

Paper for the Designing Cyberinfrastructure for Collaboration and Innovation workshop, National Academies, January 2007.

## **Social and behavioral scientists building cyberinfrastructure**

David W. Lightfoot

Assistant Director, National Science Foundation  
Social, Behavioral & Economic Sciences

The behavioral and social sciences, sometimes called the human sciences, have been transformed in fundamental ways over the last 15-20 years, largely by access to new comprehensive databases and by the new analytical possibilities that they have opened up. My own work has concerned language change and the acquisition of syntactic structures by young children. Papers in these areas of linguistics look very different now from what was written twenty years ago. However, today I will not talk about this. Rather, I want to address the matter of what the behavioral and social sciences contribute to the development of cyberinfrastructure (CI). After all, cyberinfrastructure is for humans and its development draws on the understanding of humans that the human sciences provide, dealing with them as individuals and in groups.

We are now at a unique moment in the history of science, certainly for the last four hundred years. Our modern sciences grew out of a common domain of natural philosophy. Each developed its own methodologies and tools. Physicists had their telescopes and accelerators, biologists their microscopes and sequencers, chemists their Bunsen burners and Petri dishes, social scientists had their surveys and behavioral scientists their brain imaging devices and eye trackers. Sciences continue to have their specialized tools but now for the first time all sciences are concerned with one central set of tools, cyberinfrastructure tools that, as NSF's Blue-Ribbon Advisory Panel on Cyberinfrastructure, the Atkins Report, points out, provide a "third way" for science, a path that makes use of intensive numerical computation, new types of computer-assisted meta-analysis and CI-enabled collaboration that undermines barriers of time and space. We are already moving in that direction, seeing new interdependencies among scientific and technical communities and new ways of thinking, testing, experimenting and innovating.

As for the question of how the SBE sciences contribute to the development of CI, this is covered extensively in the report of the 2005 joint CISE-SBE Airlie House conference (F. Berman and H. Brady *Workshop on Cyberinfrastructure for the Social and Behavioral Sciences: Final Report* (<http://vis.sdsc.edu/sbe/>), showing how SBE scientists can help CISE researchers design a functional and effective CI achieving its full potential and how they can assess and optimize the impacts of cyberinfrastructure on society.

Here I will tackle SBE contributions differently, briefly discussing the four major areas of new investment in the SBE sciences in the context of the NSF budget. When I say in the context of the NSF budget, I refer to SBE components that show up in the NSF Request budget that is rolled out each February as part of the President's proposed budget; for example, the 2007 budget rolled out in February 2006 outlined our Science of Science &

Innovation Policy initiative and we hope to see work on neurotechnology and the environment featuring in future budgets. Each of these areas requires new kinds of computing, but in different ways. In each case, our sciences progress but I focus here on how the SBE work stretches and develops cyberinfrastructure.

### **Neurotechnology**

Behavioral scientists are making headway in understanding how brains function. Neuroscience is not really a field but, instead, a cluster of techniques used by linguists, psychologists and even economists to understand brain function. Neuroscience advances are heavily dependent on the myriad technologies underlying such machines as EEG, fMRI, PET scans, etc, each of which depends on sophisticated computational breakthroughs and each of which permits us to image aspects of brain activity. Despite our advances over the last generation, there are aspects of brain activity that we know next to nothing about and we need to learn more about what is truly important. For example, there are almost ten times as many glial cells as neurons in a brain, yet the function of glial cells is not understood or even identified. Surely brain science is in its infancy. New tools, new imaging devices monitoring new aspects of brain physiology will be developed and change our ideas. These devices will not be invented by biologists or behavioral scientists but by collectivities of physicists, chemists, mathematicians and computer scientists, guided by biologists and behavioral scientists who study brain functions – a truly transdisciplinary enterprise. Each specialty will help to develop new imaging devices, which will certainly be computationally based. The work will open new frontiers for understanding how human beings function.

At present we have limited capacity to gather coordinated, simultaneous data from different monitoring devices, say an eye tracker and an EEG. This presents seriously challenging computational problems. With CISE/OCI, we funded a Next-Generation Cybertools award to the University of Chicago to solve those computational problems. A solution promises opportunities to synchronize study of social behavior with neurobiology. It is hard to imagine that such work will not transform our understanding of humans and it will also entail new kinds of computational machinery.

