Data Sciences Hub Proposal

datascience.wisc.edu

Michael Ferris and Brian Yandell

GOAL: Establish WID as a hub for data science at UW-Madison, with the purpose of integrating and coordinating data science activities across campus, and fostering fundamental research, teaching and outreach under its own aegis. Stake out an international leadership role in key facets of data science.

Big data, however defined, is a disruptive force in society, no less so in academia. Most campus units are affected by big data influx, or by the potential of gathering and examining data at unprecedented scales. Many individuals feel UW needs to address this somehow, but the question is how to do that in an efficacious manner. Achieving this goal will require significant investment of time and resources.

We place this challenge in a larger context. This document outlines the concept of a Data Sciences Hub, its rationale and a proposed near-term plan for implementation, setting the stage for future development. This document was initiated by a small WID-based team, recognizing that many groups on campus have been thinking and doing work along similar lines. The intent is to combine efforts in the spirit of the **Wisconsin Idea**, that education should influence people's lives beyond the classroom.

Data Sciences Hub Concept

The Data Sciences Hub (DSHub) will have a strong focus on performing and enabling research in Data Science, broadly defined, and will provide a focal point for campus research, education, and outreach in the area. The broad organizational structure of the DSHub is shown in diagrams below. An individual or team likely connects to DSHub with a problem, hoping to create a product (research results, course module, web deployment, etc.). A collaboration develops to understand the problem and identify the best way to integrate DSHub expertise into the team to achieve desired outcomes.

The Data Sciences Hub will provide a focal point for programs dedicated to research and application of modern techniques to the management, storage, and analysis of complex data sets. In addition to fundamental research activities on data science, the Hub will include campus-wide discussions of data problems, consulting services to help researchers apply the tools of data analysis and computation, provide links to educational activities in this area, and organize a Public Lecture series on Data Science and the SILO seminar series. The Data Sciences Hub will house a Data Analytics Integration Service as well as a Data Science Consulting Facility, which will collaborate in applying computational, modeling, and statistical analysis to domain-specific data projects.
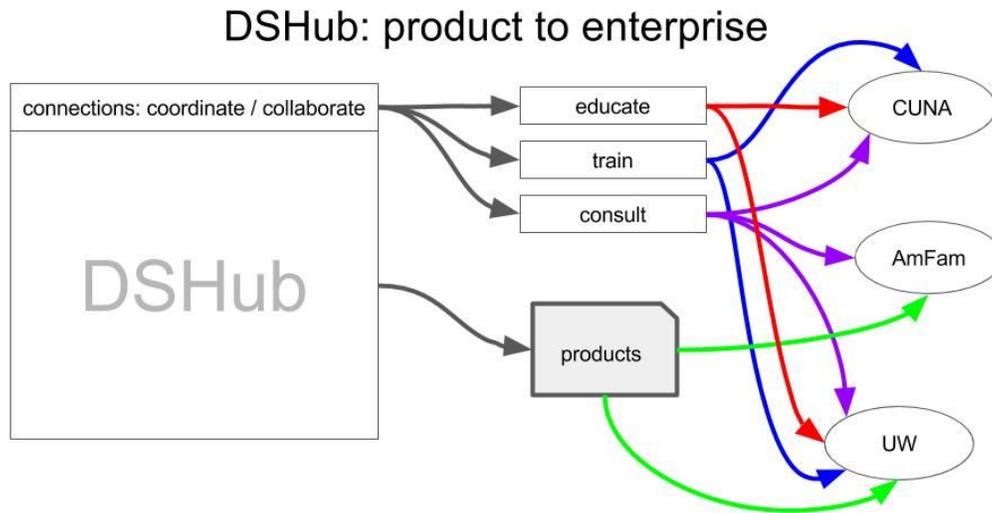
Taking this further, this DSHub will interact with enterprises, including UW-Madison as shown below. Needs may involve education programs, staff training, problem consultation, or creation of products. Training may also be an effective way to further engage industrial collaborations.

DSHub: product to enterprise

Rationale

Science is a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions about the universe. Many domains at UW, including those in the humanities, are building and organizing knowledge from big data, and finding new ways to generalize the extraction of knowledge from that data. That is, they are doing **data science**. Here at UW, the concept of many data sciences coming together in some way is emerging. This seems to be why this Data Sciences Hub (DSHub) has some traction.

Many other academic institutions are organizing their own department or center or institute or program for Data Science (perhaps with another name). While we have much to learn from these models, it is important to remember what George Box said: "All models are wrong, but some models are useful." We can, and should, improve on these early approaches. Our goal should not be to "catch up" with other institutions, or to be "better". Our goal is to be great at what we do best as an institution that values diverse, distributed leadership, and to leverage our strengths to enable all endeavors to excel.

