Supporting Data-Intensive Collaboration via Campus eScience Centers

Bill Howe
billhowe@cs.washington.edu

Cecilia Aragon
aragon@uw.edu

David Beck
dacb@u.washington.edu

Jeffrey P. Gardner
jeffpg@uw.edu

Ed Lazowska
lazowska@cs.washington.edu

Tanya McEwen
thmcewen@gmail.com

Science is being transformed by the democratization of automated, high-throughput data acquisition technology: even small labs and individual researchers now have ready access to DNA sequencers, high-resolution simulations of the Earth, satellite imagery, terabytes of telescope imagery, and a national network of environmental sensors. But there has not been a commensurate democratization of automated, high-throughput data analysis technology. Moreover, the new collaboration models afforded by these data resources have not been fully explored.

Universities nationwide have begun to establish campus organizations to explore these needs, with missions to advance both the research and practice of data-intensive and collaboration-intensive science – eScience – characterized by an explosion of data volume and complexity, significant interdisciplinary collaboration, emphasis on exploratory (as well as hypothesis-driven) discovery, and an acute need for new and newly applied algorithms, software, and cyberinfrastructure. These *eScience Centers* (eSCs) may offer a variety of services: software development, proposal authoring, computational facilities, platform matchmaking, hiring advice,funding for campus eScience projects, collaboration services (both software and organizational), consulting, algorithm design, incubation of new technology. eSCsare also characterized by a mission to conduct basic and applied research in eScience techniques and technologies stemming from computer science, computational science, statistics, and applied mathematics. Finally, eSCsemphasizeeducation and outreachfor a new generation of cross-discipline researcher, possessing depth in at least*two* areas: a domain science as well as a domain-independent technical discipline. We believe these eScience-savvy researchers will represent the core of the scientific workforce over the next 10 years, and that eSCs will be crucially responsible for training them.

We differentiate these eSCs from Campus Cyberinfrastructure organizations (c.f., http://www.educause.edu/CCI) in three focus areas: data, research, and interdisciplinary training. By *data*, we intend an emphasis on large-scale and complex data analytics – algorithms, visualization, machine learning, sensor networks, cloud computing – as opposed to conventional IT infrastructure provisioning. By *research*, we intend a peer, collaborative role with campus labs as opposed to purely a service role. Materially, we find eSCs publish actively in both technical venues (CS, statistics), domain venues, and interdisciplinary venues (e.g., Bioinformatics), and that this publication record is incorporated into their mission and performance review. By *interdisciplinary training*, we intend a long-term goal of producing a new kind of researcher with a strong technical and domain background (as opposed to either training engineers to support science, or teaching scientists programming skills.)

While the motivation for eSCs and the work they do is widely recognized, the appropriate organizational structures and funding models have been developed largely independently with little communication or sharing of ideas: There has been some discussion of "what to do" but not enough on "how to do it."

This brief position paper is intended to consider the role of eSCs for data-intensive collaboration and inform a discussion of best practices for mission, governance, and sustainability. To initiate a discussion, we provide questions rather than answers in each of these three topic areas, followed by an overview of the UW eScience Institute.

MISSION

- How should research activities be balanced by the need to produce effective, robust, and well-supported tools for researchers?
- Who sets the research agenda of an eSC? Domain science stakeholders or those in core technical discipline?
- What software services, if any, should an eSC provide to its home university? (source control, data analytics, visualization, storage/archive, cloud "on ramps", research "hub" services, collaboration tools)
- What kind of staff, with what skillsets, should an eSC possess? What does the ideal "eScientist" look like?
- ShouldeSCs hire postdocs? Research scientists? Are there faculty appointments? Do eSC faculty advise students in a home department or through the eSC?
- ShouldeSCs offer degrees apart from a home college or department? Certificates?

GOVERNANCE

- How should eSCs balance the goals of the home university with participationin the larger eScience community?
- Does it make sense to have a national body representing eSCs? What roles should this organization undertake?
- Should the eSC staff be loosely coupled (like a department) or more structured (like a software shop)?
- Can an eSC function as a virtual organization, or is a physical facility needed?
- If a physical location is deemed necessary, university libraries may have a centrally located surplus of of space and a related mission – could housing an eSC in the library be a repeatable "best practice"?
- How should eSCs coordinate activities with campus IT? What is the relationship between eSCcyberinfrastructure activities and campus IT? (Prototype vs. production?)
- What is the relationship

SUSTAINABILITY

- What kind of funding models are currently used to support eSCs? Do researchers pay for eSC services directly through recharge mechanisms? Do eSCs write joint proposals with researchers in other disciplines?
- How should an eSC be measured? (Publications? Satisfied customers? External funding?)
- What is the relationship between an eSC and other domain-independent technical disciplines? (Computer Science, Statistics, Applied Mathematics)
- How can an eSC attract, create, and retain top talent in both the sciences and the technical disciplines?
- Part of the mission of an eSC is typically to provide production deployments of cyberinfrastructure, software, and algorithms. Development and maintenance of such robust tools requires professional engineering staff. Can an eSC hope to compete with market salaries? Can an eSC attract and retain staff at lower salaries by providing a rewarding environment and an interesting set of problems?

THE UW ESCIENCE INSTITUTE

Over the last several years, and with remarkable consistency, various local and national working groups and reports, as well as less formal conversations with faculty from the University of Washington and elsewhere, have all told the same story. Faculty are drowning in data, and they need help and expertise with capturing, managing and mining all this data. They need access to computational and intellectual infrastructure to allow them to focus on the science and reduce the "tax" of working with large and complex datasets using advanced computational techniques.

In January 2008, the University of Washington responded by creating the eScience Institute, and in July operational funding was added to the state supplemental budget. Since then, the institute has hewed closely to the original mission and strategy outlined by Director Ed Lazowska, through:

- Bootstrapping a cadre of research scientists with data management expertise
- Adding faculty in key fields to assist researchers with technology and data issues, and
- Increasing the sharing of expertise and facilities.

This team of research scientists and faculty has since applied their expertise in successful collaborations in a host of fields such as Astronomy, Biochemistry, Bioengineering, Oceanography, and Physics. Additionally, the eScience Institute has initiated ongoing efforts to support the sharing of facilities including Amazon Web Services, Microsoft Azure, Google cloud services, national computational resources, campus data centers, and shared, on-premise high-end computing and storage.

The eScience Institute team is making available Web-based tools to store, compare, analyze, and share data as well as a host of online case studies of how scientists are using the techniques and technologies of eScience in their research, guides offering techniques and technology options for achieving research goals, and how-to documents for using particular technologies. Team members also regularly provide assistance with the application of computational methods, tools and other resources to data-intensive science and research effort.

The eScience Institute is led by Ed Lazowska, Professor and Bill & Melinda Gates Chair of Computer Science & Engineering, with leadership and administrative oversight from the eScience Steering Committee assembled from domain stakeholders around theUW campus.