Research on cognitive brain function is becoming increasingly cyber-intensive as the field of neuroinformatics develops. A typical functional neuroimaging data set from an individual participant may consist of 1000 3D images each consisting of around 75,000 volume elements, creating about a gigabyte of data. A busy imaging center may produce that much data every hour most hours of most days of the year, yielding tera-scale distributed storage requirements locally and peta-scale requirements across centers. Work in brain science will be one of the vehicles for *behavioral* scientists to make more extensive use of high-performance computing capacities and building the computational machinery will entail developments in cyberinfrastructure more broadly.

### **Environment**

A second area of investment is the environment and this will lead *social* scientists into new kinds of high-performance computing. The directorates of BIO, GEO and SBE just established the first interdirector standing program at the Foundation, the Dynamics of

Coupled Natural and Human Systems. This will be a vehicle for future investments in environmental matters, including work on climate change. It will also serve as a vehicle for further incorporating SBE databases into environmental observatories and collaboratories, as well as building the capacity for large-scale integrative modeling.

One of the computational foundations for SBE involvement here is work in Geographical Information Systems (GIS). SBE provided core support to the National Center for Geographic Information and Analysis (NCGIA) at three universities and, more recently, managed major infrastructure awards to support the Center for Spatially Integrated Social Science (CSISS) at Santa Barbara and the National Historical Geographic Information System at the University of Minnesota-Twin Cities. These and other efforts have resulted in the development of a suite of geospatial tools, GIS, global positioning systems, and remotely sensed data, which have transformed the study of geography. Principal component analysis, landscape change detection, and regional analysis are examples of methodologies that have been transformed due to GIS. Scientists from various disciplines have used CSISS to study many topics with societal impacts, such as environmental change, inequality, business networks, criminal justice, health and disease.

A number of the National Science and Technology Council's priorities for disaster-related research, as outlined in their *Grand Challenges for Disaster Reduction*, hinge on developing and integrating detailed information about potential bad events, the environments impacted, and the social and economic consequences of those impacts. The most promising approaches to studying these phenomena involve computationally intensive models and simulations and, again, social scientists work with natural scientists and computational scientists. Predicting potential effects of bad events in advance is critical to enabling better prevention, preparation and mitigation.

Another example of the transformative potential of high-performance computing for the SBE sciences comes from the capacity to simulate societies or aspects of them and to trace out the implications of variations in key parameters. Limited simulation is familiar in SBE research, but even the most complex models designed to capture social processes are highly simplified both in the number of parameters they contain and the relations they allow between them. Social relations, even on small scales, can be exceedingly complex, and to capture them effectively requires substantial high-speed computing resources, perhaps even exceeding what is required for weather prediction, creative software engineering and the capacity to incorporate huge quantities of data as well as techniques for systematically varying and testing assumptions about how people interact with each other and with their environments. One might think of the computational goal as building, in a machine, digital Chicago, or, on a more modest scale, developing the capacity to anticipate how changes in the price of farm land in Northern Illinois will affect population movements, birth rates and pollution in Lake Michigan.

Our scientific communities are excited by the prospect of participating in major observatories, a category of CI that enables fine-grained multidimensional recording of natural and human-built assets over time. The prospects for mapping and analyzing

trends, patterns and interactions are enormously exciting and we had a workshop on this topic at the NSF last week. All this opens up new high-performance computing capacity.

### **Science of Science & Innovation Policy**

A third area is what we call the Science of Science & Innovation Policy (SciSIP). The broad goals of this initiative are to investigate how national research and development systems work, how to nurture and measure innovation, and how to direct our investments. The long-term goal is to provide science policy makers with the same kinds of analyses and advice that economists now provide the Federal Reserve.

Of particular interest is the fundamental impact of cyberinfrastructure on scientific research and scientific culture. New databases, tools, analytical techniques and computational capabilities are changing how research questions are asked. Understanding the impact on approaches to scientific inquiry is of particular interest in tracking innovation and identifying relevant scientific metrics.

Cyberinfrastructure has undermined disciplinary barriers, increased access to digital data, and created new mechanisms for sharing computational tools. Changes in scientific culture - how different disciplines interact - create new opportunities and venues for interdisciplinary research. The human sciences study the transformative effects of cyberinfrastructure on users, organizations and scientific disciplines.

The SciSIP initiative focuses on understanding these matters, building models and developing new kinds of data, which will influence the much cited bi-annual S&E Indicators. This will involve new kinds of data extraction. Another goal is to move beyond the science-neutral approach to these matters, whereby bibliometric techniques treat sciences like chemistry and sociology alike, and to bring social and behavioral scientists together with scientists from specific domains, nanotechnology or climate change, for example. That will be facilitated by developing virtual collaboratories so that collaboration does not depend on being in the same place at the same time. The drive to develop new data extraction techniques and these interdisciplinary collaboratories will stretch existing capacities and build new CI.