Guiding principles

Before jumping into details of a Data Sciences Hub (or whatever it might be called), we must be clear on guiding principles. As a proposal, please consider the following:

- Cooperation: working together toward common, or at least complementary, goals.
- Inclusion: providing opportunities and mechanism to include all members of the campus community in discussion, actions and access.
- Equity: striving for fairness in how individuals are treated, integrity in research, and equitable ways to level the playing field for all.
- Diversity: valuing a wide palette of ideas, approaches and perspectives.

This "DSHub" should be common ground, a safe place for people to come together to discuss and make progress on all things data. And it should have a welcoming physical place, as well as a robust electronic presence. The goal is to cut across traditional silos, setting aside egos to address a larger need. DSHub would focus on the process of data science, enabling people to tell data-rich stories, recognizing that context is central to properly understanding big data.

Proposal Details

The following sections propose, in outline, strategies, actions and resources needed for our near-term plans, setting the stage for future development. The "sandwich" diagram shows more detail about some proposed DSHub components and connections to campus governance:
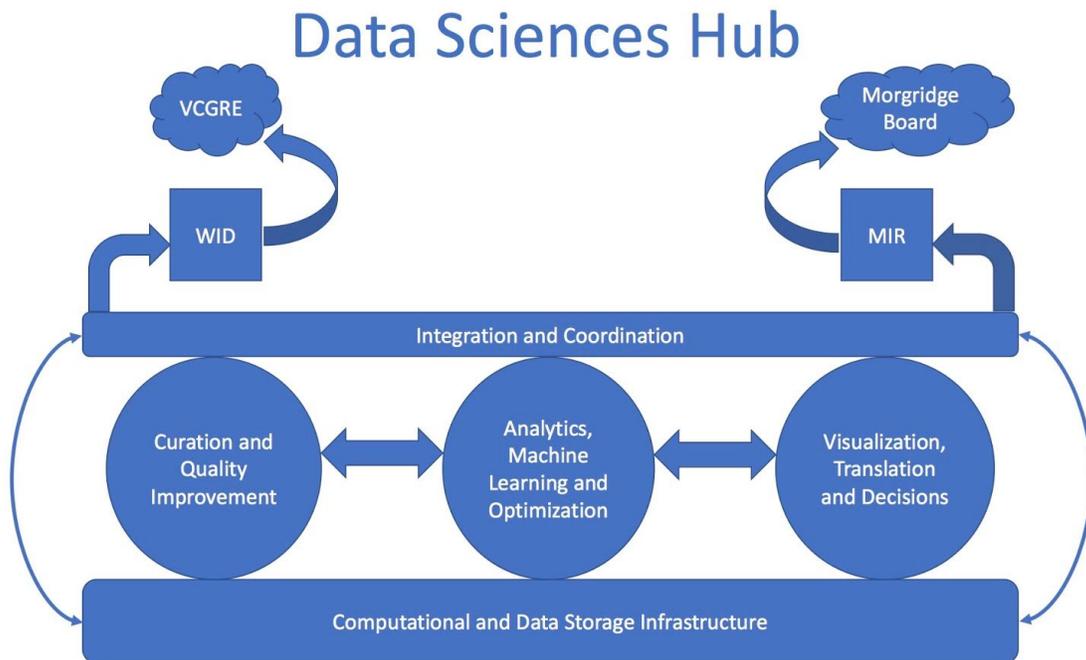
Strategies

1. Build on existing successes from WID themes (especially the Optimization theme), the SILO (Systems, Information, Learning, and Optimization) seminar series and workshops, Core Computational Infrastructure, and collaboration success in the Biometry Consulting Facility and Biostatistics & Medical Informatics.

2. Organize data science activities around three complementary areas:
   a. Mathematical foundations of data science. Modeling, algorithms, optimization, machine learning, computational statistics.
   b. Systems aspects of data science: database systems, data cleaning, data management, data integration, data visualization, computational technology.
   c. Collaborations with domain research people across campus, including health sciences, energy, agriculture and environmental sciences, education and social sciences.

3. Collaborate with others at UW to develop and provide the underlying data science infrastructure/tools for UW scientists in an R1 research institution. The DSHub will foster development of stable software systems that make state-of-the-art data science tools and methods easy for practitioners to use. Such tools are critical to facilitate transition of our research into practice. The aim is to obtain external funding for one major center in at least one key aspect of data sciences. For example, a team of 14 has recently submitted a proposal to NSF's new TRIPODS program for an "Institute for the Foundations of Data Science." We will pursue other similar opportunities as appropriate.

4. Provide a forum to advertise the broad educational activities in data sciences across campus. Collaborate with other UW faculty and UW Departments to develop Data Science education resources for the campus community. Extend some data science courses from different departments, with the design/development of these courses utilizing the DSHub. Engaging in

smaller group-defined teaching activities, such as the NSF NRT LUCID (https://lucid.wisc.edu), will provide additional mechanisms for education.

