, software, communications, institutions and personnel" (National Science Foundation, 2003, p. 5). Elsewhere, cyberinfrastructure is portrayed as

grids of computational centers, some with computing power second to none; comprehensive libraries of digital

objects including programs and literature; multidisciplinary, well-curated federated collections of scientific data; thousands of online instruments and vast sensor arrays; convenient software toolkits for resource discovery, modeling, and interactive visualization; and the ability to collaborate with physically distributed teams of people using all of these capabilities. This vision requires enduring institutions with highly competent professionals to create and procure robust software, leading-edge hardware, specialized instruments, knowledge management facilities, and appropriate training. (NSF, 2003, p. 7)

Cyberinfrastructure extends many of the trends under way since World War II, but it is distinctive in its anticipation of new scientific environments that “enable teams to share and collaborate over time and over geographic, organizational, and disciplinary distance. They enable individuals working alone to have access to more and better information and facilities for discovery and learning. They can serve individuals, teams, and organizations in ways that revolutionize *what they can do, how they do it, and who participates*” (NSF, 2003, p. 13).

Two elements of this envisioned research environment deserve special mention. The first is the expectation—based on experience to date—that cyberinfrastructure could transform science itself. Through small vignettes, the report illustrates “how scientific and engineering research will be revolutionized and the benefits that will flow from those changes” in a host of intellectual areas: atmospheric science, forestry, ocean science, environmental science and engineering, space weather, computer science and engineering, information science and digital libraries, biology/bioinformatics, medicine, physics, astronomy, engineering,

materials science and engineering, and the social and behavioral sciences (NSF, 2003, pp. 18–23). This is not about whiz-bang technology, but rather about fundamentally new directions in scientific pursuit.

Second, the cyberinfrastructure vision is consciously boundary-breaking, with coordination across education, industry, and government; across disciplines; and across national boundaries. It seeks to remove the obstacles to participation by minority-serving institutions and the physically challenged:

We have the opportunity to extend networked systems to provide comprehensive and increasingly *seamless* functional services for research and learning—to create virtual laboratories, research organizations, indeed technology-enabled research environments that offer a full spectrum of activities in the process of scientific discovery and the education of the next generation. We are at the threshold where a *collaboratory* or *grid community* can become “the place” where a research community interacts with colleagues, data, literature, and observational systems together with very powerful computational models and services. (NSF, 2003, pp. 44–45)

Much of the rest of the Atkins Report addresses the organizational and budgetary requirements within NSF that are necessary to realize the Advanced Cyberinfrastructure Program (ACP). But for our purposes, we need merely conclude that proponents of cyberinfrastructure envision a social and intellectual revolution as profound as the technological innovations that will drive it. CIOs and other campus-based IT professionals who view the future of IT in research merely in terms of network speed or computing cycles are missing the boat.

Features of the Contemporary Research Environment

At a very high level of generalization, scientific inquiry over the past millennium has progressed from

- ◆ individuals working alone, thinking, to
- ◆ individuals in face-to-face conversation with others, to
- ◆ individuals learning from others across space and time through print, to
- ◆ individuals systematizing written records into a corpus of knowledge, to
- ◆ individuals working in laboratories purposefully to extend the frontiers of knowledge, to
- ◆ teams working in laboratories based on disciplinary and institutional boundaries, to
- ◆ teams working collaboratively across laboratory, institutional, and national boundaries, to
- ◆ teams developing entirely new realms of research through high-performance computing.

Today, as *New York Times* science writer George Johnson (2001) expressed it, "All science is computer science." So what are the features of the emerging research environment, and what are the implications for IT organizations? We can distinguish six characteristics.

Every discipline is becoming computationally intense. The first discipline to become computationally intense was physics, followed by chemistry and then biology. Psychology made the computational turn in the 1990s (Austin, Scherbaum, & Mahlman, 2002). Today, all disciplines have become computational. As examples: The term *mechatronics* is used to describe the integration of computers and mechanical engineering (Bollag, 2005). In the field of English, a database of every piece of scholarship about *Hamlet*, linked to every line of the play, is used as the foundation for many of the 500 books and articles still writ-

ten on the play each year. Software written to analyze the human genome is now being used to determine which versions of Chaucer's texts are the earliest (Young, 2005). Musicians are being fully wired inside and out so that researchers can study what happens with their bodies as they play and sing (Mangan, 2004). Ethnographers are using computers to compare the results from studies conducted at diverse sites. The emphasis on digital humanities at the University of Virginia has demonstrated that there are very few academic pursuits that can escape the benefits of computation (Blustain & Spicer, 2005).

Computation is the "third branch" of science, along with theory and experimentation. For modern science as it developed over the last several hundred years, knowledge was produced through iterative interactions of theory and experimentation. Theorists, according to Kenneth Wilson (1989), focus on relationships among experimental quantities, the principles that underlie these relationships, and the mathematical concepts and techniques needed to apply the principles to specific cases. Experimentalists design and use scientific instruments to make measurements, undertake controlled and reproducible experiments, and analyze both the results of those experiments and errors that appear within them. Today, computation is a third fundamental element of knowledge production. Computational scientists use algorithms to solve scientific problems and their operationalization via software, to design computational experiments, and to identify underlying mathematical frameworks, laws, and models. Conceived in this way, computation thus changes the role of IT professionals from service providers to collaborators whose input is critical to the success of the research. Through their understanding of technology, they contribute substantive knowledge to the advancement of the methodology and the science.

