Making Research and Education Cyberinfrastructure Real

Francine Berman

Francine Berman is Director of the San Diego Supercomputer Center and is Professor and High Performance Computing Endowed Chair in the Department of Computer Science and Engineering at the University of California, San Diego.

Comments on this article can be sent to the author at <berman@sdsc.edu> and/or can be posted to the web via the link at the bottom of this page.

The *Industrial Age* transformed the world through the application of technology to research, practice, and everyday life. Technology revolutionized manufacturing, transportation, agriculture, and communications and created deep social and economic changes that continue today. In the last century, the explosion of information technologies ushered in an *Information Age* with similar transformative potential. In 2008, it is hard to imagine modern life and work without the ability to access, manipulate, organize, and understand a sea of digital information on almost every conceivable topic.

The driving engine for the Information Age is *cyberinfrastructure (CI):* the organized aggregate of information technologies (computers, storage, data, networks, scientific instruments) that can be coordinated to address problems in science and society. Fundamental to modern research, education, work, and life, CI has the potential to overcome the barriers of geography, time, and individual capability to create new paradigms and approaches, to catalyze invention, innovation,[1] and discovery, and to deepen our understanding of the world around us.

However, CI in the academic sector often falls short of its remarkable potential. Comprised of dynamically evolving information technologies, CI is both a continuous work-in-progress and a stable infrastructure driver for invention and innovation. It is this duality, as well as the challenge of creating an

environment that effectively supports both the invention and the broad use of CI for research and education, that is the focus of this article.

Cyberinfrastructure as a National Initiative

The "parents" of CI, as a U.S. national research and education initiative, are arguably Ruzena Bajcsy, who served as assistant director of the Computer and Information Science and Engineering (CISE) Directorate at the National Science Foundation (NSF) from 1998 to 2001, and Dan Atkins, who served as the founding director of NSF's Office of Cyberinfrastructure (OCI). During her tenure at NSF, Dr. Bajcsy convened a Blue Ribbon Panel to study the emergence and importance of CI, a panel chaired by Dr. Atkins. The report from the Blue Ribbon Panel still stands as a fundamental and compelling document on the promise of CI in twenty-first-century research and education.[2]

Since the publication of the Blue Ribbon Panel report in January 2003, CI has become a major priority for NSF (as described in NSF's *Strategic Plan FY 2006–2011*)[3] and also a priority for virtually every research funding agency in the United States. The importance of research and education CI is strongly underscored by the *American Competitiveness Initiative*, by the August 2007 report from the President's Council of Advisors on Science and Technology, *Leadership Under Challenge: Information Technology R&D in a Competitive World,* and by many other assessments of the state of U.S. research and education.[4] The U.S. Department of Energy (DOE), the National Institutes of Health (NIH), the Library of Congress, the National Archives and Records Administration, the National Endowment for the Humanities, and other federal agencies have prioritized CI for targeted initiatives. Clearly, as a core component of the research and education landscape, CI is here to stay.

In addition to federal agencies, universities and colleges in the United States, and indeed throughout the world, are creating CI initiatives to increase participation in, and competitiveness for, national efforts. We focus herein on CI as a national U.S. research and education initiative, leaving for others the discussion of related efforts on university and college campuses to develop research CI to support local faculty and students, international efforts in CI, and efforts within the private sector to use CI products and facilities developed for commercial use to also support research and education.

The Promise of Cyberinfrastructure for Research and Education

At its best, CI has greatly expanded the arsenal of tools and approaches for twenty-first-century academics. The following examples illustrate how key

questions from distinct academic domains are being addressed through the creative use of community CI.

Can the Progression of Parkinson's Disease Be Stopped?

Every nine minutes, an individual is diagnosed with Parkinson's, a devastating disease characterized by a decrease in limb mobility over time. The search for new drug therapies that could halt the progression of Parkinson's is an active area of research. Modern efforts to understand the behavior of Parkinson's involve molecular modeling, molecular dynamics simulations, and biochemical analysis and rely on integrated computational and data-analysis CI to support simulations of disease progression and to vet drug therapies at sufficient scale.

