WASHINGTON UNIVERSITY MCKELVEY SCHOOL OF ENGINEERING

STRATEGIC PLAN TO ACHIEVE LEADERSHIP THROUGH EXCELLENCE

RESEARCH

Data Sciences for Humanity

Data Sciences for Humanity

Executive Summary

Transformation Through Information

Computing and data are playing increasingly central roles across disciplines, creating an enormous opportunity, coupled with exciting intellectual challenges. The opportunity comes from the confluence of remarkable progress in algorithmic and computing capabilities and the ubiquitous availability of extensive datasets capturing many aspects of human life, social behavior, and scientific discovery. The challenge comes from the need to use the data to answer substantive questions about understanding human behavior in the context of the world around us, and harness that understanding to improve the human condition.

While there are many initiatives in "Data Science/Big Data," WashU is well positioned to leverage a combination of strengths to develop a unique, two-pronged initiative in Data Sciences for Humanity (DSH). The initiative will focus on applying data sciences and computing to a range of societal problems, and, conversely, on making the solutions to those problems more readily accessible to their target audience. The initiative would take advantage of WashU's existing strengths across disciplines, as embodied in world-class programs in social work, medicine, psychological and brain sciences, political science, economics, and business, among others. It would combine those strengths with existing and expanding expertise in core data science, as well as in human-computer interaction. The initiative would be synergistic with an ongoing effort to develop an interdisciplinary graduate program made possible by a new Division of Computing and Data Sciences that cuts across school boundaries.

The DSH initiative is expected to attract new funding from across a diverse set of sources. It should give rise to new interdisciplinary centers that will take advantage of the availability of a critical mass of expertise and the presence of a structure specifically intended to foster interdisciplinary collaborations. It also will appeal to a new brand of top-notch PhD students who realize the importance of data and computing, but whose interests lie not just in developing new data science tools, but also in applying them to tackle substantively important societal questions. Finally, such an initiative will create numerous opportunities to engage undergraduate students in research activities and more broadly enrich their curriculum.

Many of the core ingredients for realizing such an initiative are in place, but successfully executing this vision calls for an injection of additional resources. This includes faculty hiring to create greater critical mass in several key areas, e.g., natural language processing, causal inference, differential privacy, and human-computer interaction, while opportunistically continuing to strengthen our core machine-learning expertise. There are also challenges to overcome in instituting the interdisciplinary graduate programs necessary to support such an initiative. Catalytic investments will yield significant dividends in terms of increased external visibility, ability to attract new funding, and growth in the university's reputation.

Computing and data are increasingly fundamental across almost every discipline. The School should invest in an initiative in Data Sciences for Humanity. This initiative would include two main thrusts:

Data science for society: Taking a data science-oriented approach to improving the human condition. This encompasses problems from disciplines such as social work, public health, political science, psychology, economics, and medicine for which data is being collected but not used to its full potential. This thrust will explore real-world problems in close collaboration with domain experts from all schools across WashU while contributing new domain knowledge for the problem areas and new general techniques in machine learning, causal inference, natural language processing, etc.

Broadening access to data science: Supporting end user access to machine learning and data-science tools. To have the greatest potential impact, data science must reach the hands of domain experts rather than just computer scientists. Yet, most data science initiatives have neglected to consider how to develop data-science tools that can be used broadly. This thrust will explore the ways in which experts in noncomputing fields want to use data. This insight will develop new tools that empower noncomputer-scientists to make effective use of data science independently and identify the need for new techniques in fields like machine learning, visualization, and data exploration.

The research should be closely coupled with an interdisciplinary graduate program made possible by a new Division of Computing & Data Sciences (DCDS) that cuts across school boundaries. In addition, it will inform undergraduate curriculum development.

Background

Data science is recognized as an area of fundamental importance, and many universities have already made significant investments in the area. Given the breadth of fields that data science has the potential to change dramatically, it is strategically important for WashU to contribute. However, we need a way to distinguish ourselves that builds on our unique strengths. To that end, we propose a human-centered approach to data science including both data science for social good and enabling technologies that would allow for broader use of data science tools by stakeholder communities. The School of Engineering & Applied Science (SEAS) already houses strong core expertise in machine learning and data mining. Although this core expertise needs further strengthening (in particular in natural language processing), it offers a strong anchor and pool of technical expertise that can be brought to bear on a variety of problems. The opportunity for WashU lies in coupling this expertise with strengths in several application areas where WashU has a leadership position, e.g., through the Brown School, the School of Medicine, the Olin Business School, and various departments in Arts & Sciences.

