EDUCAUSE Center for Applied Research

Research Bulletin

Volume 2009, Issue 6

March 24, 2009

Information Technologies for eScience:

A Preliminary Report from the University of Washington

Louis Fox, University of Washington and WICHE

Cara Lane, University of Washington

Ed Lazowska, University of Washington

with

Janice Fournier, University of Washington

Greg Koester, University of Washington

William Washington, University of Washington

4772 Walnut Street, Suite 206 ◆ Boulder, Colorado 80301 ◆ www.educause.edu/ecar/

Overview

The University of Washington (UW) has a long and proud tradition as a research leader. UW is the top public university in annual federal obligations for science and engineering research,[1] with sponsored research at UW totaling over $1 billion in 2007. We recognize, though, that past performance cannot guarantee future competitiveness. In the current research climate, no institution, however successful, can afford to rest on its laurels. Therefore, we have been engaged in a large-scale effort to assess the information technology (IT) needs of UW's top researchers.

Conversations with the University of Washington's Research Leaders is a joint project of the Office of UW Technology (the central technology unit for our main campus in Seattle) and the eScience Institute (a new unit formed to address the emerging IT needs of UW's researchers). This project is unique in its scale and focus. At this time, approximately 50 technology professionals have met with more than 100 top researchers in a broad range of fields to learn firsthand about the pivotal role that technology plays in supporting their efforts. The goals of this project are (1) to increase project partners' awareness of research projects being conducted and research questions being investigated at UW, (2) to inform partners of the likely future direction of research within a variety disciplines, (3) to understand how researchers currently use technology to support their research activities and how they anticipate using technology in the future, and (4) to identify resources and services that UW Technology and the eScience Institute can offer to help researchers at UW maintain and build upon their remarkable record of success.

In this ECAR research bulletin, we discuss the key findings of the first two phases of the UW conversations, including important trends related to the need for data management expertise, computing power and network access, data collection and analysis, and communication and collaboration. This information is drawn from a preliminary project report.[2] We also discuss the short- and long-term impacts of these findings for higher education. While the priorities we discuss in this bulletin are specific to UW, many of the needs we identify are likely to apply to researchers at other universities. We anticipate that this bulletin will inspire other institutions to embark on similar investigations to better understand their researchers' needs and priorities.

Highlights of Information Technologies for eScience

In the past 20 years, many fields of science and engineering have been transformed by the introduction of large-scale simulation, as exemplified by the National Science Foundation (NSF) Supercomputer Centers program and subsequent Partnerships for Advanced Computational Infrastructure.

Today, a second computational science revolution is taking place, one that promises to be far more pervasive. This revolution is driven by ubiquitous sensors: desktop gene-sequencing machines that generate a terabyte of data per day; next-generation

telescopes such as the Large Synoptic Survey Telescope (LSST) that will generate 30 times this data rate; physics facilities such as the Large Hadron Collider with data rates similar to the LSST; ocean observatories that are seafloor networks with tens of thousands of chemical, physical, and biological sensors, transforming oceanography from an expeditionary to an observatory science; point-of-sale terminals streaming worldwide data to corporate headquarters for analysis. The enormous data volumes involved drive the need for semi-automated analysis of that data using techniques such as data mining and machine learning. This next-generation data-centered computational science has come to be called eScience.

Traditional simulation-oriented computational science has been transformative, but it has been a niche. It is our belief that eScience, in contrast, will be pervasive and that universities that do not provide institutional support for eScience—for the collection, management, and analysis of enormous volumes of data—will become less competitive. It is with this point in mind that the UW eScience Institute was established and the *Conversations with the University of Washington's Research Leaders* project was undertaken.

Methodology

The initial goals for the *Conversations* project involved providing senior technology professionals the opportunity to learn about research at UW and build relationships within the research community. During Phase I of this project, from spring 2007 to spring 2008, interviews had a fairly informal structure, with interviewers asking general questions about current projects, future directions of research, and the role of technology in those projects. Data from Phase I interviews were captured in audio recordings and interviewers' notes.