As noted in a recent article, familial studies suggest that the progression of Parkinson's is associated with "defects that cause increased aggregation of a protein known as alpha-synuclein, . . . which, in turn, leads to harmful ring-like or pore-like structures in human membranes."[5] Igor Tsigelny, Eliezer Masliah, and their collaborators used simulation studies to investigate molecules that block the propagation of alpha-synucleins into more harmful structures, providing a model for a new type of therapeutic approach. Their approach holds promise for retarding the progression of the disease and has applicability beyond Parkinson's to other diseases within the same family (e.g., Alzheimer's, rheumatoid arthritis, type 2 diabetes mellitus).

Enabling Cyberinfrastructure

Tsigelny, Masliah, and their collaborators' work involves modeling the behavior of alpha-synuclein by generating hypotheses about its structural tendency to aggregate and undergo pore formation and insertion into biological membranes. Calibrations of the disease simulation models with wet lab studies serve to vet and improve the accuracy of the computational models, making them effective tools for investigating drug therapies.[6]

The CI that enabled this breakthrough is representative of what is required for many applications in computational science. Tsigelny, Masliah, and their team used parallel versions of community codes (NAMD, DOT, and MAPAS) to develop a computational model that could simulate alpha-synuclein behavior at scale. Their application was run on resources available through NSF's TeraGrid and at IBM's research facility. Next-generation investigations by this team will involve greater resolution of the computational model, necessitating increased software scalability, more powerful machines, and/or longer run-times. This will require next-generation CI resources that are both more capable and of higher capacity than their current CI environment.

How Can Data from Field Instruments and Sensors Be Efficiently Delivered in Real Time?

The ability to access remote data from field instruments and sensors in real time is revolutionizing a broad set of disciplines including ecology, astronomy, environmental science, and biology. Moreover, the CI used to link such instruments, sensors, and sites in the field can be used more broadly for additional applications, such as distance learning or support for first-responders during environmental disasters.

The High Performance Wireless Research and Education Network (http://hpwren.ucsd.edu/), developed by Hans-Werner Braun, Frank Vernon, and collaborators, is a high-speed wide-area wireless network that links educational institutions (e.g., UCSD, San Diego State), field instruments (e.g., the Mt. Palomar telescope, Mt. Laguna Observatory, environmental sensors), and "hard to reach" areas in San Diego County (e.g., the Pala Native American Reservation, the California Wolf Center). Applications enabled by HPWREN have included the following:

- *The discovery by Caltech's Palomar Observatory astronomers of an object larger than Pluto.* The discovery eventually resulted in the "demotion" of Pluto as a planet. HPWREN supported real-time transmission of astronomical image data from the telescope to various institutions.
- *The ability to operate a variety of field equipment remotely and to transmit data and video streams from field sites in real time.* For example, HPWREN supports remote observation of wolf behavior at the California Wolf Center by biology researchers at the University of San Diego, UCSD, and San Diego State University.
- *Distance learning, tutoring, and Internet classes for participants at the Learning Center on the Pala Native American Reservation in East San Diego and UCSD.* HPWREN-initiated connectivity of Pala led to the development of networking expertise within the Native American community and ultimately to the creation of the Tribal Digital Village Network linking reservations within San Diego County.
- *Support for public response to major wildfires in the backcountry of San Diego County.* The California Department of Forestry and Fire Protection (Cal Fire) uses HPWREN cameras and data connectivity during major fires to support remote fire stations and enhance incident management.

Enabling Cyberinfrastructure

To enable the preceding and many other applications, HPWREN wireless networking CI has been designed and developed to support scientific and environmental monitoring, real-time sensor data collection, and the ability to

process, manage, and transmit data at a variety of scales. HPWREN instruments and sensors need to be operational in a challenging physical environment that is open to the elements and wildlife, and HPWREN network systems need to adapt to irregular data-transmission patterns, power and battery constraints, and emergency situations.