The intellectual opportunity is fascinating. Computing and data science has the potential to transform many disciplines, perhaps more dramatically than the introduction of the standard statistical analysis and hypothesis-testing toolbox did in the mid-20th century. The questions that are raised by data and are generated by and about human behavior are engaging and profound. However, many, if not most, of these questions can only be tackled using a multi-disciplinary approach that combines deep knowledge of the capabilities and operation of data science techniques, with the domain expertise needed to apply them effectively to the problems under consideration. For example, consider the following kinds of questions we may be able to answer by tapping into the combined expertise of partners from computer science, public health and social work, political science, psychology, medicine, economics, etc. — all areas in which WashU has significant strengths:

1. How do we discover and validate causal patterns in large datasets of human social behavior?
2. How is language used most effectively for persuasion across different types of social media on the Internet?
3. How can we use data from social service agencies to improve the provision of services to the homeless?
4. How can we incentivize the provision of private data (energy or road usage, for example) that could help improve resource allocation in modern smart and connected communities? How can we secure such data and keep it private?
5. What are the legal, economic, and ethical factors that must be taken into account when using algorithms to make decisions with real social impact on the basis of such data?
6. How can patient records be mined to develop personalized treatments with better outcomes without revealing sensitive private information?

Equally important are the questions of how we develop the right tools to democratize data science (for example, data exploration and visualization software, or automatic machine learning pipelines for risk-assessment tasks) and make these tools available to end users, whether researchers in other disciplines, practitioners in industry, or others.

The Strategic Opportunity

A school's visibility is largely determined by its flagship programs, which ideally merge scientific excellence with distinctive approaches. A Data Science for Humanity initiative has the potential to become such a flagship program, leveraging existing strengths and identifying promising directions for future research and growth. Similarly, a new Division of Computing & Data Sciences (DCDS), specifically set up to enable interdisciplinary studies, could attract top-notch PhD students and make WashU a known center of excellence in an important new interdisciplinary idea. The initiative could also help establish our presence in subareas of computer science, particularly in computational social science and in human-computer interaction (HCI), where few schools have extremely large groups, and a group of four or five faculty can have a significant impact.

In addition to leveraging strengths in the Department of Computer Science & Engineering (with 15 current faculty working across the areas of machine learning, HCI, parallel computing, and cyber-physical systems) and Electrical & Systems Engineering, the initiative would also build on existing communities around campus. In particular, several schools and departments are already involved with planning for DCDS (the Brown School and the departments of Psychological & Brain Sciences, and Political Science). In addition, there are significant opportunities to work with clinical data and practitioners at the medical school, including through the newly formed Institute for Informatics (diversifying the current research interactions between SEAS and the medical school). The Olin Business School is another natural fit, and of course, there are researchers in virtually every discipline who would like better ways of working with their data.

Pursuing the Research Opportunity

Enabling the development of a truly interdisciplinary research agenda is the primary goal for DCDS, which should eventually become the coordinating home for research that takes a data- or computation- enabled approach to problems in any domain. We envision that the Division will serve as a virtual and physical home for interdisciplinary research efforts in this area, while also enabling the development of several specific research centers. Realizing the vision behind DCDS calls for delivering on several important **short-term** goals:

1. **DCDS launch:** This involves developing plans for the structure and curriculum of interdisciplinary PhD programs in DCDS in collaboration with the Brown School and the departments of Political Science and Psychological & Brain Sciences in Arts and Sciences. DCDS needs to recruit strong initial cohorts of PhD students and to ensure the cohesiveness of the program and the development of appropriate programming (seminars, events, and workshops).

2. **Development of DCDS as an umbrella organization and research facilitator:** There are already a variety of research efforts around the university that could be categorized as Data Science for Humanity. However, most are in silos in individual departments. DCDS should identify existing efforts and curate information about those efforts to raise awareness both within the university and outside. DCDS must also facilitate new research connections and collaborations across the university. We propose to do this in two ways: 1) by running a small grants program similar to the URSA (University Research Strategic Alliance) grants that could be used to start a new project and develop initial results for use in a future grant proposal and 2) by hosting a Data Science for Humanity workshop at WashU.

3. **Recruiting an advisory board:** Senior researchers from institutions outside WashU could help to identify strengths and opportunities to develop our expertise in human-centered data science. Additionally, those same senior researchers could help to ensure our early success by spreading the word about our efforts both to colleagues and to students.

4. **Faculty hiring:** While we have significant existing strengths in human-centered data science, we believe that further investments will be necessary, particularly with respect to faculty hiring. In the short term, it is important to hire in natural language processing, causal inference, differential privacy, and human-computer interaction (in particular with a focus on enabling non-computer scientists to interface more effectively with data science tools), while opportunistically continuing to strengthen our core machine learning expertise.

In the longer term, our goals should extend to:

1. **Developing a pipeline of top-notch PhD students:** The initial efforts to launch Data Science for Humanity through DCDS will make WashU attractive to students wanting to explore the impact of data science on particular fields. DCDS will provide a programmatic home for this. However, to further aid in recruiting students and supporting the interdisciplinary nature of the research, which will call for students to complement their original training, applying for training grants (e.g., through the NSF National Research Traineeship program [NRT] or through NIH) will be an important component in developing and sustaining a vibrant research enterprise.