5. Training individuals who work with big data is a crucial process for success moving forward. The big data landscape is changing rapidly, requiring individuals to develop many competencies about tool use and ways to communicate ideas and results. People need training in how to work effectively in teams, using reproducible research principles to share emerging approaches. Project leaders need to learn how to build and evolve teams that adapt to changing needs. Big data often requires teams to learn how to maintain data confidentiality. Such training can be leveraged by research, teaching and outreach.

6. Develop campus-level consulting access to foster and help cross-disciplinary collaboration in research and teaching. It will leverage and build on successful models of the Biometry Consulting Facility and BMI-related facilities, including the Cancer ISR, CPCP and the Bioinformatics RC, to a more general campus facility serving all of campus.

7. Expand current industrial partnerships, such as the Optimization Research Consortium and the SILO Seminar sponsorship, to include a broader range of Data Science partnerships, and hold an annual Data Science Research Consortium Day at the WID.

8. Organize WID Public Lectures in Data Sciences, inviting high-profile external speakers including renowned researchers and senior figures in the major data companies.

9. Establish visitor programs in WID, including a visiting professorship in data science (usually to be held by a distinguished colleague on sabbatical) and one-year PhD student exchange programs with targeted institutions. These programs will promote new interactions and expertise beyond our group.

10. Facilitate joint graduate student recruiting in data science. Interested students enter through CS, ECE, Statistics, Mathematics, Information and other programs. We could arrange for all such student to visit on common dates, probably overlapping with the CS visit weekend, for discussions in WID with faculty and students in data sciences.

Action Items

1. Run the meeting "Towards a Strategy for Data Science at UW" at WID on 6/21/17, inviting key players from around campus. The goal will be to share information about various current and planned initiatives in data science-related research and education, and organize, strategize, and set the agenda for future activities, to the mutual benefit of all involved. We will plan the meeting carefully to maximize the chances of productivity and effective follow up. It will include panels, short invited talks and presentations, small group discussions.

2. Establish a core leadership team for the DSHub that will be responsible for coordination of DSHub activities and collaboration with the campus community and beyond.

3. Hire/engage DSHub staff, with aim to demonstrate functionality and specific competency in the major components of the Data Sciences Hub -- curation, analysis, and visualization. These staff members will provide end-to-end data to decisions integration capabilities for UW researchers. Much of the required work needs expertise that is not in the skillset (or time constraints) of a faculty member, but requires a level of programming skill and familiarity with a suite of computational tools that need (permanent) skilled support staff.

4. Develop interconnections among campus units and programs involved in research, teaching and outreach in the data sciences. This will require some staff with strong communication skills and the ability to adapt to changing needs, engage with campus individuals about problems and projects, and connect with DSHub staff on technical planning.

5. Determine computational and infrastructure needs, and processes to facilitate these.

6. Determine two to three core projects that would leverage the DSHub, identify leaders/proponents of those projects and ensure the capabilities of the DSHub facilitate advances in these projects.

7. Establish funding for the resources needed below and to provide service guarantees from the hub.

Investment for Adoption
1. Publically accessible space for the DSHub, as open-plan as possible, including space for in-house integrators, a meeting room for 10-20 people and small-group meeting rooms for 3 people.   The space is for specific application focussed postdocs or more outward facing workers in the group, and will provide coupling space to external collaborations.
2. Funding for core staff that augment the research and project-driven capabilities of the DSHub. Enabling connections and building infrastructure within the three core areas of the hub will require a staff of at least 6. (Expertise in data generation, cleaning, integration, wrangling to improve the data quality, parallel computation and algorithm design, natural language processing, machine learning and optimization processes, statistical inference, privacy and security, visualization and translational tools, data management and planning etc).
3. Funding a Coordinator position. This person would help to organize the teaching and infrastructure resources available on campus, be a visible point of contact and facilitatory tasks, and facilitate information flow among UW researchers.
4. Funding for WID Public Lecture Series. This could be a naming opportunity but will probably need seed money to initiate.
5. Endowment for additional graduate student positions in WID with emphasis on data science. Students will be admitted to existing departments, but will have funding stream from the area to allow targeted recruitment and training.   These students could be engaged for short periods within the Consulting and Data Integration Service. The funding could also be used for the recruiting visits.
6. Matching funding for new grant proposals, as appropriate to each particular call. Possible use of such matching funds could be to augment the pool of research assistants working in this area, and provision of infrastructure needed for competitive environment demonstrations.
7. Endowment for visiting professorship in data science.   This would be used to attract leading researchers in the area for half or one year visits to WID.