Developing Infrastructure for Informatics Research: Experiences and Challenges

Prem Paul, Vice Chancellor for Research and Economic Development
University of Nebraska-Lincoln

Scientific advances are generating massive amounts of data, fostering the need for new strategies to access, store, analyze, and mine data[1] [2]. The need to manage "big data" is especially acute in the life sciences, where genomes of a large number of species have been completed and efforts are underway to correlate genetic information with biological functions. Efforts are also underway to identify genes associated with health and disease. Similarly, large international collaborative experiments in physics, such as those conducted at CERN's Large Hadron Collider that recently resulted in the discovery of the Higgs boson particle, are generating large amounts of data and requiring high speed connectivity between laboratories to transfer data and to support high capacity data storage and analysis.

Most institutions are trying to deal with these challenges, which require major financial investments in infrastructure and personnel, resulting in significant economic pressures at a time when most institutions are facing budget cuts and federal funding is expected to flatten or be reduced. At the University of Nebraska-Lincoln (UNL), we recognized early on the need for enhanced cyberinfrastructure to support our researchers, and we initiated discussions on this important topic in 2005. We held an all-university workshop attended by 150 faculty members from the life and physical sciences, engineering, and the humanities. This paper summarizes our experiences, challenges, and plans for dealing with big data.

Advances in Life Sciences

Advances in nucleic acid sequencing technology have made it possible to sequence complete genomes of a large number of species. Information on thousands of genomes is now deposited in the National Center for Biotechnology database (see Table 1 for a summary of some of the major genomes).

This information makes it possible to determine biological functions coded by various genes and also to determine the significance of various genes relevant to health and disease. Nucleotide sequence data is being utilized to identify genes associated with cancer and other diseases and to develop novel therapies. At UNL, our faculty are working on plant and animal genetics and utilizing genomics in their research to improve productivity, particularly regarding traits for disease resistance and/or drought tolerance. These studies require major investments in computing and bioinformatics infrastructure and personnel trained in bioinformatics.

Cyberinfrastructure Needs in High Energy Physics

In 2005, UNL faculty identified the need for enhanced infrastructure for computing and connectivity. Our faculty competed for a National Science Foundation-funded Compact Muon Solenoid (CMS) Tier-2 site in high energy physics that utilizes data generated by CERN's Large Hadron Collider near Geneva, Switzerland, for research through the Department of Energy Fermilab. This research project required at least 10Gb Dark Fiber for data transfer and enhancement of computing infrastructure. We held a cyberinfrastructure workshop with experts from various funding agencies, including the DOE and National Science Foundation (NSF) and other institutions to learn about the importance and current state of cyberinfrastructure more broadly. NSF had published a blue ribbon cyberinfrastructure report[3] that was used as a background material. As a result of the workshop, we decided to invest in 10Gb Dark Fiber connecting Lincoln and Kansas City at a cost of over $1 million to connect with Internet2.

Species	Genome Size	Predicted Genes Coded
Arabidopsis	119 Mb	25,000 to 31,000
Fruit Fly	165 Mb	13,600
Mosquito	278 Mb	13,700
Rice	420 Mb	32,000
Corn	2300 Mb	32,000
Mouse	2500 Mb	23,000
Cow	3000 Mb	22,000
Human	3400 Mb	20,000 to 25,000

Table 1. Summary of major sequenced genomes

This gave our faculty the ability to be in a leadership position and transfer a record amount of data from Fermilab. UNL's leading capability was demonstrated at a national Internet2 meeting in 2007. The cyberinfrastructure we developed to manage big data has also been very helpful in supporting our faculty who require supercomputers. For example, leveraging these computing resources, Professor Xiao Cheng Zeng in UNL's Department of Chemistry has made several major discoveries published in top-tier journals like the *Proceedings of the National Academy of Sciences* [4,5].

UNL's Center for Digital Research in the Humanities has been a leader in the digitization of scholarly material related to Walt Whitman and the Civil War. The group of faculty in this center have so successfully competed for grants from the National Endowment for Humanities that they are recognized as national leaders in this area.

Bioinformatics Experience

One area in which we have made significant investments is bioinformatics. We have hired several faculty members during the last decade with bioinformatics expertise. They have been very successful scholars and have been extramurally funded with grants from the DOE, National Institutes of Health, NSF, and U.S. Department of Agriculture. However, they are pursuing their own scholarship and research agenda and are not able to provide bioinformatics service to others. A large number of faculty in the life sciences who do not have a background in bioinformatics need help with the analysis of sequence data, and are having a difficult time finding experts to do their work. There also is a shortage of talented people trained in bioinformatics.