In late spring 2008, a project team was assembled to initiate a more formal needs-assessment process (Phase II) to identify potential resources and services that UW Technology and the eScience Institute could provide to researchers. The project team began this process by meeting with Phase I interviewers, reviewing interview notes, and listening to interview recordings to identify common themes. The findings from Phase I informed the creation of interview questions for Phase II, which focused on four main areas: current research, future research, current technology use, and future technology needs. Data from Phase II interviews were captured in audio recordings, interview notes, and interview summaries. During both phases of this project, two interviewers met for up to one hour with each researcher. Some researchers chose to have other members of their research teams present during the interviews.

Over the course of this project, we contacted 290 of the top researchers at UW. Researchers were selected based on the number and monetary amount of grants received relative to others in their fields, as well as their status as National Academy members or as recipients of prestigious national awards for younger researchers from the David & Lucile Packard Foundation, Alfred P. Sloan Foundation, NSF, M.J. Murdock Charitable Trust, and similar entities. We also contacted researchers who were recommended by their peers. It is important to note that our interviewees represent

UW's top researchers of all ages across a wide variety of fields, rather than current "computational scientists." This is what distinguishes our study.

- **Phase I:** We conducted 37 interviews from spring 2007 to spring 2008. In addition, one researcher e-mailed responses to the project team, for a total of 38 participants in Phase I.
- **Phase II:** We conducted 84 interviews during summer 2008. In addition, two researchers e-mailed responses to the project team, for a total of 86 participants in Phase II.
- **Total Response:** Of the 290 researchers originally contacted, 124 researchers have participated in this effort to date. Since this project began, 26 researchers on our contact list have retired or left the UW. Among the 264 current UW researchers contacted, our response rate is 47%.

In our analysis to date we have examined interview summaries from half of the interviews conducted during Phase II. We tracked common themes across these interviews to understand the larger picture of how researchers are working with technology and how technology is and is not working for researchers at UW. In this bulletin we discuss preliminary findings.

Findings

In general, UW's researchers are pursuing their research in an increasingly competitive funding environment. Several researchers mentioned the low percentage of grant applications funded annually by the National Institutes of Health (NIH) and other agencies. Although NIH was a commonly cited example, these observations transcended disciplinary boundaries. While many of these researchers expressed confidence in their ability to maintain funding levels in this environment, some acknowledged that doing so would require them to spend significantly more time applying for funds, since repeated applications are increasingly required to secure a grant. Other researchers reported that they had sought or were considering seeking additional sources of funding (such as private donors, corporations, or nontraditional government agencies) or a new scale of funding (for example, funds to establish a program or center rather than funds for individual projects).

Another general trend among researchers is an increasing focus on interdisciplinary and multi-institutional projects. This trend is driven both by the complexity of research questions being raised and by granting agencies' preference for funding translational and collaborative projects. In the interviews we analyzed for this bulletin, a substantial majority of researchers indicated that they were reaching across disciplinary boundaries at UW to conduct their research. For many, these interdisciplinary partnerships represented new relationships, which they would not likely have formed in the past. For instance, a researcher in microbiology whose past work on plant diseases and genetic modification had been completely lab-based is now investigating what happens to these plants in the field, in part due to funding requirements. Due to the complexities of understanding soil composition, this shift required collaboration with a specialist in soil

science; the researcher has found partners in the UW Department of Forestry. Nearly all researchers also indicated that they were collaborating extensively with researchers at other universities, and almost a third mentioned international partners. Additionally, many researchers are working closely with local and federal government researchers and industry partners.

While at first glance the above trends may appear unrelated to technology or technology support, they provide a valuable context in which to situate researchers' technology-related needs. For instance, any additional time researchers spend pursuing funding leaves less time for them to spend conducting their research or focusing on technological issues; in addition, researchers' multi-institutional partnerships directly influence their needs regarding data management, communication, and collaboration.

Given the diversity of research projects conducted at UW, a clear consensus on technology-related priorities and needs did not emerge among researchers. The findings below represent areas of convergence where significant numbers of researchers across disciplines mentioned similar needs or where a subset of researchers voiced nearly identical needs. The five themes that follow are listed in order of priority, based on the number of researchers who reported needs in these areas: data management, shared expertise, computing power and network access, data collection and analysis, and communication and collaboration.