Moreover, HPWREN is accelerating the development of cost-effective, "green" CI. For example, recent CI research by Tajana Simunic Rosing and her students used HPWREN as a test bed to focus on techniques for maximizing battery lifetime and throughput in sensor network environments.[7] As power costs and requirements threaten to escalate out of control for modern campuses, such research is key to developing solutions that benefit both the environment and campus budgets.

Family vs. Neighborhood:
Which Has a Greater Effect on Educational Attainment?

The impact of the surrounding environment on the human condition is an important focus for social analysts, as well as for every parent who has ever considered a move to a better neighborhood to provide "greater opportunities for the kids." This was the subject of a *Review of Economics and Statistics* article that investigated the question: Which matters more in educational attainment: family or neighborhood?[8]

The authors of the article based their analysis on the Panel Study of Income Dynamics (PSID), a longitudinal data collection that has tracked information on nearly 70,000 individuals in thousands of families for over four decades. The authors measured years of education for a 1968 sample as reported in PSID interviews from 1985. Two families were defined as "neighbors" if they had matching "cluster" identifiers within the 1968 sample or had the same geocode (obtained from census tract identifiers) based on their 1969 addresses. Using a thorough statistical analysis, the researchers concluded: "Sibling correlation in years of education [is] more than .5. In comparison . . . the correlation between neighboring children [is] less than .2. . . . Sibling resemblance in educational attainment arises mostly from growing up in the same family rather than in the same neighborhood." In other words, within study parameters, the results provide scientific evidence that family has a greater impact on educational attainment than does neighborhood.

Enabling Cyberinfrastructure

This result is one of nearly 2,600 publications based on PSID data. Updated, preserved, and accessible to the community for research, the PSID is hosted as community CI by the Inter-university Consortium for Political and Social Research (http://www.icpsr.umich.edu/) and is managed by the University of

5

Michigan's Survey Research Center (http://www.src.isr.umich.edu/). Although small (less than a gigabyte of data) compared with many scientific data sets, the PSID is tremendously important to the social science community. For example, in 2007, there were about 23,000 data-extract downloads from 6,000 distinct IP addresses.

Maintaining, updating, and providing access to the PSID requires substantial human, software, and hardware infrastructure. Survey Research Center staff have collected the PSID data annually from 1968 to 1997 and biennially since 1997. In addition, they manage and preserve the PSID archive and clean, process, and disseminate the data (available at www.psidonline.org), as well as provide customized output files and codebooks for community researchers. Like the Protein Data Bank (http://www.rcsb.org/pdb/home/home.do) in the life sciences and the National Virtual Observatory collection (http://www.us-vo.org/) in astronomy, the PSID serves as a fundamental CI driver for a broad community of domain researchers and educators.

The Challenges for Cyberinfrastructure in Research and Education

Cyberinfrastructure provides an evolving foundation for twenty-first-century research and education. As such, it presents two faces: CI as a *focus for invention* and CI as an *accelerator of innovation*. These two faces of CI are linked through a trajectory that begins with invention and design and evolves to broad-based use. CI support, participant roles and responsibilities, and evaluation goals and means vary along this trajectory, and it is critical to align the CI efforts at each stage with appropriate support models and evaluation measures.

The Cyberinfrastructure Trajectory
Broad-use CI is the result of a progression that begins at conceptualization with research, design, and initial development (*CI as a research target*), evolves to further development and prototyping (*engineering CI for use*), and further evolves to become a robust and sustainable infrastructure usable by a broad constituency (*CI as an accelerator of research and education*). This progression—the "CI trajectory"—is illustrated in Figure 1.

Figure 1. The CI Trajectory

CI research, design, and initial development → Engineering and prototyping of CI → Broad-use CI

The following sections describe each stage of the CI trajectory, appropriate funding models, and support challenges and opportunities.

