MICHIGAN INSTITUTE FOR DATA SCIENCE
UNIVERSITY OF MICHIGAN

Michigan Institutes for Data Science: Successes and Challenges

Overview and history of the institute. The Michigan Institute for Data Science (MIDAS) is a virtual institute that is "the gathering place" for the University of Michigan (U-M) data science community, with ~350 affiliate faculty members whose research covers data science theory, methodology, and a wide range of application domains. In addition, our community includes >1000 data science students and a network of >100 staff data scientists. *Our mission* is to strengthen University of Michigan's preeminence in data science and to catalyze the transformative use of data science in a wide range of disciplines to achieve lasting societal impact.

MIDAS was established in 2015 as the main component of U-M's Data Science Initiative. However, individual faculty members had been influencing faculty hiring, building data science research collaboration, organizing community events and offering educational opportunities since 2008. Sufficient momentum was accumulated through such efforts that culminated in the five-year, $100M, Data Science Initiative, including MIDAS and a few existing computing and statistics infrastructure groups on campus.

Funding model. The vast majority of MIDAS funding comes from the university; external grants, industry funding and philanthropy contribute only a small percentage. For the first 3 years, funding was a combination of direct investment from the Provost and matching funds from each U-M school and college.that participated in the Data Science Initiative. Matching priorities across units was tricky, and it was difficult to ensure that all parties paid their "fair" share. Therefore, the system has since been changed to have all U-M funding coming from the Provost. The advantage of this funding model is the stability. The disadvantage is that the size is constrained and therefore limits the scope of our activities. We are currently in discussions on developing other sources to augment our funding.

Research. The most important goal of MIDAS is to promote data science research excellence within and across disciplines. We play the dual role of leader and enabler. As a leader for data science research on U-M campus, we focus on fostering excellence, coupled with responsibility. We facilitate new ideas and major grants through research working groups. As an enabler, we support the application of data science in any research domain, and assist researchers in developing the skills, teams, and tools they need to succeed. We also manage key datasets for the campus.

Success stories. 1) Strategic funding priorities. In the early stage of MIDAS we provided a percentage of start-up funds that allowed U-M to attract a number of high-profile faculty members as well as rising stars. We also provided an unusually large amount of funding ($10M total) to a few strategically chosen research areas (healthcare research, social science, transportation research and learning analytics) to capitalize on the existing research strengths at U-M and quickly demonstrate the potential of data science. Now that we have built the foundation, our funding priority has shifted to using smaller but more numerous grants to grow data science capacity in a large number of disciplines. 2) Transforming disciplinary research. Our "Data Science for Music" research initiative funded projects in music theory, performance, audience engagement and computer music. This initiative, and follow-on research activities, exemplifies a creative approach to demonstrate how data science can work with even an unlikely partner for transformative effects. 3) Using our central position to rally the entire data science community. Our rapid response to the COVID-19 pandemic includes funding 7 interdisciplinary COVID-19 projects with an expedited process (30 days from RFP release to award announcement). The projects range from theory development for epidemiological modeling, to high precision outbreak detection, to maximizing remote learning outcomes, to minimizing disparity in patient treatment and community resources. Concurrently, we organized research working groups on improving COVID-19 data quality and on data-driven campus reopening strategies. Our ability to mobilize the broad research community cannot easily be matched by traditional departments. 4) We recently organized a series of activities around reproducible data science, including a Reproducibility Challenge, a series of Reproducibility showcases, a Reproducibility Day, and an online collection of research tools and methods for reproducible data science. Research teams on campus shared their approaches and tools for reproducible research and we organized them into a conceptual framework. Our "bottom-up" approach allowed us to highlight

what issues are critical for our researchers and what approaches have been effective for them. 5) To build a diverse and inclusive data science leadership for the future, we started organizing an annual Consortium for Data Scientists in Training since 2019. Consortium members are data science graduate students and postdocs from ~30 major research universities and minority-serving institutions, with the majority of them being women and under-represented minorities. We provide an opportunity for them to receive feedback for their research, career mentoring, and a platform for them to build a professional network.

Challenges. 1) Although we have enjoyed some success in building collaboration to enable data science's transformative potential, and have demonstrated our role of intellectual leadership, there remain many issues on a very diverse campus. For example, many departments are developing their own data science efforts. Our mandate in such cases is to build strong connections with these efforts on campus so that we can provide strategic support with realistic expectations. 2) Strategically positioning resources is also challenging because researchers have diverse data needs and are at diverse skill levels while our resources are very limited. Recently we started building a staff scientist team to provide project support. But much needs to be carefully thought out in order to meet the diverse research needs.

Education. MIDAS is not a degree granting unit. Our education offerings, therefore, complement the degree programs that are mushrooming on campus and provide unique learning opportunities. These include 1) A Graduate Data Science Certificate program that allows graduate students enrolled in any program to receive data science training through classes and projects. 2) A Michigan Data Science Fellows program that supports outstanding postdoctoral fellows. They are mentored by an interdisciplinary team of faculty; in turn, their interdisciplinary research helps make the data science community more interconnected. 3) Short courses tailored to teach specific data science skills to an audience with skills in particular disciplines. Beyond these, *our major success* lies in 1) providing extracurricular training opportunities through mentoring student data science clubs and their research projects; 2) providing students with real-world research experience through industry and community projects and industry-sponsored data challenges. A few highlights last year were a March Data Madness (Basketball) Data Challenge, a Data for Public Good symposium organized by our students (with MIDAS supervision), and an industry career series.

Translating academic research into societal impact. MIDAS has sought collaboration with industry partners since the very beginning. We aimed to inspire research with real-world problems and enable better research with real-world data. As we build partnerships with more industry partners, our goal is also evolving. Recently, we have invested much more effort to work with government and community partners to help them build the capacity for data-driven policy making.

Successes: 1) We have developed a formal collaboration relationship with the National Vehicle Fuel Efficiency Lab (NVFEL). As an example of solving the "data rich, insight poor" problem that are facing many organizations, we helped NVFEL identify key questions that can be addressed through applying cutting-edge data science methods to their own data, and designed projects together to answer the key questions. Our faculty members then supervise a team of students to carry out the projects with NVFEL funding. 2) MIDAS works with the City of Detroit to develop a number of data science projects that will improve the city's public transportation, food distribution, and the relationship between communities and the police force. MIDAS provides the data science expertise, helps define the problems, organizes the data, carries out the research and ensures the validity of the take-home lessons.

Challenges: Bringing positive social change through data science is a noble cause. However, the biggest challenges that we face are that the impact of such work is hard to measure, and may not yield many publications and grants that are the "hard currency" in the research world.

Many of the challenges that we list in this document are certainly common to many data science institutes. Much of the enthusiasm, creativity and the collaborative spirit that we see in our data science community are also surely common to other data science communities. We look forward to the opportunity to learn from others' successes and tackle the challenges together.