There is a particular focus on innovation and organizations (we have a program Innovation and Organizational Change). Development of computational tools to advance our understanding of factors within and between organizations that yield effectiveness and innovation is a priority for SBE for three reasons. First, because most major human endeavors are now conducted in the context of organizations, laboratories, companies, government agencies, etc. Understanding what makes organizations innovative, effective or ineffective is pertinent to social, economic and scientific advances. Second, because some of the technical challenges that are central to organizational research tools - such as simultaneously ensuring accessibility and confidentiality of data; integrating qualitative with quantitative information in multiple media; and developing incentives, standards and policies for data access and database sustainability - are also crucial for other fields. Third, because this particular research community has historically made advances when large-scale databases became available, we expect that development of next-generation databases and other computational tools for

organizational research will have significant social and scientific value. Development in this area will have positive value for private sector vitality as well as public sector effectiveness and even advancing the management of scientific enterprises.

### **Cyberinfrastructure**

In these three domains social and behavioral scientists are building new CI, driven by the needs and opportunities of their sciences. But the fourth area of investment is cyberinfrastructure itself, an object of study for the human sciences.

The Interstate highway system provides an infrastructure to which cyberinfrastructure is sometimes compared. When we drive down a six-lane highway or see those blue lines criss-crossing a map, we are not looking at the infrastructure but its central component. To see the infrastructure one needs to examine trucking companies, bus companies, how individuals use their automobiles, how society deals with the kinds of accidents that happen on highways and the pollution that is generated, the effects on towns and cities near the highways, and so on.

So far I have focused on the computational aspirations of certain of our sciences, the highways. But then there is cyberinfrastructure itself. Our division of Behavioral & Cognitive Sciences recently appointed Terry Langendoen as our CI pointperson and he is now working halftime in OCI. There are many possibilities for collaboration as social and behavioral scientists investigate ways of developing infrastructure. Behavioral and social scientists study human behavior in many domains, including in science, and have much to contribute to developing the infrastructure associated with new computational capacities.

### *Data and confidentiality*

SBE is committed to continue developing and deploying data-oriented CI, including investments in upgrading the “gold standard” surveys, e.g. the General Social Survey, the Panel Survey of Income Distribution, the American National Election Survey, the Luxembourg Income Study; new data infrastructure projects, including sophisticated multimedia data sets; investments in facilities and technology for accessing confidential social and behavioral data resources; toolkits for facilitating data annotation, integration, mining, analysis and validation; and facilities for preserving data over the long term. Solving the interoperability problem for annotation of text, audio, image and video data of the kind generated by research on human cognitive capacity is one of the great challenges facing computational behavioral science. It is the annotation of such data that provides the basis for scientific understanding, and given the ever-increasing volume of richly structured annotated data, the annotations must be designed to enable computational devices to “understand” them in the manner that we do, and to reason over them.

An example is the Documenting Endangered Languages effort, in which the Linguistics Program in SBE has partnered for the past three years with the National Endowment for the Humanities and the Smithsonian Institution to establish sustainable data repositories for the languages of the world that are faced with extinction. SBE is supporting efforts to

provide CI resources for annotating these linguistic data sets in such a way that the results can be made interoperable, and to enable field linguists to record their data and annotations on hand-held devices in the field for upload to secure servers for subsequent, possibly collaborative, analysis.

Some of the most exciting opportunities for building and using new data sets require access to confidential micro-data about individuals, households and organizations. Breakthroughs are needed to reconcile research needs with legitimate demands for privacy and confidentiality. There are also exciting questions to answer if we can develop interoperable data systems to facilitate interdisciplinary research, with special emphasis on linking physical and biological data to socioeconomic and geographical data. With these capabilities, scientists can probe relationships between, say, happiness and career choices or between health and social ties following disasters.

#### *Broadening participation*

Participatory practice and community-based action have become increasingly important tools for decision making. Integration of CI into decision making can offer new methods for participation and improved accessibility. Questions to be addressed include: Do innovative methods for improved accessibility narrow the digital divide? Do participatory practices via the cybersphere enhance democratic processes? Does cyberinfrastructure improve participatory opportunities and experiences? We seek to develop better metrics for monitoring the understanding and use of cyberinfrastructure in the context of Broadening Participation.

It is often noted that concerted efforts by research universities to promote civically engaged science yields benefits both to society and to universities. Such science is facilitated by the rapid diffusion of knowledge and information through information technology, available for everybody with internet access.

