Michigan Institutes for Data Science: Highlights of 2021

Overview 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 and AI community, with ~430 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 AI and to catalyze their transformative use 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 a five-year Data Science Initiative, including MIDAS and a few existing computing and statistics infrastructure groups on campus. The Data Science Initiative has concluded, but MIDAS continues to carry out its mission. The vast majority of MIDAS funding comes from the university; external grants, industry funding and philanthropy contribute only a small percentage. The university's funding is a combination of direct investment from the Provost and matching funds from a number of U-M schools and colleges.

Our research and community building programs. The focus of MIDAS is research. We enhance the data science and AI research capacity of the university. We lead, facilitate, coordinate or assist, depending on the needs of the campus partners. Our approaches include:

- Leading infrastructure, training and research consortium grants.
- Enabling ethical and reproducible data science and building resources.
- Facilitating interdisciplinary research ideas and teams.
- Consulting and coordinating for major grant proposals.
- Enabling collaboration between U-M data science and AI researchers and industry, academia and public sector entities.
- Enabling Data for Social Good research projects to increase the impact of faculty research.
- Training for domain scientists.
- Providing research resources including seed funding and access to datasets.

Our community building activities include a large number of events such as the weekly data science and AI seminar series, annual data science and AI symposium, faculty research pitch events, and connecting students with faculty research projects. We also organize a number of activities specifically to increase diversity, equity and inclusion in our community, including Women+ Data Science research and career series and the annual Data Science Consortium for postdocs and graduate students at major research universities and minority-serving institutions.

We highlight below two of our activities that we consider particularly noteworthy.

Raising awareness of and building resources for reproducible data science. In 2020, MIDAS organized a Reproducibility Challenge (https://midas.umich.edu/reproducibility/). We used a functional definition of data science reproducibility: the validation of research findings, allowing for small variations in data, code, statistical assumptions, and computing environment. U-M researchers submitted reports about their work to improve reproducibility in their research projects or their fields. Research fields reflected in these submissions included biomedical and healthcare research, physical sciences, engineering, and social sciences. Based on submissions selected by an interdisciplinary judging panel, we organized a Reproducibility Showcase and a Reproducibility Day where research teams gave presentations and tutorials. We also built an online collection of tools and processes. Our Challenge revealed an admirable amount of effort from researchers toward data science reproducibility, including projects focused on: tools and methods to document and to share data, code and workflow; methods to reduce errors and variations in the study design; methods to validate their own results and others' results; understanding factors that contribute to the irreproducible results associated with a particular methodology; and definition and quantification of reproducibility. The concepts behind the majority of the submissions can apply broadly to many research fields that are data intensive; however, most of the processes and tools were developed for reproducing a project or a type of projects. Some of these tools could be adapted easily to other domains, but others would require considerable work. We are now planning a Reproducible Challenge round 2, which will focus on generalizable processes and tools.

The Reproducibility Challenge also revealed a number of hurdles for reproducible data science. 1) Researchers who developed tools for reproducible data science often don't have the resources, time, or even

the mindset to generalize their tools. 2) Some tools have been thoroughly developed with a standard procedure available for users. But some require users to explore extensively on their own. In addition, the technical skills needed to explore the tools also vary widely. 3) How extensively a tool is validated also varies greatly. On one extreme we encountered broken links; on the other extreme some tools had been used extensively by collaborators. Most were in between. This is perhaps one of the major obstacles for researchers to adopt tools that are developed by others. 4) In building tools for reproducible research, how we balance between having tools that are easy to use and having the ability to inspect, validate and modify such tools. A blackbox will deter some users because they can't validate what they do. On the other hand, packaged tools attract users who need quick solutions to reproducibility issues.

We believe academic data science institutes can play a unique role to make data science research more reproducible. Our recommendations are: 1. Data science institutes can play an important role to standardize, generalize, validate and disseminate the methods and tools across application domains, and thereby attain actionable reproducibility. 2. These centers should help researchers adopt the most efficient methods and tools to improve the reproducibility of their research. 3. They should strike the right balance between avoiding blackboxes and making the tools more transparent and easy to use. 4. Promoting best practices and tools for reproducible research will be more effective through the demonstration of how they benefit research in the long run. Researchers who responded to our Challenge indicated three main reasons why they have been involved in such work: making their research reproducible is their responsibility as scientists; developing tools and processes to make their work reproducible would greatly improve the efficiency of their own research and that of their collaborators; and irreproducible results would greatly harm their field of study. We believe these values are shared widely by our researchers, and our effort to ease the hurdle of adopting best practices will be met with enthusiasm.

Using Data for Social Good projects to promote socially engaged research. As Big Data and AI's impact on society becomes increasingly prominent, academic research needs to be not only curiosity-driven, but also use-driven – addressing challenges that our society faces, from social justice to environmental sustainability. MIDAS has been building collaboration with an increasingly larger number of government and community organizations to support their data-to-insight effort. Our goal is to enable socially engaged data science and AI research and translate academic research into positive societal impact. This is different from most of the Data for Social Good programs at our peer universities which focus on providing experiential learning for students. Our approach is to develop a trusted partnership between U-M researchers and the external partners, starting from small projects, and expanding the effort to include increasingly more U-M data science and AI researchers and external partners, and removing hurdles along the way. Three examples below illustrate our effort. In the coming year we hope to formalize our effort and seek sustainable funding.

A. MIDAS 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 designed a data analytics class with NVFEL funding, and the students carried out the research projects.

B. MIDAS is developing a collaborative relationship with the Native American tribal nations in Michigan. A MIDAS affiliate faculty member is leading a group of students to carry out the first project, which is to design a database for five tribes and to improve their ability to track and analyze their students' data. We view this as the first step to improve the tribes' and the state's capability to improve the educational outcomes of Native American students.

C. MIDAS has been collaborating with the City of Detroit to support their data effort. Last year we developed 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. Recently, we started a collaboration with both Detroit and Microsoft, funded by Microsoft Airband, to provide data analytics support for Detroit's digital inclusion effort. In all these projects, 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.