Stage 1: CI as a Research Target

CI begins with research, and CI inventions and innovations are substantively represented in the computer science, engineering, and domain literature. An excellent example of CI innovation is the project that won the "Best Research Paper" Award at the International Conference for High Performance Computing, Networking, Storage, and Analysis, 2006 (SC06), the annual supercomputing conference. Addressing the problem of simulating biochemical events on longtime scales, a team from D. E. Shaw Research focused on the development of new molecular dynamics algorithms and implementation techniques to reduce turnaround time (time between submission and completion of a computer program) on commodity clusters, a ubiquitous platform for research computing.[9] The team's *Desmond* software combined innovations in domain science (the development of new molecular dynamics algorithms with features important for research in chemistry and molecular biology) and applied computer science (novel parallelization methods for particle interactions that reduce interprocessor communication volume, new communication primitives that outperform MPI for *Desmond*'s dominant interprocessor communication patterns, and other features to promote efficiency on commodity clusters). Such efforts provide the key innovations that can ultimately evolve to broad-use community software.

Support for CI Research

Support for CI research is the focus of an increasing number of federal programs. NSF programs supporting CI research have been initiated in all NSF directorates, and the DOE, the NIH, and other agencies have active CI research programs. The key measures of success for CI research, design, and initial development are *invention* and *innovation*. Traditional measures of invention and innovation success within the academic community (e.g., publications and citations, community prizes and awards, successful demonstrations, peer approval) are relevant measures for CI research and are broadly used by research funding agencies, promotion and tenure committees, and review committees to evaluate the quality of CI research.

Stage 2: Engineering CI for Use

To progress along the CI trajectory from invention to use, promising CI research ideas must be vetted, developed, and engineered into usable prototypes. The process of engineering CI for use is critical to promoting the robustness and applicability of CI. Whereas research demonstrations are often usable primarily by tolerant "friends and family," the further engineering of software, interfaces,

and other CI components is critical to support a larger constituency. During this stage, usability is prioritized over innovation.

For example, consider the management of national production grid systems such as TeraGrid (http://www.teragrid.org), which comprises 11 sites, 19 resources, and over 900 different software package installations and services. The well-being of each TeraGrid component and its coordinating software is critical for operation in production mode. Measuring grid "health"—ensuring that all relevant components are operational and that software is up-to-date—is critical for grid stability and for effective troubleshooting.

Inca, a grid-monitoring tool developed by Shava Smallen and her colleagues, displays the state of grid "health" for TeraGrid. (Inca is also used more broadly in other national and international grid projects.) To do this, Inca deploys a set of scripts to test resources and report their status information. Inca server components then manage, integrate, and display resource data via a web interface to illustrate grid operation, functionality, and "health."[10] CI such as Inca requires substantial engineering to support real use. Inca must enable efficient automated testing of component grid software installations and services and must adapt to new resources and installations as the grid software and hardware landscape evolves. Smallen and her team provide regular software updates for Inca and have created substantial documentation and training materials to support the community. Sustaining Inca as broad-use CI will require the development of a long-range support model for staffing and continued software development in order to target Inca to next-generation grid component resources.

Support for CI Prototyping

As is the case with Inca, engineering CI for use often involves professional staff. Staff expertise and experience with increasingly complex academic and commercial software tools and systems is invaluable to creating usable CI. Such staff members are typically supported on "soft money" and are often undervalued (and comparatively underpaid) in the university research setting.

In contrast, the ability to engineer CI for use is both recognized and greatly valued within the private sector. The difficulty of pursuing and funding efforts beyond the initial innovation stage in academia has caused somewhat of a "brain drain" in recent years as entrepreneurial faculty with a desire to follow their ideas from concept to product (and to work with a team of professional staff who can help accomplish their vision) are leaving colleges and universities for the private sector rather than finding a place within the academic framework to pursue promising efforts beyond the innovation stage.