We have a bioinformatics service in our Center for Biotechnology core facility; however, the facility's staffing is not sufficient to meet the needs of all faculty because there are more users than talented experts. In our experience, life scientists want bioinformatics experts to analyze their data – much in the same way statisticians have contributed to research programs for decades. However, the majority of the bioinformatics experts want to pursue their scholarship and advance the bioinformatics field. Several research groups have created their own bioinformatics core facility, including their own computer clusters, rather than using supercomputers. We are exploring ways to add bioinformatics staff in the core facility and hire additional bioinformatics faculty, including a leader who can coordinate bioinformatics resources and services across campus.

Big Data Needs in the Social and Behavioral Sciences

There are also needs for access to big data and cyberinfrastructure to address important questions in the social and behavioral sciences. UNL has significant strengths in social and behavioral sciences, including the Gallup Research Center, which conducts research and trains graduate students in survey methodology. The Bureau of Business Research provides relevant information and insightful data on economic conditions across Nebraska, the Great Plains, and the nation. UNL's Bureau of Sociological Research and our Survey, Statistics, and Psychometrics Core Facility provide services to faculty in survey methodology and research. The University of Nebraska Public Policy Center provides the opportunity for policy makers and researchers to work together to address the challenges of local, state, and federal policy.

We also have strong programs in substance abuse and health disparities in minority populations. Though each program is highly successful, there is an opportunity for strengthening these programs through collaborations. This is critical, especially considering the importance of social and behavioral sciences in major societal challenges pertaining to food security, water security, national security, economic security, national competitiveness, and energy security. Therefore, we have launched a taskforce to better understand our institutional strengths and needs for infrastructure.

UNL faculty members have recognized the need for a Research Data Center (RDC) to access economic, demographic, health statistics, and census data. RDCs are run through the Census Bureau and NSF. Currently, there are 14 RDCs managed by the Census Bureau. RDCs provide secure access to restricted use of microdata for statistical purposes. Qualified researchers prepare proposals for approvals by the Census Bureau. Following approval, work is conducted in secure facilities where scientists may access centrally held data.

Current RDC locations include: Ann Arbor, MI; Atlanta, GA; Boston, MA; Berkeley, CA; Chicago, IL; College Station, TX; Ithaca, NY; Raleigh, NC; Stanford, CA; Washington, D.C.; Minneapolis, MN; New York, NY; and Seattle, WA. Unfortunately, there are no RDCs in the Midwest; the center nearest to Nebraska is in Minnesota.

Since RDCs are expensive to maintain and require hiring a director that is a Census Bureau employee, it might be more appropriate to pursue a regional RDC that could serve universities in Nebraska, Iowa, Kansas, and Missouri. Based on conversations with Census Bureau personnel, such an RDC would comprise secure space, including workstations for faculty and students to access data for research. We propose to build such a center at UNL that would be available to our regional partners. Access will be facilitated through proposals that are peer-reviewed by an advisory board, as required by the Census Bureau protocols.

Several years ago at the Merrill Conference, discussion took place regarding what we can do together that we cannot do alone – especially with regard to creating shared research infrastructure to support large-scale research projects and programs. The RDC concept represents such an idea for regional collaboration to access big data in social and behavioral science research.

References Cited

1. T. Hey, S. Tansley, and K. Tolle (Eds). 2009. *The Fourth Paradigm: Date-Intensive Scientific Discovery*. Microsoft Research: Redmond, Washington.
2. P.C. Zikopoulos , C. Eaton, D. deRoos, T. Deutsch, and G. Lapis. 2012. *Understanding Big Data: Analytics for Enterprise Class, Hadoop and Streaming Data*. McGraw Hill.
3. D.E. Atkins, K.K. Droegemeier, S.I. Feldman, H. Garcia-Molina, M.L. Klein, D.G. Messerschmitt, P. Messina, J.P. Ostriker, and M.H. Wright. 2003. "Final report of the NSF Blue Ribbon Advisory Panel on Cyberinfrastructure: Revolutionizing Science and Engineering through Cyberinfrastructure." Available at www.nsf.gov/od/oci/reports/toc.jsp.
4. J. Bai, C.A. Angell, and X.C. Zeng. 2010. "Guest-free monolayer clathrate: coexistence and phase transition between two-dimensional low-density and high-density ice." *Proceedings of the National Academy of Sciences USA*: 107, 5718-5722.
5. T. Koishi, K. Yasuoka, S. Fujikawa, T. Ebisuzaki, and X.C. Zeng. 2009. "Coexistence and Transition between Cassie and Wenzel State on Pillared Hydrophobic Surface." *Proceedings of the National Academy of Sciences USA*: 106, 8435-8440.