Data Management

Researchers in a variety of disciplines are collecting large and often rapidly increasing volumes of data. In many disciplines, researchers are collecting a vastly larger amount of data today than was possible to collect a decade ago—or, in some sciences, a month ago. Researchers in fisheries, for example, are continuing to add to a database on Alaska's salmon population, which they have been tracking since 1946. However, the amount and type of data they collect today are significantly more complex, detailed, and voluminous than data collected in the early decades of the project. Even more significantly, researchers in genome sciences reported that where they once ran 96 samples per DNA-sequencing machine, they can now run 10 to 20 million per machine. They anticipate this number to increase by a factor of two every few months over the next few years.

The rapid evolution and widespread deployment of sensors—sensors in gene-sequencing machines, in telescopes, on the seafloor, in point-of-sale terminals, and so forth—are causing an explosion in data volumes. We find, though, that even top-tier researchers are managing their data using approaches that badly lag behind their current needs—using spreadsheets or flat files, for example, rather than relational database systems deployed at the laboratory, institutional, or cloud level.

It is not surprising, then, that the most common challenges researchers expressed in our interviews involve one or more aspects of data management; this was mentioned in over half of the interviews we analyzed for this bulletin. The most common data management needs include access to sufficient storage, reliable backup systems, and adequate security.

- **Storage:** Many researchers reported that they need assistance with data storage, whether in storing large amounts of data for current research projects or archiving data from past projects so that the data can be easily accessed in the future. Several mentioned needing access to terabytes of storage space, while one researcher in electrical engineering asked for petabytes.

- **Backup:** Systems for backing up data were inconsistent among researchers. Several reported concerns about adequate data backup, while a few reported prior problems with lost data. Interviewers also heard of questionable practices, such as storing all data for a project on a laptop or flash drive, that were not always recognized as problematic by the researchers involved.

- **Security:** Security, especially in terms of access control, was a priority for many researchers because data collected in many studies is confidential and access needs to be limited to the research team. In addition, the expansive range of local, governmental, corporate, academic, and international partners with which researchers collaborate adds additional complexity, since institutions often rely on different authentication systems and security protocols.

Several researchers mentioned central data storage or central backup services as potential solutions to meeting their needs in these areas.

Shared Expertise

In addition to describing technological needs in areas of data management, approximately one-quarter of the researchers who were included in our analysis detailed the need for expert assistance in configuring and maintaining databases and servers. A similar number of researchers commented on a general need for local technology support, while a few others desired more access to information about current and future technologies.

- **Data Management Expertise:** Researchers asked for expert assistance handling data management problems ranging from consolidating and restructuring large data sets to setting up a system to ensure that sensitive data are stored securely and are destroyed on a set timetable. A few researchers reported needing assistance to begin envisioning a better system for data management because their teams lacked expertise in this area. One such researcher sought to migrate data off of individual researchers' computers, where data can be difficult to access if someone leaves the team, to a shared storage solution of some type; another was interested in upgrading from a system of storing data in three-ring binders to a digital solution. Some researchers outsourced data management and, at times, data collection to for-profit companies such as DatStat because they could not find a UW service that met their needs.

- **Local Technology Support:** While some research teams include dedicated support staff, it is also common for researchers to have a member of their team, often a graduate assistant, take on technology support in addition to, or in place of, his or her research work. One researcher shared a story that aptly illustrates

the challenges of this support model: when his graduate research assistant could not find a driver for a new printer, it was more cost-effective to exchange the printer for a different model then to keep looking online for the driver. In addition to relying on members of their research team, most researchers also regularly utilized departmental technology support. Many researchers reported the need for increased local technology support, whether at a departmental level or shared between research teams or departments located in close proximity to each other. A few researchers commented on needing additional support for Mac computers because their departmental support focused on PCs.

- **Information:** A few researchers desired more information about the technologies and technological expertise currently offered by UW. Researchers wanted a central site to access this type of information. One researcher offered a vision of a new service that would inform her of new technologies related to her work as they became available.