Obtaining federal research funding for CI prototyping and engineering is a challenge in the academic sector. In keeping with the federal funding agencies'

missions of research and education, most (but not all) research funding programs focus on invention, innovation, and new starts rather than on the improvement of existing demonstration efforts. (Programs that focus on further development and prototyping of research efforts include NSF's Software Development for Cyberinfrastructure [SDCI] and the NIH's Continued Development and Maintenance of Software, PAR-05-057, within the National Institute of General Medical Sciences, for example.) Compared with the support available for CI invention and innovation, the lack of funding, the scarcity of workforce opportunities, and the lack of recognition for successful efforts make it difficult to create a healthy pipeline along the CI trajectory from invention and design to CI development and prototyping within the academic sector.

Stage 3: CI as an Accelerator of Research and Education

In the third stage of the CI trajectory, CI has been engineered and targeted to be both used by and useful to a broad community. Examples of such infrastructure supported by federal agencies include NSF's TeraGrid, the DOE and NSF's Open Science Grid (http://www.opensciencegrid.org), the multi-agency-supported Protein Data Bank, and the University of Wisconsin's Condor (http://www.cs.wisc.edu/condor/).

Whereas CI research is focused on innovation, and CI prototypes are focused on moderate use, broad-use CI must exhibit reliability, interoperability, scalability, predictability, stability, and other "ilities" that allow it to support research and education efforts with low barrier to access, high robustness, and low risk of failure. By analogy, the reader relies on uniform and understandable font in reading this article, and τηε χηοιχε οφ τηε φοντ ισ νοτ νοτεωορτηψ υντιλ τηερε ισ α προβλεμ ιν ρεαδινγ ιτ (*translation:* "the choice of the font is not noteworthy until there is a problem reading it"). As illustrated in the previous sentence, where the medium becomes the issue rather than the message, broad-use CI must help the community focus on the problems of research and education rather than on the problems of using the infrastructure.

Support for Broad-Use CI

Good, broad-use CI is typically supported by a professional team whose focus is development, maintenance, and evolution of the software, as well as assistance to the user community (an aspect that users find critical for real effectiveness). The level of software engineering effort, the need for interoperability with other infrastructure, and the focus on robustness and "last mile" usability critical to successful broad-use CI typically exacerbate the problems of recruiting, supporting, and developing professional staff in the academic sector, as described for Stage 2 of the CI trajectory.

Of the many "ilities" so critical to making broad-use CI useful to and usable by the research and education community, *economic sustainability* is arguably the most challenging: the longer the time frame, the more difficult it is to create a model that supports sustainable CI over generations of hardware, software, staff, and funding. For example, stewardship of the PSID over four decades has required migrating data through generations of storage media with minimal risk, stabilizing funding for PSID staff, and evolving support for the collection with respect to community access modes and use patterns.

The kind of budget "mortgage" seemingly required for preservation of data collections such as the PSID, the Protein Data Bank, and others makes their economic support a particular challenge within a federal research framework that focuses primarily on short- to medium-term time frames and new starts. Some pioneering federal programs are beginning to address the issue of sustainable broad-use data CI support and are challenging the community to develop creative partnerships to address the problem. For example, the recent NSF DataNet Request for Proposals (http://www.nsf.gov/pubs/2007/nsf07601/nsf07601.pdf) included an explicit expectation of a viable sustainability plan post–award completion for the data CI that is developed during the award, creating a pathway to extend the time frame of federally funded data CI to alternative economic models. The economic sustainability of long-lived digital data of community value will be the subject of upcoming reports in 2008 and 2009 from the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (http://brtf.sdsc.edu). For universities and colleges, it remains difficult to obtain long-term, stable funding for CI provision. Forward-looking administrators cognizant that sustainability is critical for success are currently focusing on new line items for CI in institutional, state, or federal budgets and on other means for continuing support, rather than on short-term or one-time funding.[11]