2. **Enabling faculty to develop one or more new research centers:** While launching centers will likely call for additional, targeted investments, establishing a home and a community around Data Science for Humanity will help foster new collaborations and identify common themes and challenges across projects. Success in developing strong research initiatives in these areas should in turn coalesce into several interdisciplinary centers, for example, centers in:

a. HCI for democratizing machine learning by developing interfaces that make data sciences tools more useful to domain experts, with a focus on actionable outputs and revealing domain-specific causal relationships.

b. Data-enabled approaches to problems in public health and the provision of social services, e.g. understanding the impact and spread of health-related information and misinformation (on vaccination, smoking, etc.) on social media, and quantifying the benefits of different interventions on homelessness risk.

c. Smart and connected communities, both in urban settings as embodied in the smart city concept, or in more rural environments as in precision agriculture. The latter could leverage collaborations with local companies, e.g., Monsanto.

d. Personalized medicine with a focus on combining all sources of information (including clinical and administrative data) about individual patients to generate more effective treatments.

e. Dynamic data analysis, extending the successes of temporal modeling and systems theoretic analysis in domains like neuroscience to the social sciences.

3. **Broaden DCDS:** While we will launch with a few carefully selected partners outside SEAS, we envision the DCDS umbrella extending to incorporate other partner departments and schools in the future, including the Olin Business School, the economics department, the School of Medicine, and departments in Arts & Sciences interested in digital humanities.

4. **Create opportunities for undergraduates:** While the initial thrust of our efforts should lie in the research opportunity and PhD program, we should take the associated opportunity to develop undergraduate programs. In addition to new "CS + X" majors, we should explore introducing a new undergraduate major in data science. In addition to curricular offerings, students are likely to find many project options both within the university and in the broader St. Louis area. In addition to providing experience, these projects would also support research and our pipeline of PhD students. Another opportunity would be to create an expanded REU-style summer program in the area. The CSE department already operates such a program that brings in students both from WashU and from other institutions around the country, and has been successful for recruiting PhD students.

Challenges

While other universities have initiated a range of initiatives in the broader space of computation- and data-enabled approaches to problems with social impact, there are few that combine the structure of a graduate program and research focus across human-centered domains. One such program is MIT's recently launched Institute for Data, Systems, and Society, and PhD program in Social and Engineering Systems. Another is Penn's Warren Center for Network and Data Sciences, a loosely connected set of graduate programs and faculty that seeks "to foster research and innovation in interconnected social, economic and technological systems," and has as one of its goals to "build an interdisciplinary team of researchers, students and entrepreneurs."

Others of some relevance, but with significant differences, include highly-focused graduate programs (like Penn State's Big Data Social Science graduate program, an IGERT-funded program focused mostly on Political Science), undergraduate programs (like Penn's Networked and Social Systems Engineering), summer programs (like the Data Science for Social Good programs at University of Chicago, Georgia Tech, and the University of Washington), and individual centers (like the USC Center for Artificial Intelligence in Society and the Centre for Social Services

Engineering at the University of Toronto). It is obvious that the interest in such areas is high, and WashU has a combination of strengths in the Social Sciences and their relationship to Data Science. These strengths would allow us to rapidly become a significant force in this area, which definitely has room for more entrants, especially in the space of graduate programs and research. Our main challenges in achieving such a position are:

1. **A consistent funding mechanism for interdisciplinary PhD students:** Different schools and departments have very different expectations for how students will be funded during the course of a PhD program. While SEAS typically uses the apprenticeship model, students in political science, for example, have higher teaching expectations and an advisory relationship with their dissertation committees. To develop an interdisciplinary program that is attractive to students from diverse backgrounds and allow them the flexibility that we wish to, it will be critical to develop appropriate funding mechanisms and expectations.

2. **External visibility:** We need to make the world aware of our strengths and efforts in this area. This is important for overall success, but also in the short term to get high-quality PhD students. Therefore, a dedicated marketing effort is needed that will pull together and highlight existing strengths and activities and help us attract additional talent (both faculty and graduate students) and increase our profile.

3. **Density of relevant faculty:** While we have core strengths that can bootstrap the effort, it is critical to hire rapidly in a number of the areas mentioned above to build critical mass in methodology. It may also be important to understand how to incentivize other schools to participate actively, given that they will have to devote some resources to the program.

Outlook

WashU can become a leader in socially focused applications of computation and data science. We have among the absolute best schools of social work and medicine in the country, and well-regarded programs in the social sciences, economics, and business. With appropriate investments, especially in faculty lines, support for interdisciplinary PhD students, seed grant funding, and staff support, we would expect, in a three-to-five year timeframe, to see:

1. A steadily improving, high caliber, and diverse set of applicants to the new interdisciplinary PhD programs in DCDS. The strength of a research enterprise can only be sustained with top PhD students, and this is an important measure of success.

2. Two or three competitive large center proposals to NSF, NIH, etc., enabled by the existence of DCDS.

3. An increase in external funding for collaborative projects between DCDS PIs from different disciplines. For areas that do not traditionally receive as much grant funding (e.g., political Science), this measure of success should be appropriately adjusted and focus on different metrics, such as top-tier publications.