In general, lack of access to shared eScience expertise is a significant cause of inefficiency in UW's research enterprise. A common refrain among researchers was that they do not want to spend too much time managing data, solving technology support problems, or seeking out technological information—they would rather spend their time doing research. Many of these researchers wanted to consult with a database administrator, system administrator, or technology support person on an as-needed basis rather than having to pay for a permanent staff member with these skills. Several envisioned a communal solution to these problems, where staff with appropriate expertise would be shared among research teams or departments or where UW offered a consulting service to meet these needs.

Computing Power and Network Access

About one-third of the researchers discussed computationally intensive activities as necessary to their work, and about one-fifth stated having specific needs related to accessing high-performance computers or computer clusters. In addition, for a few researchers, access to high-bandwidth networks is an important component of their computing needs. While some researchers' computing needs are met departmentally, others have to contract for these services with other departments, institutions, or businesses.

- **Computing Power:** Many researchers require considerable amounts of computing power for activities such as generating statistics, analyzing data, creating models and simulations, and general "data crunching." While these researchers expected to have an ever-increasing need for more powerful machines, they cited a lack of resources, space, and funding as limitations to acquiring more computing power. Some researchers expressed frustration at not being able to fulfill what they perceive as the potential of their work or their data due to these shortfalls. For instance, a researcher in bioengineering stated that his research team could do more with their data if they had more CPU power. Another researcher informed interviewers that he recently had to beg for more computer time within his department. A researcher in fisheries explained that his

analysis work is limited by how long it takes to run computations—if the process exceeds 24 hours, they tend to run computations less frequently. Most notably, one researcher told interviewers about a donation of 100 cores from IBM, which sat in their boxes for nine months because his department did not have adequate electrical power to run them.

- **Network Access:** Some researchers reported that more bandwidth was always a need—one that would only continue to grow—although less than one-fifth of researchers specifically mentioned high-bandwidth network access as a requirement for their research. A few researchers cited specific needs related to network access. These researchers said that lack of bandwidth negatively impacts data transmission and analysis, collaboration, and remote access.

Many researchers saw communal resources as the primary solution to their computational needs. About one-third of the researchers in our analysis discussed communal solutions to their computational requirements as a need or interest; about one-fourth expressed interest in communal computing or centralized computing at UW. Several researchers thought UW could leverage its buying power and offer communal computing services for less money than it would cost researchers to maintain their own computing clusters or to hire an outside service. These researchers indicated they would rather focus on their research and not worry about purchasing, housing, supporting, and securing servers. Many researchers already paid other departments, institutions, or businesses to meet computing-power needs. Some also indicated that their computational needs tend to fluctuate over time, meaning they can often go for long periods of time without having a need. These hiatuses made researchers feel it was impractical for them to administer their own hardware—by the time they needed to use the equipment again, it would be outdated and underpowered. One researcher mentioned that he had looked at cloud computing services from Amazon and found them to be too expensive, but he felt that Amazon's services might be useful if he needed a lot of computational processing done in a very short time. (Encouraging the use of cloud services, where cost-effective and appropriate, is a key initiative of the UW eScience Institute.) A researcher in biostatistics said he would like UW Technology to provide fast, secure access to terabytes of data; he thinks UW Technology could offer a suite of cluster-computing environments designed and configured to provide package cycles that would meet the needs of the majority of campus research groups, similar to the timesharing service offered by campus technology units 30 years ago.

Data Collection and Analysis

While a relatively small number of researchers expressed needs related to data collection and analysis, the needs they did express echo a theme introduced earlier in this bulletin—the need for additional human and computational resources.

- **Collection:** Overall, researchers are making efforts to streamline their data-collection process. Many reported moving to electronic data collection exclusively, and some expressed the desire for UW to offer a cost-effective data-collection service. Most of the researchers we spoke with either already input and access data via the web or are trying to establish these systems.

Obstacles to creating online databases are both technological and personnel-related; researchers reported difficulty finding people with the appropriate "know-how" to assist them. Researchers also desired greater flexibility and capacity for gathering data in the field. Mobile and remote devices are becoming increasingly important to researchers for this purpose. A couple of researchers who were currently using mobile devices reported that they wanted to do more with the technology (conduct simple analyses or collect and store more data).