The process of developing long-term funding sources generally requires a compelling demonstration of a clear value proposition for CI or of sufficient return on CI investment to make it an attractive option for sponsors. To demonstrate such a value proposition and/or ROI for sustainable broad-use CI, proponents need to show that CI is "good," "useful," "usable," and "cost-effective." However, in the absence of objective and concrete metrics, it is surprisingly difficult to quantify exactly what "good" CI is. At present, there are few widely used measures for assessing the (aggregate) quality or ROI of broad-use CI. Examples of some measures that are currently being used to assess several key CI characteristics are listed in Table 1. In addition, the development of aggregate productivity measures applicable to CI is also the focus of DARPA's High Productivity Computer Systems program (http://www.highproductivity.org/).

Table 1. Measures for Assessing CI Characteristics

Measure Type	What Is Assessed	Example Measures and Metrics
Usage	Amount of use of resource by user community	Number of users of resource; utilization, throughput (computation); number of collections (data); number of hits (web); number of downloads (software)
Usability	"Ease of use" of resource by user community	Turnaround time (computation); user satisfaction as assessed by surveys; informal feedback from users; software productivity measures
Deep impact	Importance of science and engineering enabled by resource	Publication in peer-reviewed journals and conferences; community recognitions and awards; "landmark" publications
Broad impact	Extensiveness of user community; accessibility of resources	Number of disciplines, communities served; number of publications enabled; number of courses, dissertations, and other educational vehicles enabled
Workforce impact	Individuals involved in the provision of CI	Number (gender, race, creed, level) of individuals involved in CI-related professions; number (gender, race, creed, level) of individuals with CI-oriented education or training and their increase/decrease over time

Source: Drawn from Table 2 in Francine Berman, James Bernard, Cherri Pancake, and Lillian Wu, "A Process-Oriented Approach to Engineering Cyberinfrastructure," Report from the [NSF] Engineering Advisory Committee Subcommittee on Cyberinfrastructure, February 2006, p. 15, <http://director.sdsc.edu/pubs/ENG/report/EAC_CI_Report-FINAL.pdf>.

Clearly, better qualitative and quantitative measures of CI success are needed. Because it is difficult to promote and improve what isn't measured, the sparsity of representative aggregate CI success measures makes it hard to demonstrate CI's value to sponsors, and to foster clear improvement of CI over time.

***

In summary, the development of a healthy pipeline from invention to broad use along the CI trajectory is critical. Each stage must involve appropriate participants, support, and success measures. Moreover, mechanisms are needed for evolving the most promising efforts from one stage to the next and for sustaining the most critical broad-use CI over time. Without such a pipeline, academia will have difficulties building, maintaining, and evolving the CI required to propel modern research and education forward.

Cyberinfrastructure Education

In the 1980s, the use of high-performance computing and information technologies to accelerate "grand challenges" in science and engineering initiated a paradigm shift in the conduct of academic research. Over the next two decades, educational curricula and programs focusing on the methods and problems of computational science and engineering blossomed as a way to enrich the education and improve the competitiveness of students.

The development of broad-use CI tools and technologies for research and education now provides the same opportunity. Viewed as a potential driver for a new kind of "educational discipline," CI will likely form the focus of curricula, courses, and programs in modern universities and colleges over the next decade. These programs could empower new generations of students by providing training in the methods, approaches, algorithms, and models critical to conducting twenty-first-century research.