The second of our Next-Generation Cybertools awards is to a team of social scientists who are developing tools for mining the Internet Archive's 40-billion page collection of Web pages so that innovation, diffusion and other patterns in related ideas can be identified and analyzed. These tools can be used for competitive purposes - to identify market trends, for example - but they can also be applied to all sorts of general social needs: community watchdog groups can track the spread of "hate sites", citizens in communities hit by natural disasters can identify and coordinate with citizens who seek to offer aid or information, opinion researchers can trace the development and spread of new ideas and norms, and government agencies can trace past and current uses of the Web for organizing and coordinating terrorist attacks. The flood of available on-line information - from corporate web pages to news groups and blogs - has the potential to open up new frontiers in social science research. The computational challenge in this project is to advance capabilities for analyzing very large, semi-structured datasets.

#### *Cyber-savvy workforce*

Developing and supporting new CI tools and training opportunities cuts across these concerns for scientists engaged in managing, sharing and analyzing massive data sets.

Support for education and training opportunities for the development of the next-generation, cyber-savvy scientific workforce and the re-tooling of existing scientists is critical for advancing these research areas, including activities to broaden workforce participation. New access to data and user-friendly analytic programs will advance SBE science education as high school students use cyberinfrastructure advances to engage in hands-on analyses of social relationships.

To use and extend CI requires learning and workforce development initiatives. Work is needed to improve cyberlearning, a theme of one of our Science of Learning Centers in Pittsburgh. NSF Director, Arden L. Bement, Jr., has noted that cyberlearning and collaboration do not replace traditional learning contexts but they do augment them, opening new possibilities, new kinds of interaction among people, information and facilities (*Chronicle of Higher Education*, 5 January 2007). US universities must press ahead with e-learning, finding the best ways of utilizing and developing it.

**Conclusion: Computational thinking/complexity/emergent phenomena**

Cyberinfrastructure is fundamentally changing the way that scientists build and test theories of social, behavioral and economic phenomena. Before computational modeling, theories were constrained to either verbal descriptions that could only loosely capture the qualitative aspects of a complex phenomenon, or mathematical models that needed to be as simple as possible in order to be analytically tractable. New computing capacity provides a third alternative, in which complex phenomena can be simulated in a rigorous computational language, allowing for the inclusion of much more detail than is possible with purely analytic methods. Indeed, complex systems are often defined as those that are composed of many components whose dynamics are governed by non-linear interactions. In the limit, the system behavior that emerges from these interactions cannot be analytically reduced, leaving computational methods as the only viable means of theory building and testing. If one accepts that behavioral phenomena emerge from complex interactions among neural and bodily systems, and that societal and economic phenomena emerge from complex interactions among human systems and their environmental contexts, then cyberinfrastructure will play a critical role in the advancement of theory in the SBE sciences.

SBE scientists are pushing the CI envelope in order to advance their understanding and their sciences not in order to advance cyberinfrastructure. That is a consequence of the new tools that have become available and our scientists are involved in stretching and extending those tools in much the way that physicists and biologists are. All of this means that science itself is now working differently: it is more interactive, more likely to involve complex modeling, and fields are changing more rapidly, entailing new kinds of questions and new kinds of comprehensive answers.

It used to be thought that discovery and innovation were largely unpredictable and that the funders of science should pretty much follow the advice of scientists who could identify the most challenging questions. There was a distinction between basic and applied science. However, Donald Stokes (*Pasteur's Quadrant: Basic science and technological innovation*, Brookings Institution Press, Washington, D.C., 1997)

questioned these ideas and argued that the history of science showed a closer relation between fundamental and applied ideas, that many fundamental discoveries had emerged from science pursuing applied goals. Science and technology co-exist, interact and evolve interactively. Universities have departments of Science Studies and social and behavioral scientists study this, although not necessarily the same SBE scientists who are pushing the CI envelope – the SBE mansion has many rooms. Much has been written about how scientists behave in interdisciplinary research contexts and there is interesting work on where and how it yields exceptional, prize winning discoveries and innovations – I am thinking here of the work of Rogers Hollingsworth. Corporations recognize the need for such work and Intel has about forty anthropologists on its staff.

Given how science has developed with new computational capacities, scientists gravitate to areas where ideas and new technologies interact most productively and they expect that discovery and innovation will most likely come in that way. That provides a basis for prediction. Indeed, I have made some predictions here about where transformative progress is likely to come in the human sciences, and predicting innovation and discovery is what our new Science of Science & Innovation Policy initiative is all about. I hope that this perspective on how our scientists are building cyberinfrastructure may indicate how they might play a larger role in doing so.