- **Analysis:** As the amount of data collected increases, so do researchers' data-analysis needs. Nearly all of the challenges researchers reported regarding analysis stemmed from the need to make sense of vast amounts of data. Here, researchers reported a need for specialized expertise—people with computational skills as well as familiarity with the research being conducted. One medical researcher reported that bioinformatics skills are a must in his field, but finding such expertise is a challenge. Another researcher, in immunology, explained the problem of finding programmers able to write customized analytical software to address frequently changing analytical needs. Several researchers also communicated needs related to the computing power required to run analyses. (This issue was addressed in an earlier section of this bulletin.)
- **Visualization, Modeling, and Simulation:** In about one-sixth of the interviews we analyzed, researchers mentioned using visualization, modeling, or simulation to display and analyze their data. A smaller number of these researchers expressed the desire for additional resources to meet needs specific to their projects. While the use of visualization and modeling does not presently appear widespread among those we interviewed, these needs may grow in the future as the volume and complexity of data collected continue to increase.

Communication and Collaboration

Researchers in these interviews often discussed day-to-day technologies that form the backbone of individual and collaborative research conducted at UW and with their national and international research partners. When multiple institutions and organizations collaborate, it is often easiest for researchers to use basic technologies, since sophisticated technological solutions may not be available to all partners. For basic communication and collaboration tasks, the needs expressed and usages described were common across a multitude of disciplines. The common technologies that researchers deemed to be essential were reliable phone and e-mail, wikis, tools for sharing work within and across teams and disciplines (such as websites, blogs, and UWTV), and remote desktop access.

In addition to the everyday technologies described above, several researchers, particularly those with extensive partnerships beyond UW, found technologies that allow real-time collaboration to be critical to their work. Here, however, there is more variance among researchers as to which technologies are being used. As researchers' technology use becomes more sophisticated (for instance, moving from teleconferencing to videoconferencing), their expressed needs for financial or technical support also increase.

Key technologies mentioned by these researchers included shared desktops and videoconferencing technologies of all flavors (one-to-one, one-to-many, many-to-many).

What It Means to Higher Education

We are at the dawn of a revolutionary new era of eScience. eScience will transform the process of discovery in all fields of science and engineering. In order to be successful in this new era, it is vitally important to identify and address researchers' evolving IT needs.

For centuries, there were only two modes of discovery: theory and experiment. In the past several decades, a third mode has risen to equal status: simulation. Like simulation, eScience relies on the extraordinary power of the digital computer. But in eScience, the focus is on *data* rather than computation. The data can come from simulation models, but it also can come from sensors—sensors that are deployed on the seafloor, embedded in buildings and roadways, built into in telescopes and gene sequencers, or implanted in living organisms (including ourselves!). The volume of data is overwhelming, and the challenge is to store, organize, mine, visualize, and interpret these data in order to extract knowledge. This, now layered with the fundamental challenge of understanding massively complex systems in general, lies at the heart of 21st-century discovery.

To support eScience, universities will need to consider diverse strategies. New faculty lines, research scientists, postdoctoral fellowships, graduate training grants, courses, outreach, computational infrastructure, and cloud services will foment a new environment with clear outcomes.

- **Competitiveness:** Above all else, cutting-edge eScience makes our research and educational environments highly competitive, both for the best students and faculty and for increasingly scarce federal research dollars. The students and postdocs participating in this program will be competitive in the marketplace.
- **Research:** There is no doubt that our research environment will both benefit from and require new enabling technologies. Too often we are challenged by massive data in complex systems without the capacity to gather, process, and understand it. The fact is that without an institutional capability to compete for and carry out eScience initiatives, university research enterprises will suffer greatly, even in the relatively near term.
- **Learning Outcomes:** A thread throughout our discussions was the goal of augmenting and enabling computing for researchers. While the information in this bulletin focuses on established researchers, it is also important to recognize the needs of researchers at all levels of experience, from entering undergraduates to our graduate and postdoctoral students. We envision a new generation of students, equally at home with computing as they are with core sciences. Anything less puts our competitiveness at risk. Participation in eScience involves mentoring experiences for students that are more similar to experiences gained in an industry or academic laboratories than to traditional graduate school classes. Problem solving, career discussions, and handling

communication between disparate disciplines are all natural parts of supporting team efforts in research.