What would a CI curriculum consist of? One could imagine such a curriculum to be based on a solid grounding in computer science, mathematics, and engineering concepts, as well as an appreciation for the issues that arise in domain research enabled by CI. The CI curriculum might also include the following components:

- *Solid understanding of statistics and probability.* The incorporation of sensors into everything from the tagging of laboratory animals to the structural analysis of seismic stresses on bridges, as well as the development of programming environments for high-performance computing architectures with millions of cores, highlights CI environments for which statistical analysis will become increasingly important. A solid understanding of statistics and probability will be critical to designing the next generation of CI methods and tools.
- *Knowledge of economics and social science.* Modern CI environments are set within a larger landscape and often require sophisticated assessment of relational dynamics among and between components and people. Distributed environments in which users have "currency" (e.g., time, cycles, bytes) may require analysis of complex trade-offs for optimization

strategies and/or organizational frameworks that maximize aggregate behavior. A working knowledge of key concepts from economics, as well as from the behavioral and social sciences, will form an important base for understanding such environments.

- *Awareness of policy*. Policy frames what can and cannot be done in the world of CI. Knowledge of requirements or regulations such as Sarbanes-Oxley (for financial reporting constraints leading to the need for data preservation) and HIPAA (for health privacy constraints leading to the need for increased system security and data anonymization) is key to designing CI that fits within a larger environment. An awareness of the organizational and business management models that provide the context for modern policy provides a foundation.
- *Grounding in the real world*. Tracking modern CI trends (e.g., the prevalence of collaborative technologies and environments such as Facebook, the rise of cloud computing options for research and education, the emergence of multicore technologies for high-end machines) expands the set of options available for CI design and provisioning. A solid grounding in these real-world trends is an important component of a well-rounded CI curriculum.

The development of CI programs and curricula will challenge traditional educational modes of delivery and assessment. Moreover, the focus on applied *infrastructure* will push the envelope further if CI is to become a legitimate academic research and education discipline. The next decade will provide an opportunity to address this challenge head-on, with the potential to help evolve the academic system to address the needs of CI research and education in the Information Age.

A Call to Action

At this point, it is hard to imagine modern research and education without the transformational influence of CI. Yet, as described in this article, the academic community continues to struggle to provision and sustain broad-use community CI within traditional academic frameworks. Changing this will involve a paradigm shift in the way we think about designing, evolving, provisioning, and learning about CI; new partnerships between academics, the federal government, and the private-sector-focused CI; and new strategies to incorporate CI within academic infrastructure.

Much of this will be new territory for many faculty researchers and educators, as well as for university and college administrators, CIOs, and librarians. To make research and education CI real, we need creative, broad-scale

community initiatives, as well as plans for their sustained implementation, to include the following:

- *Development of programs to support and accelerate the trajectory of research in CI to broad-use CI.* Academic venues and models that support a healthy pipeline along the CI trajectory—from promising research to usable prototypes, from promising prototypes to broad-use CI, and from widely used CI to sustainable community infrastructure—must become an integral component of the research and education landscape.
- *Creation of sustainable institutional and community economic models for research and education infrastructure.* Strategic and sustainable partnerships between the academic, private, and public sectors will be critical to addressing key questions such as: Who will pay the "data bill" over the next decade? Who will support national and university/college computational environments (including their critical human infrastructure) over the long term? Realistically, we should expect a spectrum of solutions, from "endowed" resources to low-barrier-to-access user fees.
- *Incorporation of university and college CI as campus infrastructure.* Over the next decade, new roles and responsibilities will be key to making universities and colleges competitive: the roles of academic libraries may expand to include stewardship of faculty research data; the roles of CIOs may expand to oversee campus research CI; and networking from the laboratory to the data center on campus and beyond to national facilities should provide an enabler, rather than a roadblock. Embedded university/college infrastructure in 2018 should include data centers, co-location or condominium clusters, and adequate networking, in addition to power and electricity.
- *Development of CI curricula and programs.* A CI-savvy workforce requires education and training. Courses and programs in CI will need to be developed and incorporated into university and college curricula just as computational science courses and interdisciplinary programs began to be developed a decade ago.

These and other initiatives will be critical to ensuring that the academic community can conduct twenty-first-century research and education with twenty-first-century tools and infrastructure. Only then will cybeinfrastructure become a real foundation for research and education and achieve its transformative promise to accelerate the next generation of invention, innovation, and discovery.