Large research projects are more likely to transcend disciplinary boundaries. The expectation that computationally intense research projects would be interdisciplinary in nature has been inherent in the design of the Internet since the 1960s (Abbate, 1999) and has been enunciated as a basic principle for all IT-intensive research since the appearance of the “grand challenges.” The notion of computing and communications as infrastructure, articulated in the Bardon Report (Bardon & Curtis, 1983), has been helpful in this regard, reorienting attention from disciplinary boundaries to the problems themselves. An important corollary of this approach is the heterogeneity of research teams. The implication is that IT professionals who have traditionally oriented their services and resources toward departmental needs (for example, how do I work with the chemistry department?) will need to think about supporting variable teams of faculty built around specific research questions.

Large research projects are more likely to involve data from multiple studies at multiple sites across multiple time periods. The notion of a single research study generating and analyzing its own data in a vacuum is dying. High-performance computing has made it possible to analyze much larger aggregates of data, expanding methodologies and data sets across space and time in ways previously not possible. Researchers working with such data sets need support to develop software capable of analyzing patterns across types of data. They also require institutional arrangements that ensure their access to that data irrespective of where it resides or in what form. Again, this requires IT professionals to be consultants as much as service providers.

The distinction between “basic” and “applied” research has fallen away. Although this distinction, introduced by Vannevar Bush in 1945, is still common in public discourse, some sociologists of knowledge reject that distinc-

tion as one-dimensional, to be replaced with a more complex array of linkages among various kinds of scientific activities. Similarly, the distinction between science and technology is fading, since IT is simultaneously the subject and the tool of inquiry. Technological innovation can develop out of existing technologies without additional fundamental research, and the use of technologies can itself generate new basic knowledge. For aggressive IT units, these realities can provide additional motivations for engaging in problems such as the development or adaptation of software to serve research needs, because doing so might result in patents and other research successes of their own.

Research is more likely to be framed around specific problems and applications. The promulgation of “grand challenges” started the process of shifting research funding away from theory-driven research and toward specific social, environmental, and other problems. A prime example is the emphasis on national and homeland security, which has emphasized the shorter-term, utilitarian salience of research. As a result, researchers are increasingly required to respond and be accountable to social and political demands rather than being able to isolate themselves within an “ivory tower.” Academic research, in other words, has become less, well, academic. This trend affects IT professionals because they too, like the researchers they support, are more likely to be drawn into pressing social issues that have a multitude of policy concerns.

The Current Environment

The research community and the IT professionals who support it now find themselves at an interesting juncture in their co-evolving history. The past five decades have seen a configuration of trends that include data-intensive computation as a core element of research, growing complexity and costs, the intensification of research as a stamp of institutional

distinction, and the need for sufficient money to fuel the system. In FY 2004, higher education institutions reported R&D expenditures of \$42.9 billion, a 7.2 percent increase over 2003 (NSF, 2006).

In 2004, federally financed research and development grew by 10.7 percent, the third straight year of double-digit growth. Federal funding that year was \$27.4 billion and accounted for 64 percent of total academic R&D support for science and engineering (NSF, 2006). The increase in research funding has slowed. Much of the focus has been on the declining growth of the NIH budget, which for 2006 exhibited the smallest growth in 36 years. At the same time, the NSF is limiting the number of applications it will accept from any one institution, creating and intensifying competition within large research universities as well. Further, an increasing proportion of federal research R&D money is going into development rather than general scientific research, a trend that favors corporate over academic research (Bender, 2005). However, the news from Washington is not all bleak. Under President Bush's 2007 budget request, the NSF would see a 7.9 percent increase over the current 2006 fiscal year, with three-quarters of the proposed increase funding 500 new grants (Field, 2006). Still, universities that have linked their missions and prestige with research are eyeing the federal environment with unease.

Many universities are seeking to increase their support for research from industry. But in FY 2004, industrial support for science and engineering declined for the third consecutive year, to \$2.1 billion. Industry's share of academic R&D support was 4.9 percent, the same as in 1983 (NSF, 2006). Any hoped-for growth from industry is likely to come with stiff resistance from other institutions. More institutions are seeking to grow their research capacity, creating an even keener competitive environment.

Universities themselves have invested heavily in research, but there are signs that funding is slowing here as well. In 2002 and 2003, institutions spent \$7.6 million to construct 16 million square feet of academic research space, the most since 1988. Over half of this space was in the biomedical field, and it was financed with a declining percentage of federal funding (Brainard, 2006). But in FY 2004, institutional funding increased by only 1.5 percent, to \$7.8 billion, a leveling off after average annual increases of 9 percent over the previous seven years (NSF, 2006). At research universities, the average annual job growth for PhD-level scientists and engineers for the period from 1993 to 2003 was 1.2 percent, compared with 2.3 percent job growth over the previous decade (Brainard, 2006).

Summary

The past 60 years have seen a remarkable progression both in the technologies that support research and in the nature of research itself. The integration of those technologies has been shaped to a large extent by a series of federal programs directing resources toward greater integration across the research enterprise. As a result, computation has had a profoundly liberating effect on research by breaking the shackles of time, distance, discipline, and the need for physical experimentation.

At the same time, all signs indicate that research funding is becoming tighter and more competitive. As always during times of retrenchment, people will seek ways to do more with less. Given the importance of data-intensive research and the ways in which it has become indispensable to the conduct of science, researchers and the IT units that support them will be more accepting of approaches that leverage scarce resources.

It is within the context of this environment that we now proceed to explore IT-intensive research as it is lived in our institutions of higher education. We begin with how the research landscape looks from the central IT leader perspective.

Endnote

1. The term *computational science* arose in the 1970s to refer to the use of computers to do science; that is, the use of an algorithm to solve scientific problems (see <<http://www.shodor.org/refdesk/Help/whatiscs.html>>). The phrase distinguishes computational science from science that depends on experimentation, observation, or theoretical reliance on mathematical logic to derive results.