- **Deliverables:** We imagine web-based tools that can be shared across higher education, connecting activities on one campus to similar ones developing elsewhere. In fact, this multi-institutional reach should become a model for how to engage our students in teams for research.

Key Questions to Ask

- How does our institution track the IT needs—current and future—of its leading researchers?
- To what degree are we meeting the IT needs of our researchers in various disciplines, and in which critical areas are we falling short?
- How well does our technology staff understand the institution's research and disciplinary directions and the IT implications?
- What potential resources, other than those currently in place, can be used to provide broad-based IT support for eScience and eScholarship?

Where to Learn More

- Atkins, Daniel E., Kelvin K. Droegemeier, Stuart I. Feldman, Hector Garcia-Molina, Michael L. Klein, David G. Messerschmitt, Paul Messina, Jeremiah P. Ostriker, and Margaret H. Wright. "Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure." Washington, DC: National Science Foundation, January 2003, http://www.nsf.gov/cise/sci/reports/atkins.pdf.
- Emmott, Stephen, and Stuart Rison, eds. "Towards 2020 Science." Microsoft Research, 2006, http://research.microsoft.com/towards2020science/.
- Indiana University Cyberinfrastructure Newsletter, http://racinfo.indiana.edu/newsletter/archives/2007-03.shtml.
- Katz, Richard, ed. *The Tower and the Cloud: Higher Education in the Age of Cloud Computing*. Boulder, CO: EDUCAUSE, 2008, http://www.educause.edu/thetowerandthecloud/133998.
- Klingenstein, Ken, Kevin Morooney, and Steve Olshansky. "Final Report: A Workshop on Effective Approaches to Campus Research Computing Cyberinfrastructure." Ann Arbor, MI: Internet2, 2006, http://middleware.internet2.edu/crcc/docs/internet2-crcc-report-200607.html.
- Montana State University, Center for Computational Biology. "NeuroSys Data Management System." http://neurosys.msu.montana.edu.

Endnotes

1. John V. Lombardi, Elizabeth D. Capaldi, and Craig W. Abbey, “The Top American Research Universities: 2007 Annual Report” (Tempe, AZ: The Center for Measuring University Performance), 8, http://mup.asu.edu/research2007.pdf.
2. Janice Fournier, Greg Koester, Cara Lane, and William Washington, “Conversations with the University of Washington’s Research Leaders: Preliminary Report” (University of Washington: Learning & Scholarly Technologies and eScience Institute, September 2008), http://escience.washington.edu/PI_Preliminary_Report_9-24-2008.pdf.

Acknowledgments

The authors would like to acknowledge more than 50 colleagues who contributed to this project: the project committee, the logistics team, and the interviewers.

About the Authors

Louis Fox (lfox@wiche.edu) is vice president for Technology & Innovation at Western Interstate Commission for Higher Education (WICHE) and previously was associate vice president for technology at the University of Washington; Ed Lazowska (lazowska@cs.washington.edu) is the Bill & Melinda Gates Chair of Computer Science & Engineering at the University of Washington; Cara Lane (cgiacomi@u.washington.edu) is a research scientist at the University of Washington; Janice Fournier (fournier@u.washington.edu) is a research scientist at the University of Washington; Greg Koester (koester@u.washington.edu) is a project manager at the University of Washington; William Washington (scumby@u.washington.edu) is an interaction designer at the University of Washington.

Copyright

Citation for This Work

Fox, Louis, Cara Lane, and Ed Lazowksa, with Janice Fournier, Greg Koester, and William Washington. “Information Technologies for eScience: A Preliminary Report from the University of Washington” (Research Bulletin, Issue 6). Boulder, CO: EDUCAUSE Center for Applied Research, 2009, available from http://www.educause.edu/ecar.