Notes

1. In this article we use the following distinction between *invention* and *innovation:* "Invention is the first occurrence of an idea for a new product or process, while innovation is the first attempt to carry it out into practice." From Jan Fagerberg, "Innovation: A Guide to the Literature," in Jan Fagerberg, David Mowery, and Richard Nelson, eds., *The Oxford Handbook of Innovation* (New York: Oxford University Press, 2005), pp. 1–26.

2. *Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure,* January 2003, <http://www.nsf.gov/od/oci/reports/CH1.pdf>.

3. National Science Foundation, *Investing in America's Future: Strategic Plan FY 2006–2011,* <http://www.nsf.gov/pubs/2006/nsf0648/nsf0648.jsp>.

4. Domestic Policy Council, Office of Science and Technology Policy, *American Competitiveness Initiative: Leading the World in Innovation,* February 2006, <http://www.whitehouse.gov/stateoftheunion/2006/aci/aci06-booklet.pdf>; President's Council of Advisors on Science and Technology, *Leadership Under Challenge: Information Technology R&D in a Competitive World,* an Assessment of the Federal Networking and Information Technology R&D Program, August 2007, <http://www.ostp.gov/pdf/nitrd_review.pdf>.

5. Igor Tsigelny, Pazit Bar-On, Yuriy Sharikov, Leslie Crews, Makoto Hashimoto, Mark A. Miller, Steve H. Keller, Oleksandr Platoshyn, Jason X. J. Yuan, and Eliezer Masliah, "Modeling the Molecular Basis of Parkinson's Disease," *SciDAC Review,* issue 6 (Winter 2007), <http://www.scidacreview.org/0704/html/parkinsons.html>.

6. Yuriy Sharikov, Ross C. Walker, Jerry Greenberg, Valentina Kouznetsova, Sanjay K. Nigam, Mark A. Miller, Eliezer Masliah, and Igor F. Tsigelny, "MAPAS: A Tool for Predicting Membrane-Contacting Protein Surfaces," *Nature Methods,* vol. 5, no. 2 (February 2008), p. 119.

7. Daeseob Lim, Jaewook Shim, Tajana Simunic Rosing, and Tara Javidi, "Scheduling Data Delivery in Heterogeneous Wireless Sensor Networks,"

Proceedings of the Eighth IEEE International Symposium on Multimedia (Washington, D.C.: IEEE Computer Society, 2006); Gaurav Dhiman and Tajana Simunic Rosing, "Dynamic Voltage Frequency Scaling for Multi-Tasking Systems Using Online Learning," *Proceedings of the 2007 International Symposium on Low Power Electronics and Design* (New York: Association for Computing Machinery, 2007).

8. Gary Solon, Marianne E. Page, and Greg J. Duncan, "Correlations between Neighboring Children in their Subsequent Educational Attainment," *Review of Economics and Statistics,* vol. 82 (August 2000), pp. 383–92.

9. Kevin J. Bowers, Edmond Chow, Huafeng Xu, Ron O. Dror, Michael P. Eastwood, Brent A. Gregersen, John L. Klepeis, et al., "Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters," Best Research Paper, SC06, November 14, 2006, <http://sc06.supercomputing.org/schedule/pdf/pap259.pdf>.

10. Shava Smallen, Kate Ericson, Jim Hayes, and Catherine Olschanowsky, "User-Level Grid Monitoring with Inca 2," *Proceedings of the 2007 Workshop on Grid Monitoring* (New York: Association for Computing Machinery, 2007).

11. A good description of some of the economic issues these administrators face can be found in Rosio Alvarez, "Developing and Extending a Cyberinfrastructure Model," *EDUCAUSE Center for Applied Research (ECAR) Research Bulletin,* vol. 2008, issue no. 5 (March 4, 2008), <http://connect.educause.edu/Library/ECAR/DevelopingandExtendingaCy/46339>.